The mnemonic basis of subjective experience

Hakwan Lau, Matthias Michel, Joseph E. LeDoux and Stephen M. Fleming

Abstract | Conscious experiences involve subjective qualities, such as colours, sounds, smells and emotions. In this Perspective, we argue that these subjective qualities can be understood in terms of their similarity to other experiences. This account highlights the role of memory in conscious experience, even for simple percepts. How an experience feels depends on implicit memory of the relationships between different perceptual representations within the brain. With more complex experiences such as emotions, explicit memories are also recruited. We draw inspiration from work in machine learning as well as the cognitive neuroscience of learning and decision making to make our case and discuss how the account could be tested in future experiments. The resulting findings might help to reveal the functions of subjective experience and inform current theoretical debates on consciousness.

Every conscious experience comes with distinctive qualities, such as associated colours, sounds, smells or emotions. What an experience feels like is called the 'subjective quality' or 'subjective character' of the experience. These subjective qualities are sometimes said to be ineffable. For example, most people are familiar with the simple perceptual experience of seeing red. But it would be quite difficult to explain what it is like to see red to someone naive to this experience. If pressed to describe the experience, one would probably use comparative statements to relate it to other experiences: "seeing red is a bit like seeing pink or orange; it is a bit like seeing purple, more so than it is like seeing blue; it is nothing like seeing black, or white," and so on. People use these comparative statements among similar experiences to tell others what it is like to have a particular experience.

According to a relational or structural account of conscious experience, the subjective character of a conscious experience can be exhaustively described by a set of precise comparative statements relative to all other possible experiences[1–3]. For example, the subjective character of seeing red can be fully described by statements relating red to other experiences, including all perceptible colours. These comparative statements might be difficult to express verbally, but can be succinctly captured in terms of distances within a similarity space. If this view is taken seriously, the brain must encode, maintain and exploit this spatial relational structure between all experiences. Current work on the cognitive neuroscience of memory, mental imagery, decision-making and emotions can help us to answer whether and how this scheme is implemented in the human brain.

The relationship between consciousness and memory has long been investigated. In particular, memories have been categorized on the basis of accompanying subjective experiences during memory recall, distinguishing between implicit and explicit memory[4]. Implicit memories are procedural and they have little or no consciously accessible content, whereas explicit memories can be defined by their conscious content. Somewhat less attention has been paid to how memory mechanisms might contribute to conscious perception[4–6]. However, some theorists have argued that implicit procedural memory might underlie the ability to compare sensory experiences and thereby determine simple subjective qualities[7]. These implicit mnemonic mechanisms could explain the subjective qualities involved in simple perceptual experiences, as in seeing red.

More complex experiences are likely to involve both explicit memories and implicit memories. For example, some people are familiar with the experience of tasting steamed grouper fish with soy sauce. Thinking about the qualities associated with this experience invokes complex processes including semantic categorization and autobiographical memory, without which the experience would be strange and unfamiliar. In the late nineteenth century, the sensory physiologist Ewald Hering noted that memory holds consciousness together. Around the same time, Hermann von Helmholtz proposed that past experiences influence perception by supporting unconscious inferences. In the twentieth century, there were numerous demonstrations of how expectations, which are extrapolations from memory, influence what people consciously see and hear[8–14]. This literature highlights how explicit memory provides semantic knowledge, which in turn forms the conceptual basis of perceptual meaning.

In this Perspective, we suggest that conscious experiences derive their subjective character from both implicit and explicit memories. We review current work on the cognitive neuroscience of memory, mental imagery, decision-making and emotions to explore how memory could underlie subjective experience. First, we introduce the concept of a mental quality space that defines the similarity between a particular experience and all other experiences. Next, we consider the degree to which theories of consciousness incorporate this space using self-monitoring. Then we focus on a specific higher-order theory of consciousness that depends on implicit self-monitoring and explicit memory replay to generate rich subjective experiences. We discuss the computational advantages of memory replay and conclude with recommendations for testing the higher-order mnemonic view in future work.

Mental quality space

According to some accounts of consciousness, the subjective quality of a conscious experience is determined by its comparison to other experiences[1–3]. If one can tell precisely how similar an experience is to all other possible experiences, perhaps that

is a complete description of that experience. One version of this view, known as quality space theory[1], defines similarity between two experiences as the inverse of the pairwise discriminability between the relevant stimuli, as assessed empirically. Discriminability refers to the extent to which one stimulus can be distinguished from another, as assessed using psychophysical testing. This concrete definition avoids the circularity of defining the subjective quality of one experience in terms of the subjective character of other experiences, which might also need explaining. Instead, mental qualities are identified functionally. For example, scarlet and crimson are subjectively similar to each other because they are less discriminable from each other than from blue. This relationship can be shown graphically: the subjective quality of a colour experience is determined by its position in a space defined by the discriminability of each colour from all other colours (FIG. 1a). Because the pairwise distance is defined by each person's perceptual experiences, the space encapsulates a great deal of knowledge that the subject has about their own perceptual capacities. Thus, each person's quality space would be different.

It is implausible that humans explicitly access and visualize a mental quality space in everyday life, owing to its vastness and multidimensionality. Unlike the caricature of a mental quality space for colour, a full mental quality space would contain all possible experiences, across multiple modalities (such as vision and hearing). Different modalities are defined on the basis of these similarity relations; experiences from the same modality are subjectively and functionally more similar than experiences from different modalities. Although cross-modal similarities might seem less important for determining subjective quality than similarities within a modality, they are still relevant. First, qualities can often be meaningfully compared across modalities along some dimensions, such as intensity[15]. Second, the contributions of cross-modal similarities might be subtle yet meaningful. For instance, a face can be perceived as more similar to the sound of thunder than to a gentle stroke on the back of one's hand. These cross-modal similarity relationships partly constitute the subjective experience of seeing that face, which might be troubled rather than serene.

Despite the complexity of the quality space, the full space containing all possible experiences can be plausibly implemented in the brain. The nature of sensory representations in the human brain permits an implicit implementation of quality space through two properties of sensory representations. The first relevant property is sparseness: very few sensory neurons need to be activated to signal the presence of a certain feature in the environment[16]. In a hypothetical scenario of extreme sparsity, each neuron would have a unique label based on what stimulus it primarily responds to (and therefore can be said to represent). In reality, sparseness comes in degrees. Human sensory cortices show a relatively high degree of sparsity[16]. For example, in the visual cortex, a typical simple stimulus excites only a small number of neurons (FIG. 1b). By contrast, the prefrontal cortex demonstrates a relatively complex coding scheme. Neurons in this region have a relatively high degree of mixed selectivity[17,18], meaning that most neurons respond to many different stimuli, to varying degrees. In the prefrontal cortex, neurons with mixed selectivity encode multiple aspects of a stimulus, task or motor response simultaneously.

The second property of sensory representations is smoothness. In the mammalian sensory cortex, coding is smooth[19,20]; the content typically conforms to a continuum, rather than to discrete, absolute categories. For example, humans subjectively see individual colours as falling within a continuous space; purple is on the colour continuum between red and blue. This continuity is a consequence of similarity across neural representations (FIG. 1b). Smoothness is especially obvious within a single modality such as vision. But even across sensory modalities, there is considerable interdependence and interaction between sensory cortices[21]. Importantly, some modality pairs (such as vision and hearing) are probably more mutually interdependent than other pairs (such as olfaction and touch). Certain senses have strong interconnections at the neuronal level and are similar in containing information regarding the spatial location of external stimuli. As such, there might also be smoothness across sensory modalities, which could allow the similarity between two stimuli in different modalities to be meaningfully assessed.

However, sensory coding is not necessarily smooth. For example, in the visual system of the mantis shrimp, different colours are coded symbolically as absolute categories[22]. Any colour is either detected or not by the relevant sensors. Different colours are categorically different and cannot be meaningfully compared in a fine-grained manner. In such a non-smooth space, there is no sense in which red is more similar to purple than to green.

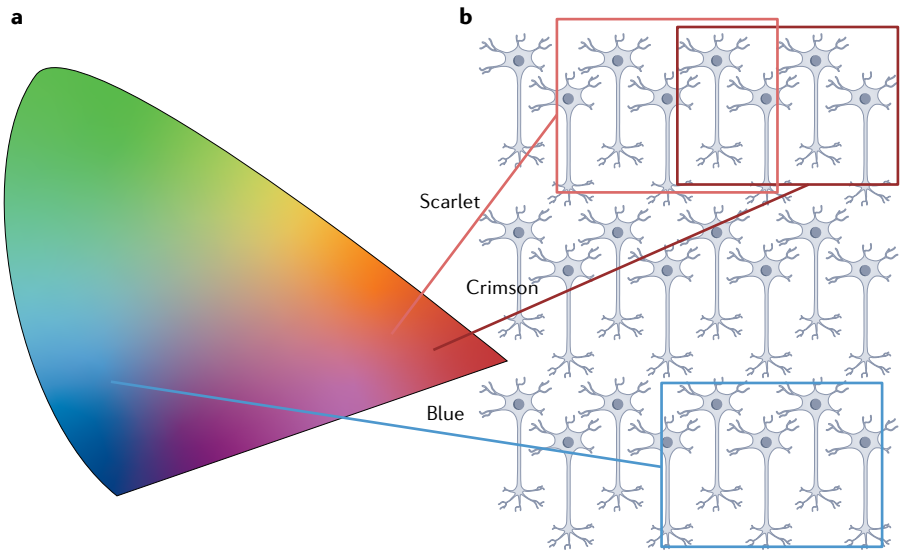

Fig. 1 | **An example mental quality space for colour. a** | In this space, the distance between two points reflects how subjectively discriminable the relevant stimuli are, for a given person at a certain time. Each point reflects the quality of the subjective experience of perceiving a stimulus, determined by how subjectively similar that stimulus is to all other stimuli in the space. This space is smooth, in that the content is continuous rather than categorical. **b** | Smoothness can be achieved by neuronal population coding. For example, two shades of red (scarlet and crimson) are represented by partially overlapping neuronal populations, indicating their high degree of similarity. A distinct colour, blue, might be represented by a rather different set of neurons. This coding scheme is also sparse; only a small set of the neurons is active to signal a specific colour. Taken together, knowing all the tuning properties of the relevant sensory neurons is tantamount to knowing the mental quality space.

Taken together, sparsity and smoothness enable the sensory quality space to be efficiently implemented in the brain. Sparsity means that two sensory representations can be approximately referred to by their neuronal 'addresses'— the location of their neural representations in sensory quality space. Two neural populations with different addresses probably represent highly discriminable contents. Owing to smoothness, one can also meaningfully infer the discriminability of the content of two neuronal populations. Thus, a functional brain mechanism that has access to the precise functional layout of human sensory cortices would possess an implicit grasp of the human mental quality space. Given a percept and the layout of the quality space, such a mechanism would 'know' exactly how similar the relevant experience is to all other stimuli encoded within the entire sensory space.

Theories of consciousness

The mental quality space is one account of how the subjective character of conscious experiences are represented. This idea has been explored by theorists holding otherwise contrasting views about consciousness. Existing cognitive neuroscience theories of consciousness fall into three broad classes: local theories, global workspace theory and higher-order theories. These theories can be distinguished along three main dimensions: the brain mechanisms responsible for subjective experience (the neural correlates of consciousness), the degree to which consciousness is important for executive cognitive functions, and whether self-monitoring is required for consciousness. Self-monitoring here refers to the implicit, automatic evaluation of the reliability of one's own sensory processes. We focus on the role of self-monitoring across theories because this capacity may entail an implicit grasp of the quality space. Local theories and global workspace theories do not consider self-monitoring to be necessary for consciousness, but it is a key component of higher-order theories. For further comparison between theories (including some not discussed here), see REFS[23–25].

Theories without self-monitoring. According to local theories of consciousness[26–32], neuronal activity in sensory cortices solely determines the subjective character of an experience. These theories are called 'local' because consciousness of a given perceptual feature does not require any cognitive processes beyond the sensory areas where that feature is first processed. Consciousness does not require higher cognitive functions such as self-monitoring and working memory (the ability to hold information in mind for short-term maintenance and manipulation), nor does it necessarily facilitate such functions. To explain subjective qualities, proponents of local theories have adopted views similar to the mental quality space account described above[33]. One proposal is that the qualitative character of an experience within a given category (such as colour), is determined by the relative differences between the patterns of cortical activation generated by other stimuli within that category[33]. According to this account, local lateral connections in sensory brain areas implicitly capture a quality space for each category of experience.

However, some neural findings are difficult to explain using local theories. For instance, subliminal stimulation (such as visual presentation of stimuli that participants are not aware of) activates the same sensory circuits as fully visible stimuli[31,34]. Because similar activity can be triggered both consciously and non-consciously, activity in early sensory circuits seems insufficient to support consciousness.

To meet this challenge, one could postulate that the neural dynamics or activity level within sensory areas determine whether subjective experience occurs. Some local theories posit that consciousness arises only when the relevant activity crosses a threshold[33]. Others hypothesize that recurrent activity between sensory areas is necessary for consciousness[26,27]. An account along either of these lines can explain the activation of the early sensory cortex by subliminal stimuli. However, conscious experience is also found in individuals who lack the early visual cortical area V1[35–38]. Thus, at least in the case of visual awareness, neural recurrent activity would need to take place outside the early sensory cortex[33].

Considering the downstream impact of local sensory activity can help to explain why such activity needs to reach a certain threshold for conscious experiences to arise; it is possible that the threshold is required for signals to be transferred to other areas. An influential alternative to local theories of consciousness is the global workspace theory. According to this theory, sensory signals become conscious through entry into a 'global workspace.' The global workspace is a hypothesized system akin to working memory. In the global workspace, neural signals from sensory areas are amplified and stabilized across the brain[39–41]. The theory postulates that non-conscious perception involves neural activity only in early sensory areas, whereas conscious perception involves a more widespread pattern of activity across the brain referred to as 'global ignition.'

Although multiple findings have been interpreted as confirming that neural activity is more widespread during awareness than during lack of awareness[42], there are arguments against the global workspace theory[23,24]. A central problem with the theory is that broadcast seems to determine the ability to cognitively access information and to perform tasks, rather than to enable subjective experience per se. For example, when holding information in working memory, this information seems to be globally accessible by different processes. Yet, working memory content is not confusable with normal perception of that content. For instance, simply thinking about a friend's face is not the same sensory experience as seeing it. Thus, entry into the global workspace does not seem to always entail subjective experience.

Further evidence regarding the role of working memory in consciousness comes from individuals with aphantasia, who do not experience vivid mental imagery. According to a preprint, these individuals can perform comparably to individuals with intact mental imagery in working memory tasks, such as mental rotation, which are often interpreted as involving the maintenance and manipulation of perceptual representations[43]. And yet, they do not consciously experience vivid imagery while doing so[44]. The mechanisms by which individuals with aphantasia accomplish working memory tasks should be investigated further, because they probably differ from the mechanisms used by individuals with intact visual imagery[45]. Still, these findings suggest that the function of global broadcast, which seems to be intact in individuals with aphantasia when they perform working memory tasks, is not always accompanied by conscious experience. Additional discussion of the relationship between consciousness and higher cognitive functions, such as working memory, can be found elsewhere[46,47].

Higher-order theories. A third class of theories, known as higher-order theories, avoids these problems of local theories and global workspace theories. According to higher-order theories, subjective experience is determined by specific self-monitoring mechanisms in the prefrontal and parietal cortices[23,24,48,49]. These mechanisms are implicit, so consciousness does not require

explicit cognition about oneself. Because these implicit self-monitoring mechanisms are highly specific — unlike in the global workspace theory — conscious experiences are not broadly associated with higher cognitive functions.

In this context, monitoring refers to the process of determining the source or reality of a memory or percept. People generally do not confuse the contents of working memory with normal perception. For instance, holding visual images in mind does not result in mistaking those images for the outside world. The same is true when recalling long-term memories: people do not mistake the recalled content for a current percept. Under normal circumstances, one can also easily judge whether the recalled image was seen or whether it came from one's imagination. These distinctions all depend upon monitoring. In the memory literature, monitoring the source of mnemonic representations is called source monitoring, whereas monitoring whether an event actually occured is known as reality monitoring[50–52]. A failure of reality monitoring would mean that one confuses actual past experiences with one's imagination.

Following this tradition, the notion of perceptual reality monitoring focuses on the ability to identify the nature of ongoing perceptual (rather than mnemonic) representations[53]. When a perceptual signal occurs, it could reflect the presence of an external stimulus. However, it could also be driven internally; for example, by imagination or working memory maintenance. Although the content of working memory can reflect reality (rather than imagination), it does not reflect the present state of perception. Alternatively, such a perceptual signal might only reflect spontaneous noise. Perceptual reality monitoring is the function of distinguishing between these possibilities. Although closely related to source and reality monitoring in memory, perceptual reality monitoring probably depends on distinct mechanisms[54,55].

Perceptual reality monitoring is important for determining conscious experiences, because the mere presence of a perceptual signal does not create a conscious experience. Patients with damage to the primary visual cortex lack corresponding conscious experiences in the visual modality. However, the internal perceptual signal seems to be sufficient for them to perform well in some visual tasks[35–38]. One possibility is that the perceptual reality monitor fails to recognize the source of such a signal as reflecting current external reality, so conscious experience is absent. When a perceptual signal is generated internally, as in mental imagery, there is a conscious experience, but not one of typical vision reflecting the outside world. Importantly, in aphantasia, internally generated perceptual signals do not come with vivid conscious experiences.

However, some internally generated perceptual experiences do form part of conscious experience. Although external stimulus input is absent during hallucinations and dreams, the corresponding experiences feel as though they represent reality, and are highly similar to the normal conscious experience of seeing. The explanation of this feeling is that the mechanisms contributing to perceptual reality monitoring can malfunction. There are two main ways in which failures of perceptual reality monitoring could lead to hallucinations[56]. The reality monitoring mechanisms could be dysfunctional — as found in certain clinical populations[52]. Alternatively, anomalies in the systems that generate inputs to the reality monitoring system, such as hyper-activation of sensory cortices or an absence of cognitive control, could also lead to impaired reality monitoring[57]. In either case, the perceptual reality monitor misinterprets internally generated perceptual signals as reflecting stimuli in the external world, leading to hallucination. Such a failure could also explain conscious experiences in dreams[53].

It is important to note that perceptual reality monitoring mechanisms operate implicitly: they are automatic and largely not subject to volitional control. The distinction between explicit and implicit reality monitoring can be illustrated by the phenomenon of lucid dreaming[58]. During lucid dreams, explicit reality monitoring enables one to recognize the illusory nature of the experiences. Despite this explicit knowledge, these experiences continue to feel as though they reflect the present reality, presumably because the perceptual reality monitoring mechanism still implicitly categorizes the relevant perceptual signals as reflecting external reality.

Perceptual reality monitoring forms the basis of our version of a higher-order theory of consciousness. According to this account, consciousness does not depend solely on local activity in sensory brain areas, nor on activity in a global workspace across the brain. Rather, consciousness depends on implicit self-monitoring of the nature of perceptual signals. The mechanisms of perceptual reality monitoring are discussed in the following section.

Self-monitoring in the brain

Mechanisms for perceptual reality monitoring are likely to depend on prefrontal and parietal areas, similar to mnemonic reality monitoring[52,54,59–61]. To infer the presence or absence of an external signal, one needs to implicitly monitor the statistical properties of one's internal sensory responses[62]. A computational framework has been proposed in which self-monitoring tracks the signal-to-noise statistics of sensory representations[63]. For example, if baseline noise in a sensory area is very high, a certain level of activity might be less meaningful than if baseline noise is low. Along these lines, monitoring might be achieved by tracking the precision or clarity — rather than content — of sensory representations within the mental quality space[64].

Proponents of local theories of consciousness could argue that the activation profile within the sensory cortices is sufficient to account for different perceptual sources without a dedicated monitoring mechanism. During working memory, mental imagery, and episodic recall, neural activity is entirely top-down, from frontoparietal to sensory cortex. Because top-down (feedback) projections terminate at different cortical layers than bottom-up (feedforward) projections (from sensory cortex to frontoparietal areas), they lead to spatially distinguishable patterns of activity. In principle, these local differences could explain the phenomenological differences between imagery and normal perception because imagery is associated with top-down signals and perception is associated with bottom-up signals[65]. However, in dreams and hallucinations there is no external stimulus and therefore a lack of bottom-up, feedforward input, yet the percepts associated with dreams and hallucinations are misintrepreted as reflecting reality. Thus, it is unclear whether the layer profile of local sensory activities alone can explain the phenomenology during these conditions.

Returning to the role of prefrontal areas, atypical prefrontal activity has been associated with dreams and hallucinations[66–68]. Reduced activation levels are often found, relative to the level found during ordinary conscious experience, although the extent of such reduction has been debated[69]. During dreaming and hallucinations, reduced prefrontal activity might lead to the perceptual reality monitoring system going 'offline', such that internal signals are misinterpreted as being externally triggered and dreams and

hallucinations are accordingly misjudged as reflecting reality.

Further data suggesting the importance of the prefrontal cortex for perceptual reality monitoring come from transcranial electrical stimulation. Stimulation of the prefrontal cortex changed not only the rates of lucid dreaming, but also the reported realism and frequency of dreams (see the supplementary information in REF.[58]). Thus, both explicit and implicit perceptual reality monitoring seem to depend on activity in the prefrontal cortex. This conclusion is congruent with the general motif that explicit and implicit forms of the same computation often depend on similar brain regions[31,34].

Further critical evidence localizing perceptual reality monitoring to the prefrontal cortex comes from neuronal recordings in non-human primates. Activity in the dorsolateral prefrontal cortex in macaque monkeys distinguishes contents maintained in working memory from current percepts[70]. Whereas perceived and memorized contents can both be decoded from this region, they are represented by largely distinct neuronal populations. Accordingly, it has been theorized that prefrontal activity plays a part in distinguishing between internally generated and externally triggered sensory activity. This finding might help to explain why the atypical prefrontal activity during dreams and psychosis in humans[66–68] can lead to confusion between internally generated activity and reflections of external reality. In such conditions, alterations to prefrontal cortex function could lead to a failure to properly distinguish between internally and externally generated sensory activity, and therefore cause confusion between the two.

One potential function of a perceptual reality monitoring mechanism is to enable the smooth running of predictive processing. Predictive accounts of perception suggest that perception is a product of top-down and bottom-up interactions that create an internal model of the environment. In such schemes, it is important to keep track of primarily top-down (imagination) or primarily bottom-up (perception) states to generate appropriate predictions[53,71].

The putative existence of perceptual reality monitoring mechanisms, whether or not they are localized to the prefrontal cortex, supports higher-order theories of consciousness. Perceptual reality monitoring serves the role of implicit self-monitoring in determining whether perceptual awareness arises in a given situation. However, current accounts of perceptual reality monitoring do not address how subjective qualities are determined. Although reality monitoring might determine whether mental qualities occur consciously or unconsciously, this function alone does not explain why experienced qualities feel the way they do.

Higher-order mnemonic view

Combining the concepts of quality space theory and perceptual reality monitoring, we hypothesize that the subjective quality of an experience is not determined within the local sensory circuitry alone. Instead, higher-order mechanisms support how people automatically know what an experience is like without effort. The local theorist might argue that when early sensory signals are strong enough, the relevant knowledge is easily accessible by higher-order mechanisms. However, such access tends to be task-dependent, subject to attentional modulation, and often involves cognitive effort. This access seems incompatible with direct and effortless grasp of the subjective qualities of an experience, regardless of the task required at present. The higher-order mnemonic approach is an evolution of higher-order theories, and posits that consciousness depends strongly upon both implicit and explicit forms of memory. Implicit memory supports direct access to the mental quality space, whereas explicit memory provides complex categories, schema and emotions for everyday subjective experience (FIG. 2).

Implicit mnemonic process. Information relevant to subjective qualities is not easily verbalized and therefore might not be represented explicitly. Instead, humans might have some degree of implicit familiarity regarding how similar a conscious percept is to all other experiences. We propose that this access to the mental quality space depends on procedural memory, akin to a skill — one can be highly skilled in a procedure without being able to articulate how it is done[7].

Because the mechanisms for perceptual reality monitoring are involved in determining whether a subjective experience occurs, it is possible that the same mechanisms might also support the implicit procedural mnemonic process for accessing the mental quality space. Prefrontal and parietal cortices send top-down signals to various sensory areas targeting specific representations for the purposes of attentional modulation and inhibition of those representations[72–75]. As such, these prefrontal and parietal mechanisms must contain some implicit knowledge of what different sensory neurons represent. Computational models have highlighted how prefrontal circuits could store this knowledge abstractly, much like how computer programmes use variables and pointers to store memory addresses and to reference specific locations in memory[76]. Advances in the decoding of ensemble activity in the prefrontal cortex also show that neuronal populations can encapsulate rich information and enable meaningful abstraction and generalization over different contexts[18]. Thus, it is not implausible for prefrontal areas to implicitly track the mental quality space encoded in sensory areas.

Importantly, because the mental quality space reflects knowledge of one's perceptual capacities, an implicit grasp of this space probably depends upon learning. Early in development, the brain adapts to its sensory milieu and extracts regularities that wire cortical sensory circuits[77–79]. Sensory plasticity continues throughout life as an individual encounters novel sensory events[80–83]. Building and maintaining a

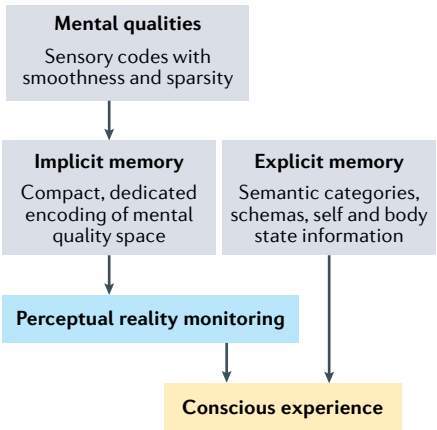

Fig. 2 | **The higher-order mnemonic account of consciousness.** Because of the smooth and sparse coding in sensory cortices, perceptual representations can be described in terms of subjective mental qualities. However, for these qualities to be meaningfully related to consciousness, the information needs to be succinctly encoded at a stage of processing that is readily available for cognitive access. We propose that an implicit mnemonic process represents the mental quality space in some specialized format. A higher-order mechanism with the relevant information regarding the functional layout of early sensory coding is key to this process. This mechanism is also hypothesized to support the function of perceptual reality monitoring, which determines whether perceptual awareness occurs at all. At some stage after this process of implicit, automatic self-monitoring, explicit mnemonic information is also taken into account. Together, both implicit and explicit memories contribute to subjective experience.

current mental quality space to support subjective experience might require continued learning about the world and how experiences relate to each other[62].

One intriguing hypothesis is that certain prefrontal areas might develop dedicated, specialized representations of the quality space. Insights come from how physical space is represented for spatial navigation in rodents. Neural representations of space often use a grid code, in which neurons respond to specific locations of the external environment in a hexagonal grid pattern[84–87]. These 'grid cell' neurons fire to specific landmark locations, which follow a regimented pattern evenly distributed across a physical space. Thus, their firing patterns can tell an animal how far it has moved in space. These cells have also been identified in humans[88–91].

There is evidence that similar grid cells also encode abstract conceptual spaces[89,92,93], as well as non-spatial sensory quantities such as the frequency of an auditory stimulus[94]. In one human study, the authors asked participants to imagine trajectories through a two-dimensional sensory space defined by the mixing of two different odours[95]. For instance, one region of odour space could be defined by a high quantity of odour A and a low quantity of odour B, and another region by a low quantity of A and high quantity of B. High quantity of both odours defined a third region. Voxels in the ventromedial prefrontal cortex showed activity reflecting hexagonal grid coding as participants imagined moving from one odour to another within odour space. Thus, this brain region might encode the relationships between different sensory stimuli. Relational coding in areas such as the ventromedial prefrontal cortex could represent the subjective similarity relationships between sensory experiences. This way of representing the mental quality space could underlie the human capacity to effortlessly access what an experience is like.

Specifically, we hypothesize a general mechanism that represents the entire mental quality space, including all possible subjective experiences across modalities (such as colour, sound, and touch). This space reflects the nature of neuronal coding in sensory cortices, with specialized 'spatial' representations in higher-order brain regions (such as the prefrontal cortex) enabling one to 'know' what it is like to have a certain subjective experience. Importantly, this knowing is not explicit: it happens automatically and the content is often hard to articulate. Neural codes that track multiple points within a quality space (such as hexagonal grid codes) could support implicit access to the quality of experience by making the relational information within the sensory cortices more readily available. This possibility remains to be tested carefully.

In summary, we propose that the sense of what an experience is like depends on implicit, procedural memories, stored within higher-order mechanisms that are also responsible for monitoring and determining the source of early sensory signals. The medial prefrontal region identified as possibly representing sensory quality space information[95] is distinct from the lateral prefrontal areas linked to reality monitoring. However, these areas are densely connected. We believe that a complex network of multiple prefrontal areas could interact to determine subjective experience, each having different and complementary roles (FIG. 3). The interactions between these areas might be particularly important for the contribution of explicit memories to conscious experiences and/or for monitoring the precision of quality space information.

The role of explicit memory. Whereas implicit quality space memories might be sufficient for determining the subjective quality of simple experiences — such as seeing red — explicit memories probably play a large part in determining the

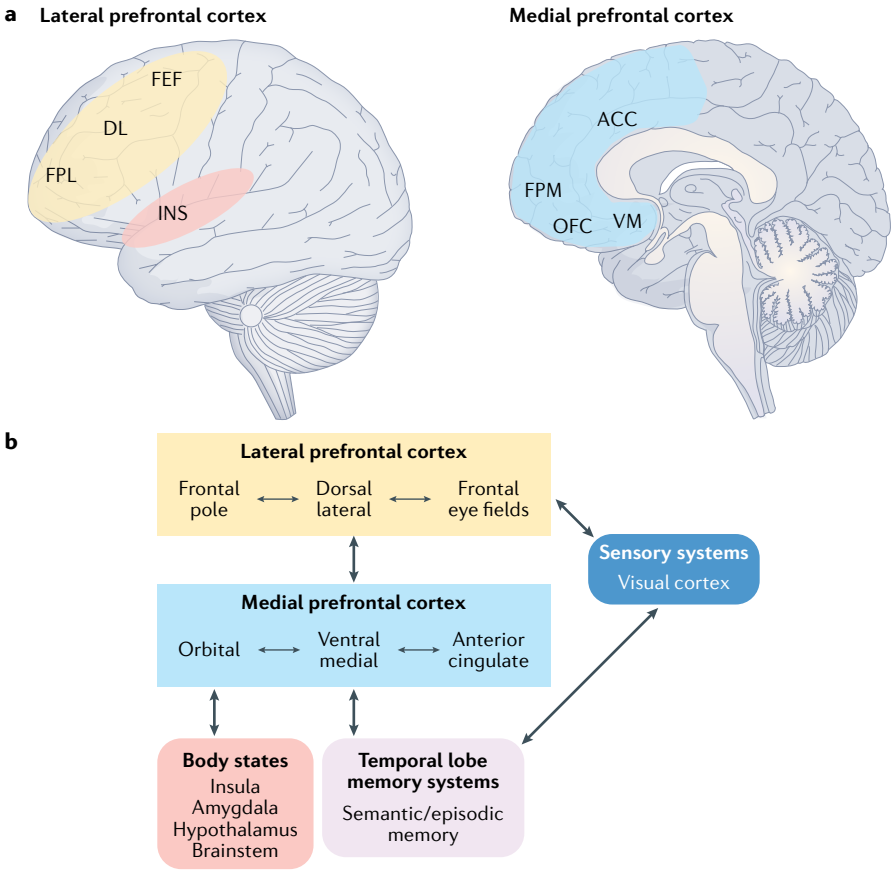

Fig. 3 | **Prefrontal cortex connections with memory systems. a** | Anatomical locations of some key areas within the prefrontal cortex, as well as the insula. **b** | Connectivity of prefrontal areas with sensory, memory and body state areas. Lateral prefrontal areas receive feedforward input from sensory systems and send feedback signals back to them. This pathway has been linked to perceptual monitoring functions. Although only connectivity with the visual cortex is shown here, other sensory areas also connect with the lateral prefrontal cortex. Medial prefrontal areas are also intricately interconnected with lateral prefrontal areas and they probably work together in some contexts. Medial prefrontal areas might be involved in representations of the quality space. Medial prefrontal areas also receive mnemonic information from the temporal lobe, which might be how semantic memories, including schemata, give sensory information meaning and how episodic memories contribute to the experience of everyday perception. Medial prefrontal areas are also involved in self-related processing and receive inputs from other cortical and subcortical brain processes for monitoring bodily state information. As such, these pathways might also be important for emotional experiences. FEF, frontal eye fields; DL, dorsolateral prefrontal cortex; FPL, lateral frontal pole; INS, insula; ACC, anterior cingulate cortex; VM, ventromedial prefrontal cortex; FPM, medial frontal pole; OFC, orbitofrontal cortex.

subjective quality of complex everyday experiences. Everyday experiences often concern emotions and thereby, one's bodily states and oneself over time. Everyday experiences also are not isolated incidents but rather form a coherent narrative within which individual experiences are interpreted. These complex processes require explicit memories.

Medial prefrontal areas that mediate between semantic and episodic memory circuits and lateral prefrontal areas that have direct inputs from sensory systems are crucial for integrating explicit memories and conscious experience (FIG. 3). For example, medial prefrontal areas might use memory to rapidly form predictions in perceptual inference[73,74]. Another related role of medial prefrontal areas might be to construct schemata that underlie humans' conceptual understanding of the world and themselves[7,96–98]. These schemata are collections of semantic memories about recurring objects and situations[98]. For example, when perceiving a restaurant scene, one naturally expects certain objects and sensations, such as seeing menus, touching a table, and smelling food. These conceptual templates help humans to acquire and organize memories and enable a better understanding of the present situation relative to one's goals.

Schemata, especially schemata concerning one's self, have also been proposed to play an important role in emotional processing. For example, stimuli associated with danger activate self and fear schemata via interrelations between explicit memory circuits and medial prefrontal cortex (PFC) (in addition to eliciting behavioural and physiological fear responses). Conscious emotions, in this view, are higher-order states that emerge in biologically or psychologically important situations[7,23,96,97,99].

According to this higher-order theory of emotions, lateral prefrontal integration of perceptual and memory signals with signals related to brain and body states to form situational, self-related and emotional schematic memories are key to emotional experience. These processes might be supported by medial frontal brain areas, which have been linked to processing of self-related information[100–104]. In particular, the anterior cingulate receives episodic memory inputs from the middle temporal lobes[105]. The ventromedial prefrontal cortex integrates these inputs. Moreover, both the cingulate cortex and the ventromedial prefrontal cortex connect with the dorsolateral prefrontal cortex[106]. Body state information[107] also reaches the dorsolateral prefrontal cortex via the insula[108].

Areas of the parietal cortex also mediate between memory and prefrontal circuits[109–114]. Several parietal areas are densely connected with specific prefrontal regions, so regions across the two lobes might work together on some functions[115]. These functions include the incorporation of autobiographical episodic memories[116] into prefrontal emotional mental models via temporal–parietal connections with medial prefrontal areas[106].

In summary, a broad network of prefrontal and parietal brain areas incorporate explicit mnemonic information into everyday conscious experiences, providing a high degree of informational richness (FIG. 3). These include brain circuits outside the prefrontal cortex that route through and interact with the implicit mnemonic functions that depend on specific prefrontal mechanisms.

Memory replay

To summarize, consciousness relies on implicit memory processes for the automatic access of mental quality space information, and on explicit memory to integrate complex schemata regarding one's body and oneself. As long as the perceptual reality monitor is involved in determining whether a perceptual signal should give rise to conscious experience, that signal is interpreted in terms of the relevant position on the mental quality space. Thus, one always 'knows' what a conscious experience is like.

Given that the explicit memory system itself does not require the perceptual reality monitor (FIG. 2), it might seem advantageous for the brain to store explicit memories in a compact internal format that would not necessarily contain fine-grained quality space information[117,118]. However, humans seem to project memory details back into the quality space during conscious recollection. Memory replay in computational models has various functional benefits, including enhancing learning, preventing memories from being overwritten, and prioritizing certain events for future planning[119]. However, replay in modelling contexts does not typically involve reactivation of detailed sensory representations. Instead, these replays take place via abstract internal states that are symbolic and categorical, rather than smooth and graded like the subjective qualities of conscious experiences. We hypothesize that memory projection in humans takes advantage of the quality space to improve generalization in learning and novelty detection. In turn, graded generalization and novelty detection might be some of the key advantages of having subjective conscious experiences, especially in memory recall and in comparing concurrent percepts with past experiences.

The potential benefits of projecting episodic memory information back to sensory space might relate to the computational characteristics of sparse and functionally smooth sensory coding. These properties are observed in humans and a wide variety of non-human animals, including in the olfactory system in fruit flies[120,121]. Although fruit flies are not necessarily conscious — they probably lack the relevant higher-order mechanisms — the organization of their olfactory coding might still inform the computational functions of the human sensory system. In particular, the fruit fly olfactory system has been compared with artificial neural networks[122,123]. Researchers have identified an active 'sparsification' architecture in the projections of the olfactory receptors to another anatomical structure known as the mushroom body, in which relatively few neurons project to a higher number of neurons, leading to a neuronal code that is sparser than at the initial stage.

A sparsification architecture facilitates high performance in two kinds of computational problems. The first is similarity search, in which an agent spontaneously comes up with similar examples to a stimulus. Sparsification can be useful for generalization in learning because graded generalization comes naturally if learning is performed on representations that smoothly link similar stimuli together[124]. The second problem that sparsification aids is determining whether a stimulus has been encountered previously, known as novelty detection. If a certain sparse and smooth representation has not been activated previously within a context, the relevant stimulus is likely to be novel. And yet, this novel stimulus can be meaningfully related to other past stimuli in terms of its relative similarity. Owing in part to these benefits, the sparse and smooth circuit architecture found in fruit flies can outperform even some of the best current computer algorithms for similarity search and novelty detection[122,123]. These benefits cannot in principle arise from sensory coding like that found in the mantis shrimp, which is symbolic and unsmooth[22].

Importantly, the smooth and sparse architecture of the fruit fly olfactory system is also present in human sensory systems.

We speculate that smooth and sparse architecture enables efficient 'model-based' learning when humans project episodic memories back to sensory space. Model-based learning involves constructing mental models (also called cognitive maps) about how events are causally related in the world[125].

Using the retrospective re-evaluation paradigm, one can assess whether people can understand the causal relations between events to maximize reward[126,127]. Conceptually, the retrospective re-evaluation paradigm tests updating the consequences of an initial action. As an everyday example, imagine that you have experienced dining in your favourite restaurant. The restaurant might be famous for a particular chicken dish, which you regularly order and which is rewarding to eat. Suppose at some point you get food poisoning from eating chicken at home, which leads to aversion to chicken in general. Consequently, you might lose the strong motivation to go to your favourite restaurant again. Rather than being due to direct experience at the restaurant (which has itself never been aversive), this change might be driven by reasoning through the potential consequences of eating chicken and weighing them accordingly in a mental model. The retrospective re-evaluation paradigm mimics a similar situation under gamified experimental settings. It was found that this process of indirectly updating the reward likelihood of an initial action requires the replay of the relevant experiences[126,127]. For example, researchers have tried to block episodic memory replay by requiring a concurrent effortful task — such as a working memory task — and found that retrospective re-evaluation was impaired[126].

Humans typically subjectively re-experience the specific details of the experience during memory replay[6,128,129]. Memories of a bad meal, when vividly recalled, are often about a specific dish, rather than a general category of food. As one recalls episodes in the sensory space where vivid details are represented, one re-experiences the perceptual experiences and how similar they are to other stimuli. Projecting information back into sensory space might enable a generalization across the fine-grained subjective similarity between specific experiences, rather than coarser generalization across different conceptual categorizations without replay. Future experiments could test this possibility, specifically in decisional and learning contexts involving episodic recall.

The functional benefits of replay might also apply to the prospective role of episodic memory, also known as mental time travel[128,129]. When thinking about the future, people tend to think in sensory terms, rather than in purely conceptual and categorical terms. One possibility is that imagery enables one to easily compare the similarity between future potential experiences and memories of past experiences. Projecting future scenarios in sensory space might facilitate fine-grained generalizations from past experiences, enabling people to maximize future reward and avoid threatening outcomes[124]. Additionally, by imagining the future in subjective sensory terms, one might more easily predict whether a certain scenario would lead to outcomes that have not been experienced in the past (within a certain context). It might be beneficial to anticipate novel scenarios to minimize unexpected risks and to optimize the balance between exploration (of new options) and exploitation (of known outcomes)[130]. These speculations remain to be tested.

In summary, quality space representations underpinning conscious experience also allow projection of memories back into sensory space. Projection has a range of computational advantages — including the capacity to generalize reward expectations across experiences that share sensory qualities, recognize novelty and similarity, and to plan within sensory, rather than conceptual, space. More broadly, projection enables relationships between experiences to be encoded within the naturalistic space of stimuli, rather than only between abstract concepts.

Conclusion

Conscious perception of a stimulus involves automatically and implicitly remembering how similar or dissimilar it is to other experiences, past and present. We propose that this comparison is a consequence of higher-order mechanisms in the brain that have learned the organization of the corresponding sensory representations. These higher-order representations can be considered an implementation of a mental quality space of similarity among previous simple experiences. These mechanisms enable one to effortlessly compare experiences of the past, present and future, in a fine-grained analogue manner. This automatic, implicit comparison gives even the simplest conscious experiences subjective richness, because a single percept encapsulates a great deal of self-knowledge. In everyday experiences beyond simple perception, explicit mnemonic processes further embellish this complexity.

If this account of the subjective quality of experience is correct, it pressures current theories of consciousness to make room for the relevant mechanisms of self-monitoring. Major cognitive theories such as the global workspace view have largely neglected these issues, especially regarding the role of implicit memories in determining subjective qualities. By contrast, local theories might be correct that sensory circuits could be described in the context of relational mental qualities, but this fact alone does not explain why people have immediate and automatic access to the relevant information. The subjective quality of conscious experience cannot meaningfully exist in a vacuum. The scientific challenge is to explain how people seem to effortlessly know what experiences are like. Existing theories are somewhat silent on these issues, and they also leave little room for explicit memories to be substantively involved in consciousness.

We speculate that implicit procedural memories of the mental quality space might be represented spatially via grid-like coding[84] in the ventromedial prefrontal cortex[95]. However, it is currently unclear whether such activity is spontaneously employed outside experimental settings, in which participants were trained to navigate over the space. Future experiments could test whether these representations spontaneously contribute to subjective experiences. For instance, outside the context of a navigation task, experimenters could causally manipulate the relevant brain activity and evaluate reported changes in experience.

The relation between the representation of mental quality space and perceptual reality monitoring also remains unclear. One hypothesis is that a medial prefrontal grid-like code represents location in a quality space informed by sensory activity, whereas other regions such as the frontopolar cortex might track the reliability of these signals. This view would align with these latter regions' role in tracking confidence in perceptual and memory tasks[131,132], and work showing that lateral frontopolar cortex monitors the uncertainty of medial prefrontal representations during decision-making[133,134]. Alternatively, subjective confidence in the presence of a stimulus, as given by the perceptual reality monitoring process, might be inherent to the quality space representation[135]. For instance, confidence in a discrimination judgement task has been reliably linked to the activation profile of medial frontal areas[55,136–138].

Finally, to further investigate the functional benefits of conscious memory recall, studies of aphantasia might be

particularly informative. Whereas memory recall can be blocked by concurrent working memory tasks[126], such a procedure probably also impairs cognition and decision-making in general. Individuals with aphantasia, on the other hand, generally have intact cognitive replay capacities. Although they lack the conscious experience of vivid mental imagery, this absence seems relatively selective. It will be useful to pin down any specific cognitive disadvantages of aphantasia, especially in the context of memory recall and comparing current and past experiences.

To understand the nature of subjective experiences in psychological terms, one must understand their functions. We have outlined here how higher-order theories might be extended and further tested to meet this challenge. We hope that fruitful research avenues can be opened up by integrating consciousness studies into the burgeoning literature of the cognitive and computational neurosciences of memory, mental imagery, emotions and decision-making.

Hakwan Lau[1]✉, *Matthias Michel*[2], *Joseph E. LeDoux*[3] *and Stephen M. Fleming*[4,5,6]

[1]*Laboratory for Consciousness, RIKEN Center for Brain Science, Wako, Japan.*

[2]*Center for Mind, Brain and Consciousness, New York University, New York, NY, USA.*

[3]*Center for Neural Science and Department of Psychology, New York University, New York, NY, USA.*

[4]*Department of Experimental Psychology, University College London, London, UK.*

[5]*Wellcome Centre for Human Neuroimaging, University College London, London, UK.*

[6]*Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK.*

✉*e-mail: hakwan.lau@riken.jp*

https://doi.org/10.1038/s44159-022-00068-6

Published online: 01 June 2022

1. Rosenthal, D. How to think about mental qualities. *Phil. Issues* **20**, 368–393 (2010).
2. Clark, A. *A Theory of Sentience* (Oxford Univ. Press, 2000).
3. Sellars, W. *Science, Perception and Reality* (Humanities Press, 1963).
4. Squire, L. R. *Memory and Brain* vol. 315 (Oxford Univ. Press, 1987).
5. Schacter, D. L., Buckner, R. L. & Koutstaal, W. Memory, consciousness and neuroimaging. *Phil. Trans. R. Soc. Lond. B* **353**, 1861–1878 (1998).
6. Tulving, E. in *The Missing Link In Cognition: Origins Of Self-reflective Consciousness* Vol. 364 (ed. Terrace, H. S.) 3–56 (Oxford Univ. Press, 2005).
7. LeDoux, J. E. & Lau, H. Seeing consciousness through the lens of memory. *Curr. Biol.* **30**, R1018–R1022 (2020).
8. Bruner, J. S. & Leigh Minturn, A. Perceptual identification and perceptual organization. *J. Gen. Psychol.* **53**, 21–28 (1955).
9. Allport, F. H. Theories of perception and the concept of structure: a review and critical analysis with an introduction to a dynamic-structural theory of behavior. *Optom. Vis. Sci.* **33**, 216 (1955).
10. Gregory, R. L. *Concepts And Mechanisms Of Perception* (Charles Scribner's Sons, 1974).
11. Graham, D. J., Friedenberg, J. D., Rockmore, D. N. & Field, D. J. Mapping the similarity space of paintings: image statistics and visual perception. *Vis. Cogn.* **18**, 559–573 (2010).
12. Summerfield, C. & de Lange, F. P. Expectation in perceptual decision making: neural and computational mechanisms. *Nat. Rev. Neurosci.* **15**, 745–756 (2014).
13. Murray, E. A., Wise, S. P. & Graham, K. S. *The Evolution of Memory Systems: Ancestors, Anatomy, and Adaptations* (Oxford Univ. Press, 2017).
14. Lamy, D., Carmel, T. & Peremen, Z. Prior conscious experience enhances conscious perception but does not affect response priming. *Cognition* **160**, 62–81 (2017).
15. Marks, L. E. et al. Magnitude-matching: the measurement of taste and smell. *Chem. Senses* **13**, 63–87 (1988).
16. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14**, 481–487 (2004).
17. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
18. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020).
19. Rosca, M., Weber, T., Gretton, A. & Mohamed, S. A case for new neural network smoothness constraints. Preprint at *arXiv* https://arxiv.org/abs/2012.07969 (2020).
20. Jin, P., Lu, L., Tang, Y. & Karniadakis, G. E. Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness. *Neural Netw.* **130**, 85–99 (2020).
21. Bauer, A.-K. R., Debener, S. & Nobre, A. C. Synchronisation of neural oscillations and cross-modal influences. *Trends Cogn. Sci.* **24**, 481–495 (2020).
22. Thoen, H. H., How, M. J., Chiou, T.-H. & Marshall, J. A different form of color vision in mantis shrimp. *Science* **343**, 411–413 (2014).
23. Brown, R., Lau, H. & LeDoux, J. E. Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* **23**, 754–768 (2019).
24. Lau, H. & Rosenthal, D. The higher-order view does not require consciously self-directed introspection: response to Malach. *Trends Cogn. Sci.* **15**, 508–509 (2011).
25. Michel, M. & Lau, H. On the dangers of conflating strong and weak versions of a theory of consciousness. *PhiMiSci* https://doi.org/10.33735/phimisci.2020.II.54 (2020).
26. Lamme, V. A. F. Towards a true neural stance on consciousness. *Trends Cogn. Sci.* **10**, 494–501 (2006).
27. Lamme, V. A. F. Why visual attention and awareness are different. *Trends Cogn. Sci.* **7**, 12–18 (2003).
28. Malach, R. Conscious perception and the frontal lobes: comment on Lau and Rosenthal. *Trends Cogn. Sci.* **15**, 507 (2011).
29. Zeki, S. The disunity of consciousness. *Trends Cogn. Sci.* **7**, 214–218 (2003).
30. Zeki, S. Localization and globalization in conscious vision. *Annu. Rev. Neurosci.* **24**, 57–86 (2001).
31. Macknik, S. L. Visual masking approaches to visual awareness. *Prog. Brain Res.* **155**, 177–215 (2006).
32. Macknik, S. L. & Martinez-Conde, S. in *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 1165–1175 (MIT Press, 2009).
33. Malach, R. Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci. Conscious.* **2021**, niab028 (2021).
34. Kouider, S. & Dehaene, S. Levels of processing during non-conscious perception: a critical review of visual masking. *Phil. Trans. R. Soc. Lond. B* **362**, 857–875 (2007).
35. Mazzi, C., Savazzi, S. & Silvanto, J. On the 'blindness' of blindsight: what is the evidence for phenomenal awareness in the absence of primary visual cortex (V1)? *Neuropsychologia* **128**, 103–108 (2019).
36. Silvanto, J., Cowey, A., Lavie, N. & Walsh, V. Making the blindsighted see. *Neuropsychologia* **45**, 3346–3350 (2007).
37. Weil, R. S., Plant, G. T., James-Galton, M. & Rees, G. Neural correlates of hemianopic completion across the vertical meridian. *Neuropsychologia* **47**, 457–464 (2009).
38. Weiskrantz, L. Prime-sight and blindsight. *Conscious. Cogn.* **11**, 568–581 (2002).
39. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
40. Dehaene, S. *Consciousness And The Brain: Deciphering How the Brain Codes Our Thoughts* (Penguin, 2014).
41. Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1988).
42. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).
43. Pounder, Z., Jacob, J., Evans, S., Loveday, C., Eardley, A. F., & Silvanto, J. Only minimal differences between individuals with congenital aphantasia and those with typical imagery on neuropsychological tasks that involve imagery. *Cortex* **148**, 180–192 (2022).
44. Keogh, R. & Pearson, J. The blind mind: no sensory visual imagery in aphantasia. *Cortex* **105**, 53–60 (2018).
45. Kay, L., Keogh, R., Andrillion, T. & Pearson, J. The eyes have it: the pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength. *Elife* **11**, e72484 (2022).
46. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).
47. Lau, H. Volition and the functions of consciousness. *Neurosci. Res.* **65**, S28 (2009).
48. Rosenthal, D. *Consciousness And Mind* (Clarendon, 2005).
49. Lycan, W. in *The Stanford Encyclopedia of Philosophy* (Stanford University, 2019).
50. Johnson, M. K. & Raye, C. L. Reality monitoring. *Psychol. Rev.* **88**, 67–85 (1981).
51. Johnson, M. K. Reality monitoring: an experimental phenomenological approach. *J. Exp. Psychol. Gen.* **117**, 390–394 (1988).
52. Simons, J. S., Garrison, J. R. & Johnson, M. K. Brain mechanisms of reality monitoring. *Trends Cogn. Sci.* **21**, 462–473 (2017).
53. Lau, H. Consciousness, metacognition, & perceptual reality monitoring. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/ckbyf (2019).
54. McCurdy, L. Y. et al. Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J. Neurosci.* **33**, 1897–1906 (2013).
55. Morales, J., Lau, H. & Fleming, S. M. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**, 3534–3546 (2018).
56. Garrison, J. R. et al. Testing continuum models of psychosis: no reduction in source monitoring ability in healthy individuals prone to auditory hallucinations. *Cortex* **91**, 197–207 (2017).
57. Dijkstra, N., Mazor, M., Kok, P. & Fleming, S. Mistaking imagination for reality: congruent mental imagery leads to more liberal perceptual detection. *Cognition* **212**, 104719 (2021).
58. Voss, U. et al. Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nat. Neurosci.* **17**, 810–812 (2014).
59. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).
60. Ye, Q., Zou, F., Lau, H., Hu, Y. & Kwok, S. C. Causal evidence for mnemonic metacognition in human precuneus. *J. Neurosci.* **38**, 6379–6387 (2018).
61. Miyamoto, K. et al. Causal neural network of metamemory for retrospection in primates. *Science* **355**, 188–193 (2017).
62. Cleeremans, A. et al. Learning to be conscious. *Trends Cogn. Sci.* **24**, 112–123 (2020).
63. Fleming, S. M. Awareness as inference in a higher-order state space. *Neurosci. Conscious* **2020**, niz020 (2020).
64. Whyte, C. J. & Smith, R. The predictive global neuronal workspace: a formal active inference model of visual consciousness. *Prog. Neurobiol.* **199**, 101918 (2021).
65. Koenig-Robert, R. & Pearson, J. Why do imagery and perception look and feel so different? *Phil. Trans. R. Soc. Lond. B* **376**, 20190703 (2021).
66. Muzur, A., Pace-Schott, E. F. & Hobson, J. A. The prefrontal cortex in sleep. *Trends Cogn. Sci.* **6**, 475–481 (2002).
67. Zmigrod, L., Garrison, J. R., Carr, J. & Simons, J. S. The neural mechanisms of hallucinations: a quantitative meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* **69**, 113–123 (2016).
68. Narayanan, N. S., Rodnitzky, R. L. & Uc, E. Y. Prefrontal dopamine signaling and cognitive symptoms of Parkinson's disease. *Rev. Neurosci.* **24**, 267–278 (2013).

69. Fazekas, P. & Nemeth, G. Dream experiences and the neural correlates of perceptual consciousness and cognitive access. *Phil. Trans. R. Soc. Lond. B* **373**, 20170356 (2018).
70. Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat. Commun.* **8**, 15471 (2017).
71. Gershman, S. J. The generative adversarial brain. *Front. Artif. Intell. Appl.* **2**, 18 (2019).
72. Barceló, F., Suwazono, S. & Knight, R. T. Prefrontal modulation of visual processing in humans. *Nat. Neurosci.* **3**, 399–403 (2000).
73. Bar, M. et al. Top-down facilitation of visual recognition. *Proc. Natl Acad. Sci. USA* **103**, 449–454 (2006).
74. Bar, M. A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* **15**, 600–609 (2003).
75. Tsushima, Y., Sasaki, Y. & Watanabe, T. Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* **314**, 1786–1788 (2006).
76. Kriete, T., Noelle, D. C., Cohen, J. D. & O'Reilly, R. C. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl Acad. Sci. USA* **110**, 16390–16395 (2013).
77. Held, R. & Hein, A. Movement-produced stimulation in the development of visually guided behavior. *J. Comp. Physiol. Psychol.* **56**, 872–876 (1963).
78. Hubel, D. H., Wiesel, T. N. & LeVay, S. Plasticity of ocular dominance columns in monkey striate cortex. *Phil. Trans. R. Soc. Lond. B* **278**, 377–409 (1977).
79. King, A. J., Hutchings, M. E., Moore, D. R. & Blakemore, C. Developmental plasticity in the visual and auditory representations in the mammalian superior colliculus. *Nature* **332**, 73–76 (1988).
80. Weinberger, D. R. From neuropathology to neurodevelopment. *Lancet* **346**, 552–557 (1995).
81. Merzenich, M. M. & Sameshima, K. Cortical plasticity and memory. *Curr. Opin. Neurobiol.* **3**, 187–196 (1993).
82. Fiser, J. & Aslin, R. N. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* **12**, 499–504 (2001).
83. Recanzone, G. H. Rapidly induced auditory plasticity: the ventriloquism aftereffect. *Proc. Natl Acad. Sci. USA* **95**, 869–875 (1998).
84. Moser, E. I., Kropff, E. & Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* **31**, 69–89 (2008).
85. O'Keefe, J. & Burgess, N. Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus* **15**, 853–866 (2005).
86. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
87. Qasim, S. E., Fried, I. & Jacobs, J. Phase precession in the human hippocampus and entorhinal cortex. *Cell* **184**, 3242–3255.e10 (2021).
88. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
89. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: spatial codes for human thinking. *Science* **362**, eaat6766 (2018).
90. Nau, M., Navarro Schröder, T., Bellmund, J. L. S. & Doeller, C. F. Hexadirectional coding of visual space in human entorhinal cortex. *Nat. Neurosci.* **21**, 188–190 (2018).
91. Bellmund, J. L., Deuker, L., Navarro Schröder, T. & Doeller, C. F. Grid-cell representations in mental simulation. *eLife* **5**, e17089 (2016).
92. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
93. Mark, S., Moran, R., Parr, T., Kennerley, S. W. & Behrens, T. E. J. Transferring structural knowledge across cognitive maps in humans and models. *Nat. Commun.* **11**, 4783 (2020).
94. Aronov, D., Nevers, R. & Tank, D. W. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543**, 719–722 (2017).
95. Bao, X. et al. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron* **102**, 1066–1075.e5 (2019).
96. LeDoux, J. *The Deep History of Ourselves: the Four-Billion-Year Story of How We Got Conscious Brains* (Penguin, 2019).
97. LeDoux, J. E. Thoughtful feelings. *Curr. Biol.* **30**, R619–R623 (2020).
98. Johnson-Laird, P. N. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Harvard Univ. Press, 1983).
99. LeDoux, J. E. What emotions might be like in other animals. *Curr. Biol.* **31**, R824–R829 (2021).
100. Passingham, R. E., Bengtsson, S. L. & Lau, H. C. Medial frontal cortex: from self-generated action to reflection on one's own performance. *Trends Cogn. Sci.* **14**, 16–21 (2010).
101. Wagner, D. D., Haxby, J. V. & Heatherton, T. F. The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdiscip. Rev. Cogn. Sci.* **3**, 451–470 (2012).
102. Sutherland, K. & Bryant, R. A. Autobiographical memory and the self-memory system in posttraumatic stress disorder. *J. Anxiety Disord.* **22**, 555–560 (2008).
103. Olsson, A., Knapska, E. & Lindström, B. The neural and computational systems of social learning. *Nat. Rev. Neurosci.* **21**, 197–212 (2020).
104. Isoda, M. The role of the medial prefrontal cortex in moderating neural representations of self and other in primates. *Annu. Rev. Neurosci.* **44**, 295–313 (2021).
105. Ritchey, M., Libby, L. A. & Ranganath, C. Cortico-hippocampal systems involved in memory and cognition: the PMAT framework. *Prog. Brain Res.* **219**, 45–64 (2015).
106. Gilboa, A. & Marlatte, H. Neurobiology of schemas and schema-mediated memory. *Trends Cogn. Sci.* **21**, 618–631 (2017).
107. Damasio, A. & Carvalho, G. B. The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **14**, 143–152 (2013).
108. Yeterian, E. H., Pandya, D. N., Tomaiuolo, F. & Petrides, M. The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex* **48**, 58–81 (2012).
109. Curtis, C. E. Prefrontal and parietal contributions to spatial working memory. *Neuroscience* **139**, 173–180 (2006).
110. Vilberg, K. L. & Rugg, M. D. Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia* **46**, 1787–1799 (2008).
111. Cabeza, R., Ciaramelli, E., Olson, I. R. & Moscovitch, M. The parietal cortex and episodic memory: an attentional account. *Nat. Rev. Neurosci.* **9**, 613–625 (2008).
112. Fischer, M., Moscovitch, M. & Alain, C. A systematic review and meta-analysis of memory-guided attention: frontal and parietal activation suggests involvement of fronto-parietal networks. *Wiley Interdiscip. Rev. Cogn. Sci.* **12**, e1546 (2021).
113. Wagner, A. D., Shannon, B. J., Kahn, I. & Buckner, R. L. Parietal lobe contributions to episodic memory retrieval. *Trends Cogn. Sci.* **9**, 445–453 (2005).
114. Berryhill, M. E., Picasso, L., Arnold, R., Drowos, D. & Olson, I. R. Similarities and differences between parietal and frontal patients in autobiographical and constructed experience tasks. *Neuropsychologia* **48**, 1385–1393 (2010).
115. Katsuki, F. & Constantinidis, C. Unique and shared roles of the posterior parietal and dorsolateral prefrontal cortex in cognitive functions. *Front. Integr. Neurosci.* **6**, 17 (2012).
116. Berryhill, M. E., Phuong, L., Picasso, L., Cabeza, R. & Olson, I. R. Parietal lobe and episodic memory: bilateral damage causes impaired free recall of autobiographical memory. *J. Neurosci.* **27**, 14415–14423 (2007).
117. Tanaka, K. Z. & McHugh, T. J. The hippocampal engram as a memory index. *J. Exp. Neurosci.* **12**, 1179069518815942 (2018).
118. Teyler, T. J. & Rudy, J. W. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus* **17**, 1158–1169 (2007).
119. Wittkuhn, L., Chien, S., Hall-McMaster, S. & Schuck, N. W. Replay in minds and machines. *Neurosci. Biobehav. Rev*. https://doi.org/10.1016/j.neubiorev.2021.08.002 (2021).
120. Endo, K., Tsuchimoto, Y. & Kazama, H. Synthesis of conserved odor object representations in a random, divergent-convergent network. *Neuron* **108**, 367–381.e5 (2020).
121. Pashkovski, S. L. et al. Structure and flexibility in cortical representations of odour space. *Nature* **583**, 253–258 (2020).
122. Dasgupta, S., Stevens, C. F. & Navlakha, S. A neural algorithm for a fundamental computing problem. *Science* **358**, 793–796 (2017).
123. Dasgupta, S., Sheehan, T. C., Stevens, C. F. & Navlakha, S. A neural data structure for novelty detection. *Proc. Natl Acad. Sci. USA* **115**, 13093–13098 (2018).
124. Dunsmoor, J. E. & Murphy, G. L. Categories, concepts, and conditioning: how humans generalize fear. *Trends Cogn. Sci.* **19**, 73–77 (2015).
125. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
126. Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
127. Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018).
128. Schacter, D. L. & Addis, D. R. On the constructive episodic simulation of past and future events. *Behav. Brain Sci.* **30**, 331–332 (2007).
129. Schacter, D. L. Constructive memory: past and future. *Dialogues Clin. Neurosci.* **14**, 7–18 (2012).
130. Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L. & Platt, M. L. A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology* **42**, 1931–1939 (2017).
131. Mazor, M., Friston, K. J. & Fleming, S. M. Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *eLife* **9**, e53900 (2020).
132. Miyamoto, K., Setsuie, R., Osada, T. & Miyashita, Y. Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. *Neuron* **97**, 980–989.e6 (2018).
133. Martino, B. D., De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
134. Donoso, M., Collins, A. G. E. & Koechlin, E. Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481–1486 (2014).
135. Morales, J. & Lau, H. in *Qualitative Consciousness: Themes from the Philosophy of David Rosenthal* (ed. Weisberg, J.) (Cambridge Univ. Press, 2021).
136. Gherman, S. & Philiastides, M. G. Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife* **7**, e38293 (2018).
137. Bang, D. & Fleming, S. M. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc. Natl Acad. Sci. USA* **115**, 6082–6087 (2018).
138. Wittmann, M. K. et al. Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91**, 482–493 (2016).

Acknowledgements
S.M.F. is funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society.

Author contributions
H.L. led the writing and all other authors contributed equally to the remaining aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Psychology thanks Peter Fazekas, Rafael Malach and the other, anonymous, reviewer for their contribution to the peer review of this work.