Check for updates

Capturing the objects of vision with neural networks

Benjamin Peters [1] ✉ and Nikolaus Kriegeskorte [1,2,3,4] ✉

Human visual perception carves a scene at its physical joints, decomposing the world into objects, which are selectively attended, tracked and predicted as we engage our surroundings. Object representations emancipate perception from the sensory input, enabling us to keep in mind that which is out of sight and to use perceptual content as a basis for action and symbolic cognition. Human behavioural studies have documented how object representations emerge through grouping, amodal completion, proto-objects and object files. By contrast, deep neural network models of visual object recognition remain largely tethered to sensory input, despite achieving human-level performance at labelling objects. Here, we review related work in both fields and examine how these fields can help each other. The cognitive literature provides a starting point for the development of new experimental tasks that reveal mechanisms of human object perception and serve as benchmarks driving the development of deep neural network models that will put the object into object recognition.

Vision gives us a rapid sense of our surroundings that exceeds the information in the retinal image and provides a structured understanding of the scene. The structure imposed on the basis of prior knowledge is central to perception as an inference process[1,2] and to a causal and compositional understanding that enables us to consider counterfactuals and act intelligently[3]. The basic building blocks of our perceptual representation are objects. Our percepts include parts of objects that are occluded by other objects or behind us. Out of sight, for a mature primate, is not out of mind[4]. Relevant objects that become invisible remain represented, a memory trace, and may even be animated in our minds according to a rough approximation of the laws they obey in the world.

Human behavioural researchers have quantitatively investigated these phenomena using a wide range of ingenious experimental paradigms. They have condensed the insights gained from the data in cognitive theories, which describe separate mechanisms for seeing 'stuff'[5] and seeing 'things'[6]. Stuff has come to refer to parts of the visual scene represented in terms of summary statistics[7–9] that capture textures, materials and perhaps categories at an aggregate level. Things are the objects that our brains pick out for individuated representation. An object representation may explicitly bind together the parts of each object and the image features that each part accounts for[10]. An object's missing information may be filled in by inference using prior information[11]. Cognitive scientists have described how bottom-up and top-down processes interactively determine the formation of a limited number of object representations that are accessible to higher cognition[12].

The object representations may have a life of their own, simulating trajectories and interactions among objects to predict the future. Short of foreseeing the future, even being on time in representing the present requires prediction to compensate for signalling delays in the nervous system. The perceived world emerges from the confluence in the inference process of prior information and present sensory signals[13,14]. Our brains combine past experiences over multiple time scales to best predict the present and the future[1,2,15,16].

Cognitive scientists want to understand these dynamic and constructive inferences and the representations of objects in the human mind. Object representations abstract from the sensory features and cast the world as a composition of entities that can be acted on and named. This places object representations at the nexus of perception, action and symbolic cognition (Fig. 1).

Engineers may not be interested in modelling the human mind. However, engineering also benefits from models that have concepts of objects because they promise, for example, to enable a robot to understand the structure of the world, and to reason, plan and act on this basis. For humans and machines alike, decomposing the world into objects may facilitate the modular reuse of learned knowledge and simplify complex inferences. An object-based representation provides a radical abstraction from the stream of sensory signals, a predictable scaffold of reality and a basis for causal understanding. Building models with object-based representations is therefore a crucial challenge for engineering[17,18] as well as for cognitive science.

Parsing the world into objects requires an operational definition. What is an object? A key criterion is physical cohesion[19]. As written previously[20], "If you want to know what an object is, just 'grab some and pull'; the stuff that comes with your hand is the object.". This operational definition grounds objects in the physical structure of the world. Sensorimotor interactions, such as grabbing and pulling, may help us to acquire the perceptual ability to parse the world into objects in early development[4]. They also continue to serve us in maturity, enabling us to confirm, through direct experimentation, our perception that something is an object. The operational 'what if' nature of this definition reveals that objects are rooted in a causal understanding of physical reality[3].

Object-based representations carve the scene at its physical joints. Reducing a million retinal signals to a few behaviourally relevant objects requires prior knowledge of the physical world, prior percepts from the present scene and selection of what is relevant in light of the current behavioural goals. The present sensory evidence does not solely determine the percept; it is just one of a number of
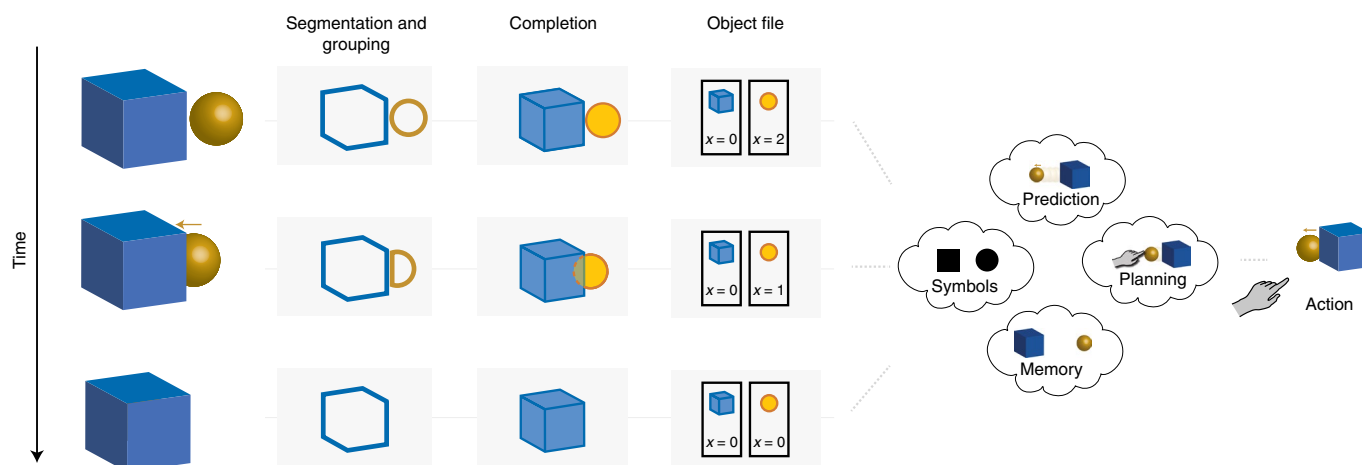
Fig. 1 | Stages of untethering human visual object perception from the sensorium. As the golden ball moves behind the blue box (left; top to bottom), it is first unoccluded, then partially occluded and finally fully occluded. It remains represented at the level of its object file even when fully invisible. The initial segmentation parses the scene into groups of features, each corresponding to one of the objects. Amodal completion may occur for partially occluded objects, completing the invisible portion of the object on the basis of short-term or long-term memory of its shape. A subset of the objects may be encoded in a non-retinotopic object-based representation (for example, object files). Object files can sustain information about the presence and properties of objects across temporary occlusions, untethering the object representations from the sensorium. Untethered object representations can be considered to be an interface between perception and symbolic thought, prediction, mental planning and action.

constraints. Thus, object representations untether and emancipate perception from the stream of sensory signals.

Engineering has made substantial inroads towards this type of dynamic and constructive perceptual inference. The integration of sensory data over multiple timescales is captured by the Bayes filter, a recurrent mechanism that stores a compressed representation of recent experience for optimal representation of the present moment[21]. Recurrent neural networks (RNNs) provide a universal model class for such inferences that can implement Bayes filters[22]. However, getting RNNs to perform this kind of inference for natural dynamic vision (video) remains challenging. Computer vision therefore heavily relies on feedforward convolutional neural network (CNN) models, which analyse each frame separately through a hierarchy of nonlinear transformations[23,24]. Feedforward deep CNNs can learn static mappings from images to category labels or structural descriptions of the scene. However, the representations in these models remain tethered to the input and lack any concept of an object. They represent things as stuff[25]. They cannot combine information over time such as to condition current perceptual inferences on past observations. They may also not be ideal for parsing scenes into objects. These limitations may explain why the performance of feedforward convolutional networks is brittle, breaking down when the models must generalize across domains[26]. The models lack what humans have—a generative structural and causal understanding of the world—to stabilize their perception[27–29].

A generative mental model is a model of the process that generates the sensory data. A mind that uses a generative model is challenged to comprehensively explain all aspects of the sensory data, rather than taking a shortcut and selectively extracting only behaviourally relevant information[30]. In the context of a generative model that captures our prior assumptions about the world, perception can be conceptualized as inference[1]. Probabilistic inference provides a normative perspective on how perception should work to make optimal use of limited sensory data. In particular, human vision is often conceptualized as an approximation to probabilistic inference on a generative model[2,16,31]. However, given limited neural hardware and compute time, it is difficult to implement the normative ideal. The cognitive theories and neural network mechanisms

that we review here can be understood as heuristic approximations to inference on a generative model.

Cognitive scientists and engineers have begun building models that can maintain an internal state and dynamically map the sensory input to internal object representations that have their own persistence and dynamics. Brains and models must decide what qualifies two bits of the visual image to be grouped together as parts of the same object[32,33]. Containment within a closed contour and persistence over time of shape, colour and motion are key factors that determine how humans segment a scene into objects[19,20]. These factors are encapsulated by the more general notion of spatiotemporal contiguity, which provides evidence for an underlying physical property—cohesion. But how are the sensory indications of spatiotemporal contiguity combined and their conflicts resolved? How are the object representations untethered from the sensorium, and made to persist when the object disappears behind an occluder? How are they animated jointly by sensory data and generative models of the world? These remain computational mysteries of the human mind and brain.

The focus of this Review is on the general computational mechanisms of object-based representations, which are generative and recurrent and complementary to the discriminative feedforward mechanism underlying the initial sweep of activity through the visual hierarchy. We describe these mechanisms in the context of generic rigid bodies. However, these general mechanisms could be replicated in the brain in domain-specific modules that are adapted to the particular properties of behaviourally important objects. Similar to the feedforward mechanisms that learn the appearance of objects in different domains (such as faces, people, animals, buildings, food and tools), the object-based mechanisms will also adapt to the behaviour of the objects, including their ways of moving (for example, facial expressions), their rigidity (for example, for rocks and buildings) or articulation (as for bodies and tools), their interactions with other objects (be it according to the laws of classical mechanics or theory of mind) and their behavioural relevance.

We first review behavioural phenomena and cognitive theories of human object representations, and then the current state of neural network modelling. Our goals are to highlight parallels between cognitive concepts and neural network model mechanisms

and to discern what characteristics of human object representations are missing in current neural network models. We hope that this Review will (1) help modellers to understand the behavioural literature, (2) help behavioural researchers to understand the computational literature and (3) help both groups to develop tasks that can serve simultaneously as probes of human cognition and as benchmarks for computational models.

Cognitive theories

Cognitive scientists have explored object vision with behavioural experiments, and their concepts and theories summarize the insights gained (Fig. 1). Grouping of visual features and amodal completion yield a rapid initial scene segmentation that transcends the static filters of the feedforward visual hierarchy, but remains tethered to the retinal reference frame. This retinotopic representation forms the basis for selection of a limited set of objects for representation in an object-based reference frame, known as object files or slots. At this level, object representations are untethered from the retinal reference frame and may enter central cognition[34–36] and interaction with other cognitive systems[37,38]. To date, the cognitive concepts that we review here lack full mechanistic specification. However, they help to summarize the behavioural phenomena, decomposing the cognitive processes and providing essential stepping stones towards their implementation in neural network models.

Tethered to the retinal reference frame: pixels to proto-objects. *Grouping features.* The simplest way to combine evidence over space is using static filter templates. This is the mechanism of models of V1 simple and complex cell responses[39]. A hierarchy of such filters[40] yields texture statistics at different spatial scales, as used in convolutional feedforward neural networks[23]. However, there is evidence that the visual system also uses lateral recurrent signal flow to relate collinear edges[41–44]. Dynamic recurrent processing through lateral interactions may provide a more flexible mechanism for grouping features at larger scales. Imagine, for example, the set of all smooth closed contours. The combinatorics of feature configurations forming a smooth closed contour may render the representation of this set with a basis of static filters unrealistic. However, the regularity of smooth continuation can be exploited by a model using lateral recurrent connectivity.

Principles of perceptual grouping were first identified by Gestalt researchers[45–47], who noted that people perceive visual elements as grouped by principles including continuity, proximity, similarity, closure, prägnanz and common fate. One of these principles, continuity, involves the detection and integration of contour elements[42], and the computation of border-ownership for the creation of surface representations[48]. Feedforward[49] as well as recurrent operations[50,51] that incrementally group contours by spread of activation[52] have been proposed. Perceptual grouping is influenced by several factors, such as binocular disparity[53], textures[7], temporal coincidence[54] and knowledge about object appearances[55].

Local integration processes may give rise to a mosaic stage[56], in which each connected set of visible parts of an object forms a group. The mosaic stage is similar to Marr's[57] full primal sketch, in which contour integration gives rise to an initial grouping. In Marr's theory, the primal sketch is followed by the 2.5-dimensional sketch, which represents the visible portions of objects as surfaces and assigns a depth to each patch of the image. Once surfaces and depth relationships are represented in the 2.5-dimensional sketch, the visual system can infer how objects may extend behind occluders. Disjoint mosaic pieces belonging to the same object (disconnected by occlusion) can be grouped together and the occluded parts filled in.

Amodal completion. Visual scenes often contain objects that are partially occluded by other objects. Moreover, objects always occlude their own back sides. We nevertheless perceive them as three-dimensional wholes. It has been proposed that this subjective experience might result from a process that explicitly fills in the missing parts of an object in our mental representation. The process has been called amodal completion[58] because, in contrast to perceptual filling-in (that is, modal completion)[59], it transcends the sensory modality—the occluded part or back side of an object is not visually perceived, yet it is part of the percept.

Beyond the phenomenology of subjective experience, the hypothesis of an amodal completion process suggests testable behavioural predictions. A partially occluded object should elicit priming effects that match those elicited by its complete form, rather than those elicited by its visible fragments (Box 1). This prediction has been confirmed in behavioural experiments[56]. Similar predictions have been confirmed for discrimination[60] and visual search tasks[61,62]. These studies have also shown that it takes time for amodal completion to emerge, suggesting that it relies on recurrent processing[56,60].

Amodal completion must rely on prior knowledge. It could use general knowledge about the statistics of images (for example, the knowledge that edges tend to extend smoothly) or about the shape of objects (for example, the knowledge of the shape of an occluded part of a letter). It could also rely on knowledge gleaned moments earlier from having observed the now occluded parts of the object. There is evidence that amodal completion extends edges behind occluders if a continuous smooth connection exists[63]. Amodal completion is also thought to fill in missing parts of surfaces[61] and volumes[64]. Local completion extends and connects object contours mostly linearly according to the Gestalt principle of good continuation (Fig. 2b). Global completion refers to completion that prefers symmetric solutions[65] (Fig. 2c) probably occurring in higher visual areas such as the lateral occipital complex[66,67]. More generally, the term perceptual closure[68,69] refers to completion based on prior knowledge about the shape or appearance of an object (Fig. 2d).

Amodal completion may best be construed as an inference process—the visual system's best guess about the missing part, given the current evidence and prior knowledge. The computational function of making the inferred information explicit might be to support further inferences about the object.

Proto-objects. The initial input segmentation occurs in parallel and pre-attentively across the visual field[35,70]. These processes are largely independent of conscious cognition, in the sense that our conscious thoughts cannot penetrate and interfere with them[71]. For example, consciously thinking that the horse pattern in Fig. 2e should extend regularly behind the occluder does not prevent the visual system from generating the percept of an elongated horse.

These initial segmentations are thought to be tethered to the retinal reference frame. As a consequence, they are subject to change whenever we move our eyes or the world evolves. Moreover, the grouping of features might not yet be definitely established at this early stage. It might be best understood as a set of tentative feature associations than a full parse of the scene into object representations[72]. Thus, these representations have been termed proto-objects[12], to acknowledge their volatile and tentative nature. Transforming a proto-object representation into a stable and spatiotemporally coherent object-based representation will require selection by higher cognitive processes and untethering from the retinal reference frame.

Untethered from the retinal reference frame: object files and pointers. To individuate objects and combine the distributed evidence about them, the visual system has to overcome a fundamental challenge—how to group the spatiotemporally disjoint pieces into a coherent object representation? In the retinal reference frame, the pieces had to be grouped in space. Now the grouping problem extends in space and time. Rather than segmenting retinal space, the system must carve out a 'space–time worm'[20] from the spatiotemporal input (Fig. 2f)).

Box 1 | Probing human perception of untethered objects through tasks in behavioural experiments

Cognitive scientists have developed a variety of ingenious tasks to examine human perception of untethered objects with behavioural experiments. **a–d**, Grouping tasks. Four different tasks for contour integration and grouping. **a**, Decide as fast as possible whether two dots lie on the same or different lines[265]. **b**, Decide whether the dot lies inside a closed contour[264]. **c**, Decide whether both red dots lie on the same object[291]. **d**, Detect the direction of the horizontal offset between the central vertical lines in the presence of flankers. The task is more difficult if the flankers are also isolated (crowding, left) and is easier if the flankers are part of a coherent object (uncrowding, right)[241]. **e**, Amodal completion. A partially occluded shape (here, a circle) is presented as a prime. Subsequently, the participants are presented with two shapes and have to decide whether they are identical[56]. Responses are faster if these shapes match the percept of the prime (for example, the circles if the percept was amodally completed). **f**, Object-reviewing paradigm. In a typical object-reviewing trial[85], two objects containing a letter are presented during the previewing display. In the test display, only one letter is presented and needs to be identified. Reactions are faster if the letter is in the same object as in the previewing frame. Here, the objects also switch positions. **g**, Object-based attention. In the object-based attention task[100], one end of one object is briefly flashed to attract attention to this position. After a brief delay, the participants have to react as quickly as possible to a target (red dot). Reactions are faster when the target appears in the same object (top) as the flash compared with when the target appears in the other object (bottom). **h**, Multiple-object tracking. A set of targets is flashed initially and has to be tracked among identical distractors. After the tracking phase, participants have to select the identity of the tracked targets[91]. **i**, Violation of expectation. Violation of expectation to study object permanence and physical reasoning. Here, a solid ball disappears behind a wall that subsequently folds down. The observer's surprise is measured (for example, by measuring the looking time) in response to this physically impossible sequence of events[107]. **j**, In the block–copy task, participants have to reconstruct a model visual pattern in a workspace area using building blocks from the resource area[266].
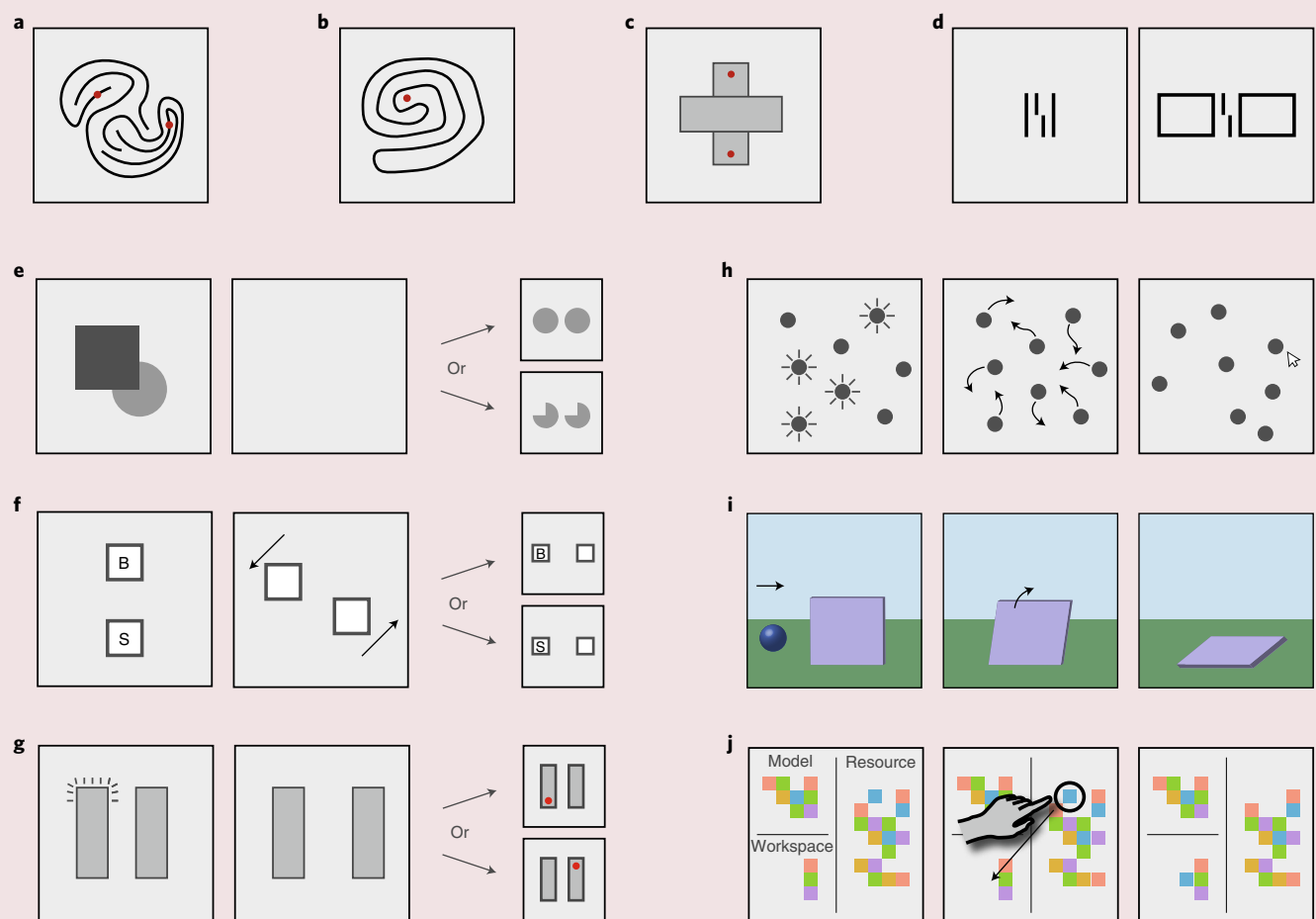
How does the visual system link distinct sensory inputs across occlusions or saccades to a single object-centred representation? In many situations, this correspondence problem[73] is solved by assessing the spatiotemporal continuity of objects[20,74,75]. A notable example is the 'tunnel effect'[76]. An object that moves behind an occluder and reappears with a different appearance (such as a different colour or even category) may still be considered to be the same object by the visual system instead of two different ones[11,77]. A single object is more likely to be perceived if the pre-occlusion stimulus is similar to the post-occlusion stimulus[78,79], suggesting a general mechanism that flexibly weighs object feature dimensions to infer correspondence[80]. If correspondence is inferred, we perceive a single object of which the appearance combines pre- and post-occlusion sensory signals. The post-occlusion appearance of the object is biased
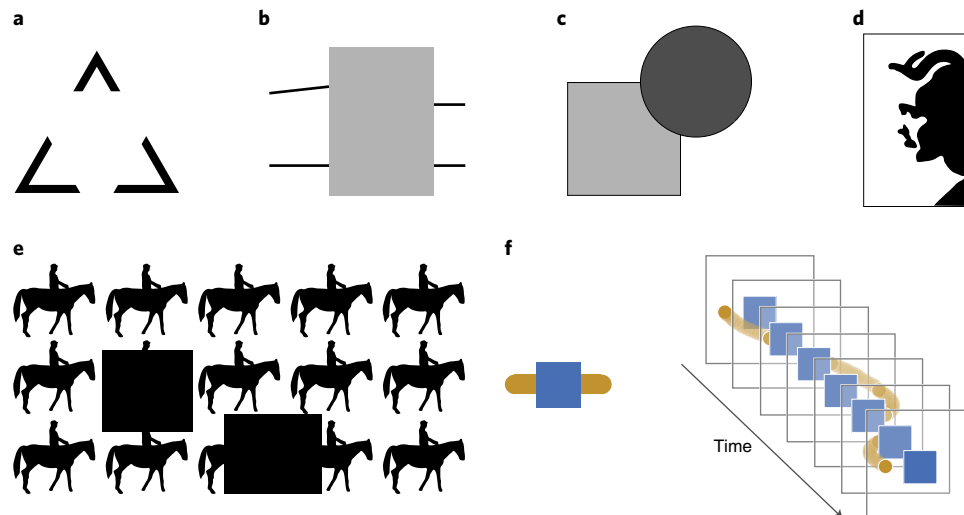
Fig. 2 | Completion phenomena. a, There appears to be a solid white triangle occluding the black contours of another triangle. The percept of the occluding white triangle is an example of modal completion, because the inferred contours appear as though they were present in the visual modality. The percept of the occluded black triangle is an example of amodal completion, because the missing black contours are perceived to exist, but are not visually perceived[283]. **b**, The bottom black line segments appear to be connected behind the grey box. This is an example of amodal completion because the inferred continuation is not perceived as visible in the image. **c**, A complete grey square appears to be present. This is an example of amodal completion on the basis of global shape cues. **d**, We perceive a face lit from the right. This is an example of perceptual closure[68]. **e**, People may perceive a giraffe-like rider (top left black box) or an elongated horse (bottom right black box)[284]. These percepts are inconsistent with both the global repetitive pattern and our prior knowledge about the anatomy of horses and people. Such illusions demonstrate that local cues can override global cues and prior knowledge in the perceptual inference process. **f**, We perceive a single golden object extending behind the blue occluder (left). This is an example of amodal completion that requires grouping of all of the golden bits across space. Perceptual inference can also group bits of visual evidence across space and time simultaneously. Right, frames of a video in which a golden ball oscillates behind a blue occluder. When watching such a video, we perceive a persistent object, the presence of which continues across periods of total invisibility. Our visual system groups the golden bits into a space–time worm. This is an example of spatiotemporal amodal completion.

towards the pre-occlusion stimulus[81,82]. Similarly, eye-movement studies[83] suggest that both the locations and appearances of stimuli are used to establish correspondences across saccades[84].

Correspondence computations support stable internal representations of individuated, untethered object representations that transcend the retinal or spatial reference frame. Different cognitive theories have been proposed that encapsulate empirical findings of how object representations might interact with the retinal bound proto-object representational level[12,85,86]. These theories emphasize the importance of space over other features to individuate and keep track of objects. Different objects tend not to occupy the same portion of space simultaneously. The natural domain to uniquely track objects across time is therefore the spatial domain. Feature integration theory suggests that segregation of the input into objects and binding of object features to coherent representations occurs through space[70]. Pylyshyn[86] proposed an indexing system that individuates and tracks objects through spatial pointers or indices. Although visual indexes are pointers to locations, they themselves encode no object properties. Thus, Pylyshyn termed his theory fingers of instantiation (FINST), as indices work like physical fingers; without knowing anything about the tracked (pointed to) object, spatial information such as a location or spatial relations between different fingers can be extracted.

Similarly, Kahneman et al.[85] proposed that our visual system individuates each object by creating an object file that groups a subset of the proto-objects carved out in the retinal reference frame on the basis of spatiotemporal factors. In contrast to visual indices, object files are thought to also store information about the properties of the object (such as colour and shape), therefore rerepresenting and binding essential sensory information in a coherent object representation[85]. This process is termed identification because the

feature information defines the identity of each object. Evidence for separate processing of object features bound into a coherent object representation comes from studies in which humans perceive illusory conjunctions of features of two different objects[72] under some conditions, demonstrating the failure of the process. The individuation of an object is thought to precede the identification of its appearance, as famously captured by the observation of Kahneman et al. (p. 217 of ref. [85]) that humans can conceive of something as the same thing while its identity remains in flux and might change substantially over time: "Onlookers in the movie can exclaim: 'It's a bird; it's a plane; it's Superman!' without any change of referent for the pronoun".

One of the hallmark features of human cognition is that the number of simultaneously maintained object files is highly limited. These capacity limitations are often phrased in terms of limited attentional resources. Spatiotopic maps may encode the distribution of attention over the visual field. These spatial attention maps[87] may be the access point of the spatial indexing system in which object files could be created from saliency peaks by centre-surround inhibition. Multiple object files can then each be tracked by top-down attention in the spatial attention map[88]. A mechanistic explanation for the capacity limitation of the object-file system is therefore surround inhibition[89] between spatial pointers in these maps[90]. One influential class of tasks that now has been used in hundreds of empirical studies is multiple-object tracking[91] (Box 1h). Humans can track a limited number of objects (perhaps three or four) even through full occlusions[91–94]. Subsequent research found that the tracking limitations can be better described by a flexible resource[95] that is independent across hemifields[88]. For example, if a slower object speed reduces spatial crowding, up to eight objects can be tracked[96].

The selection of an object for tracking entails a processing advantage for all of its elements and for the spatial positions it occupies[94,97]. This manifests in faster and more accurate detection of targets that appear on tracked compared with untracked objects. The processing advantage extends across the whole representation and suggests that objects are the fundamental units of attentional selection[98]. Object-based attention benefits both dynamic and static objects[34,99–102], objects that are only partially visible and completed amodally[103], and even objects that are completely invisible and retained in memory for a brief duration[104,105].

Object permanence, visual working memory and mental simulation. Objects can transiently cease to elicit retinal responses, for example when they become occluded and when we shift our gaze. However, internal object representations can remain stable even with their links to the input momentarily severed. The knowledge that out of sight is not out of mind has been termed object permanence by Piaget[4]. In infants, artificial stimuli that violate object permanence elicit longer looking times, consistent with surprise (violation of expectation; Box 1i). The results of such experiments support the idea that a kernel of object permanence may be either innate or established within 3 or 4 months after birth[106–108]. However, the ability to represent objects that are not currently in view probably matures over early development[109–111].

Adults can track objects through full occlusions without noticeable performance decrements[93]. This suggests a remarkable ability of our visual system to attribute spatiotemporally disjoint sensations to the same coherent object representation. An object representation can better track the sensory signals elicited by its object if it captures the dynamics of its object and predicts its future location and state[27]. Evidence for mental simulations of object dynamics comes from studies of representational momentum, which show that people incorrectly estimate the angle of a suddenly disappearing rotating object as slightly advanced along the rotational motion trajectory[112]. The mental simulations seem to be confined to first-order dynamics: humans appear to use velocity, but not acceleration, to simulate objects behind occluders[113,114]. From a normative perspective, prediction of the dynamics should be important for an object representation to track its object through longer periods of occlusion so as to find the sensory signals elicited by the object as it re-emerges. However, in most real-world scenarios that humans encounter, a coarse approximate prediction of the dynamics might potentially suffice to successfully track objects. Indeed, psychophysical evidence suggests that human perceptual inferences rely heavily on coarse spatiotemporal heuristics[115].

The fact that object representations can bridge occlusions implies that some information about the object is stored during occlusion. But what is the nature of this internal untethered representation? Another frequent event that momentarily severs the object representations from the sensorium is the saccade, during which input into the visual system is suppressed (saccadic suppression[116]). Asking people to detect changes in visual patterns across saccades reveals that their transsaccadic memory is capacity-limited and does not retain detailed spatial information but rather abstract and relational information[83,117].

The limits of human object representations are also evident in multiple-object tracking tasks. When the objects are suddenly occluded, people can recall location and velocity (including direction) information, but not the detailed identifying features of the objects[118,119]. In particular, shape and colour are difficult to consciously recall a moment later[120], although information about them (along with location and velocity) is maintained across occlusions[78,79].

These findings suggest that the human visual system does not maintain an object representation that fully specifies all its features. Instead—for the purpose of bridging disruption of the input as caused by saccades or occlusions—only a small subset of the features of an object is maintained.

A candidate system that can encode and maintain visual information for a limited amount of time during occlusions or saccadic remapping is visual working memory[77,121,122]. This system is severely limited in its capacity. Visual working memory capacity was originally conceptualized as a limited number of slots for individual objects (similar to object files)[123–126]. Subsequent research has questioned strong versions of the slots hypothesis. For example, remembered objects do not fail as a unit—rather, object features and their bindings to the object can be forgotten independently for the same object[127,128]. The memory representations may better be characterized as hierarchically structured feature bundles[129] in which bindings and features can fail independently. The capacity of visual working memory has also been characterized as a limited continuous resource that can be divided up among the objects with a different portion allotted to each[130–132]. A related hypothesis is that the object representations interfere with each other within the same substrate[133,134]. Importantly, the concept of working memory goes beyond mere storage. The 'working' part refers to flexible access and control of information for the purpose of higher-order cognitive processes such as visual reasoning[135–137].

Neural network models

The cognitive theories capture the human behavioural phenomena and provide a blueprint for computational models. However, they fall short of fully specifying the algorithm or how it might be implemented in a neurobiologically plausible way. We now discuss attempts to implement untethered object representations in neural network models. Ever since the inception of the first artificial neuron models[138], researchers have studied how cognitive capacities can arise from the interaction of neurons in a network[139]. The classic models were designed for small toy problems, raising the question of whether their computational mechanisms scale to real-world vision. Modern computer hardware and software enable us to test these mechanisms in large-scale models that perform real-world visual tasks. A successful example is the deep convolutional mechanism, which was first implemented in the neocognitron[24] 40 years ago and which, in the past decade, has enabled deep neural networks to perform image recognition[23,140].

Neural network mechanisms and cognitive phenomena. Multilayer perceptrons[141–143] and their convolutional variants[24], including modern deep CNNs[23], lack mechanisms for untethered object representation. However, the classic literature also has a rich history of models that implement mechanisms for untethered object representations, such as completion, grouping, object files and working memory. We first outline some elemental mechanisms for associative completion, gating, routing and grouping and describe how neural networks may represent untethered objects and perform probabilistic inference. We then consider how these elements may interact to implement the cognitive functions of modal and amodal completion, object files and slots, and object permanence.

Associative completion. If a neuron or model unit were to implement a feature detector, it would be useful for it to listen to its neighbours for evidence that its feature is present or absent. When two features are correlated in natural visual experience, bidirectional connections with equal weights between the neurons representing the two features can help both neurons to detect their features in the presence of noise (Fig. 3a). Such connectivity could be acquired by Hebbian learning[144].

The prevalence of smooth contours in natural images renders approximately collinear edge detectors correlated under natural stimulation[43]. There is evidence that V1 neurons selective for collinear edge elements are preferentially connected by excitatory
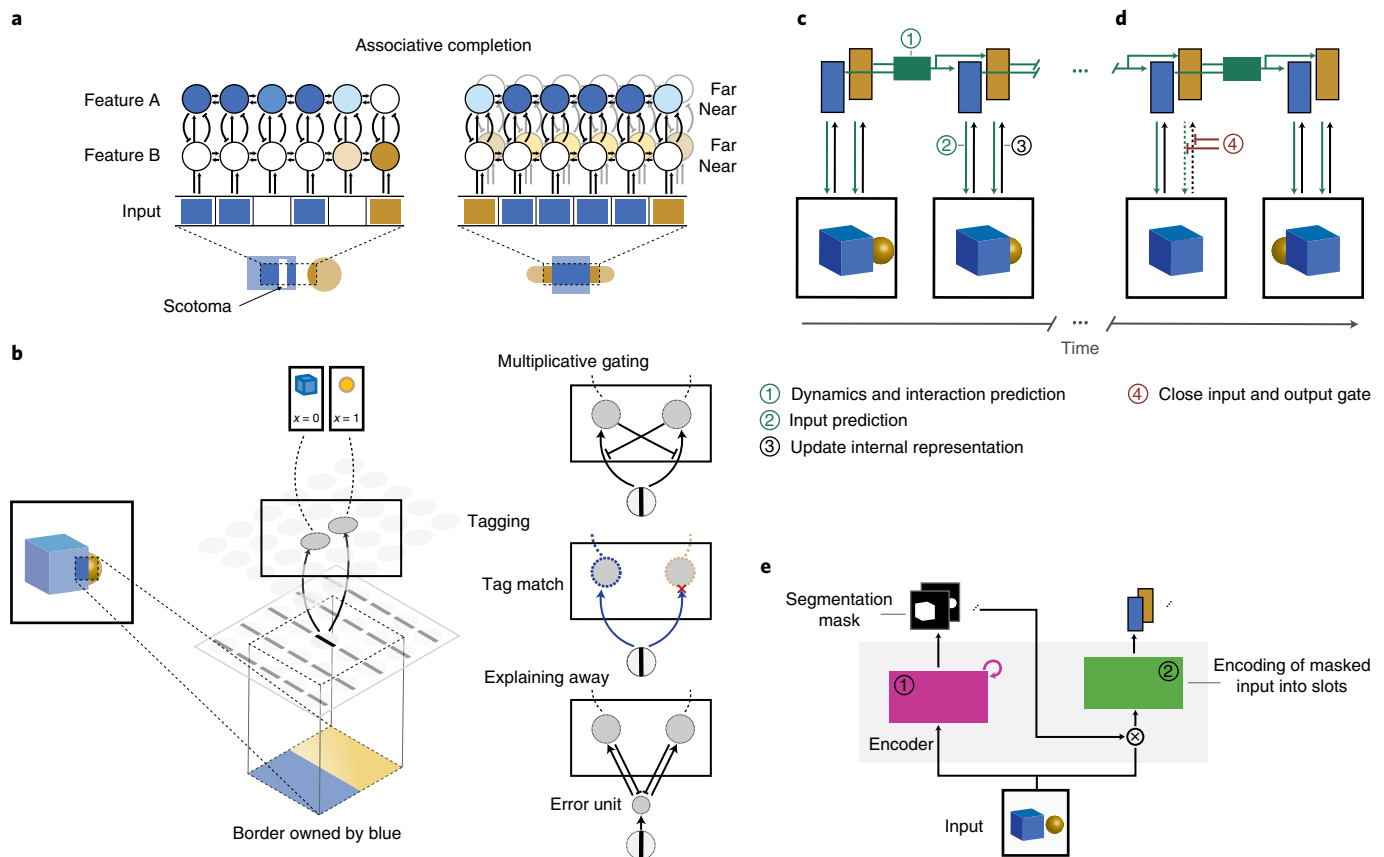
Fig. 3 | Neural network mechanisms for untethering. a, Associative completion can fill in missing information. Here, a scotoma is bridged in the representation via lateral connections that perform modal completion (left). Right, associative processes may also contribute to amodal completion, which also requires units for different depth planes. **b**, Local routing mechanisms enable context-dependent local modulation of the connectivity between units at the time scale of inference. The network layer detects the presence of an edge, which in this case belongs to the blue object. Gating mechanisms selectively route information to the part of the network that represents the blue object. Three gating mechanisms are illustrated. Multiplicative gating suppresses the input to the units that do not represent the target object. Tagging adds a label (for example, a temporal or phase tag) to the activation (the blue lines indicate a tag corresponding to the blue object), which is used by upstream units to filter their inputs. Explaining away subtracts already explained parts from the input[157]. **c**, Predictive processing with structured representations engages multiple mechanisms. Prediction of dynamics and interactions between objects occurs at the level of object representations (for example, slots) (1). The prediction at the abstract level of the latent representation may be decoded into lower-level predictions that are closer to the input at the next time step (2) and object representations are updated depending on the prediction error (3). **d**, Memory gating. During occlusion, the yellow object is persistent and has to untether its connection with the input (4). **e**, Global routing through recurrent spatial attention. Example of separate localization and encoding of objects in a deep neural network[220,222]. A recurrent attention network computes segmentation masks that select portions of the image for routing into separate object slots.

synapses[44]. The lateral connections may implement a diffusion process that regularizes the representation, shrinking it back towards a prior over natural images or collapsing behaviourally irrelevant variability, so as to ease the extraction of relevant information by downstream regions.

Symmetric lateral connectivity can also implement autoassociative completion of complex learned patterns[145]. The weight symmetry enables us to understand the dynamics of the network in terms of an energy function. An activity pattern far from all of the learned patterns will have high energy. From such a point in state space, the dynamics will descend the energy landscape until it reaches a fixed-point attractor, a local minimum of the energy function, corresponding to one of the learned patterns[146]. Associative completion can more generally be understood as predictive regularization. When the predictions are not only across space (as in the example above) but also across time, they can approximate a Bayes filter, which optimally combines past and present evidence. The connection weights between two units will not be symmetric and the dynamics, rather than converging to fixed-point attractors,

can model the dynamics of the environment[22]. Such a mechanism might implement the cognitive phenomenon of representational momentum[112].

Associative completion processes could be used not only within, but also across levels of the visual hierarchy. In either case, associative completion involves interactions between units that directly adjust what we may think of as the units' representational content. We next consider a complementary set of mechanisms that operate at a higher level—modulating interactions between units, rather than unit activity, so as to gate, route and group the representational content.

Gating, routing and grouping. Object representations could be inferred from the input by a set of static filters. However, this approach would require filters for all possible shapes, sizes and locations of objects and their interactions when one partially occludes another. A more efficient solution with respect to the number of units needed is to use static filters for parts (in particular, parts that are frequently encountered) and to dynamically compose the

parts to represent a given object. The composition can be implemented by selectively routing lower-level part representations to the higher-level representation of the object. Architectural connections in a neural network between units representing parts are therefore potential connections, a subset of which is instantiated to represent a specific object. This requires a routing mechanism: a rapid modulation of the connectivity between units at the time scale of inference[147]. An example of routing is a neural-shifter circuit that dynamically maps retinal input from varying locations into a location-invariant (that is object-centred) representation[148–150].

Routing can be implemented by multiplicative modulation of the input gain to a unit[151,152]. During grouping, units can influence the gain functions of other units that compete to explain the same lower-level input. The unit that wins responsibility for the input may end up closing the gate between the input and the other competing units (Fig. 3b).

Instead of attenuating the connectivity between units, a neural network might also use explicit tagging of messages. For example, the message that a neural activation conveys (for example, the presence of a feature) could be tagged with a signal indicating which group it belongs to[153]. A receiving unit could then selectively combine information over inputs with the relevant tag (Fig. 3b). One such mechanism that has been investigated in neuroscience is binding-by-synchrony, in which a temporal tag is provided by the time of firing, and units that fire synchronously are considered to be signalling features of the same object[33,154–156].

Another form of gating is subtractive gating, whereby input to a unit is cancelled by inhibition from a gating unit. For example, predictive coding[157] uses a process of subtractive explaining away, whereby higher-level units explain their lower-level input and subtract their predictions out of the lower-level representation (Fig. 3b). What remains are the unexplained portions of the lower-level representation, the residual errors, which continue to drive the higher-level units. The resulting recurrent dynamics can implement an iterative inference process, in which higher-level units converge to a state in which they jointly account for the input. A higher-level unit that explains a part of the input (for example, an object that clutters or partially occludes another object) will explain away its portion of the image, preventing that portion from interfering with the recognition of the other portions. Predictive coding combines forms of routing and grouping, processing the image in parallel, but successively accounting for more of the objects and their interactions as it progresses from the easy to the difficult parts.

Untethered representation of objects. We refer to object representations as untethered if they are free from immediate control by the sensory stimulus. Untethered representations can combine information over time scales, including recent sensory information (for example, about the trajectory of an object as it moved behind an occluder) and prior knowledge (for example, about the behaviour of objects of a category). To exploit the objects' relative independence in the world, untethered object representations must disentangle the information about different objects[158]. One approach is to dedicate a separate set of units, a neural slot, to the representation of each object. Alternatively, multiple objects can be represented in a shared population of units as distributed representations. Each unit might have mixed coding for different objects, but the information about different objects could still occupy separate linear subspaces. For both slot and mixed representations, the object representations may be distributed across hierarchical levels that jointly encode a scene-parsing tree,[10,155,159] with lower levels encoding detailed features and higher levels more abstract aspects of the object.

Probabilistic inference on a generative model. A neural network implementation of probabilistic inference on a generative model must combine probabilistic beliefs[160] about the latent variables (the prior) with the probability of the sensory data given each possible configuration of latents (the likelihood)[16,157,161]. The generative model would need to specify the prior over the object-level representation and how to generate an image from that representation. Perception then amounts to inversion of the generative model, inferring the object-level representation from an image. Assuming that we are given the generative model, we might train a feedforward neural network to approximate the mapping from data to posterior, using training pairs of images and latents obtained either by drawing latents from the prior and generating images[162] or using a generic inference algorithm to infer latents from images drawn from some distribution. Speeding up inference by memorizing past inferences is called amortization[163]. A feedforward neural network can memorize frequently needed inferences and generalize to new inferences to some extent. However, for complex generative models, the stochastic inverse may not lend itself to efficient representation in a feedforward network with a realistic number of units and weights. Fully leveraging the generative model for generalization may require generative model components to be explicitly implemented and dynamically inverted during perceptual inference, which requires recurrent computations[164]. Challenges with probabilistic inference include the acquisition of the generative model and the amount of computations required for inference. Brains and machines must strike some compromise, combining the statistical efficiency of generative inference with the computational efficiency of discriminative inference. For example, instead of evaluating the likelihood at the level of the image, the inference may evaluate the likelihood at a discriminatively summarized higher level of representation. Furthermore, short of inference of the full posterior, a network may use a generative model to infer only the most probable latent variable configuration for a specific input, the maximum a posteriori estimate[157]. One approach is to seed the inference with a first guess about the objects and their locations computed by a feedforward computation. The initial estimate can then be iteratively refined towards the maximum a posteriori estimate. At each step, the likelihood can be evaluated by synthesizing a reconstruction of the sensory data using a top-down network that implements the generative model.

Inferring object properties beyond the visible input. The associative completion described above can fill-in missing pieces or otherwise repair a representation corrupted by undesirable variability (including internal and external noise, as well as behaviourally irrelevant variation of the objects). Perhaps surprisingly, elaborating the representation through memory regularizes the representation and, therefore, reduces the information about the stimulus. This may be desirable if the information lost is not relevant. If associative completion is to collapse undesirable variability, it should overwrite the sensory representation. This may explain illusory contours and other modal completion phenomena[165] (Fig. 3a). Associative completion might also contribute to amodal completion. For example, the occluded portion of a contour of a simple convex shape could be extrapolated locally using prior assumptions about contour shape (for example, an assumption of smoothness). However, whether associative completion can by itself explain amodal completion phenomena is questionable[166]. An associative mechanism for amodal completion would require dedicating a different set of units to the inferred, but invisible, features. Separate units for inferred features would enable the system to represent the occluder and the occluded parts of the back object simultaneously in different depth planes. More generally, separate units for inferred features might help a probabilistic inference process to avoid confusing inferred features for independent sensory evidence.

Alternatively or in addition to associative completion, amodal completion phenomena may arise through the representation of the object as a whole at a higher level. The same mechanisms[167–170] that

Box 2 | The binding problem

The binding problem refers to a set of computational challenges of how different elements can be flexibly and rapidly linked to each other in a network, in which connections change only at the slow time scale of learning. Binding has often been studied in the context of vision, in which it refers to the binding of parts and properties of objects, objects to locations and objects across time[32]. Binding is not a problem that is intrinsic to vision but results from the specific implementation of a visual system. For example, when different features of the same object (for example, colour and shape) are preferentially analysed in separate, specialized regions, they might need to be linked or recombined together subsequently again. Several solutions to the binding problem in neural networks have been proposed[18,33,292,293]. For example, specialized neurons could signal the presence of specific feature combinations (that is, conjunction coding)[294]. However, this approach is limited due to the combinatorial explosion of possible feature combinations and the fact that only previously learned combinations can be represented. Humans however can perceive and act on arbitrary and previously unseen feature combinations (for example, "Consider seeing a three-legged camel with wings, or a triangular book with a hole through it" (p. 108 of ref. [295])). Distributed representations of conjunctions that encode feature combinations in a coarse code[296] or through tensor product coding[297], or dynamic interunits[298], could alleviate these downsides. Instead of using feature combination detectors, a network could dynamically adapt its weights to bind features of the same object together[147]. Another binding challenge arises when simultaneously perceiving multiple objects. As a consequence of increasing receptive field sizes, higher-level visual neurons receive input from the full visual field and potentially from multiple objects at the same time. This superposition in neuronal populations is problematic if the information cannot be uniquely attributed to the different objects (that is, the superposition catastrophe[299]). How does the brain distinguish between these multiple objects in a distributed representation? One solution may be to sequentially process individual objects[148,177,300]. In the brain, such temporal multiplexing of object representations could be implemented in theta rhythmic neural activity[156]. Furthermore, this selective processing of individual proto-objects might be necessary to bind constituent features into a structural description of the object[41,70,72]. A prominent and highly debated proposal of how the brain solves the binding problem is the idea that binding is expressed through correlated activity of neural assemblies that encode the same object[33,154,155]. Neurons could operate as coincidence detectors of synchronous incoming spikes of feature detectors that represent parts that should be bound together, temporarily increase synaptic efficacy for these inputs and decrease sensitivity to asynchronous inputs (although see ref. [301]). The temporal phase at which feature detectors spike then represents a dimension that labels the temporary grouping that a neuron belongs to.

group the visible features, by combining priors about object shape with sensory information, might also give rise to the percept of an amodally completed object. Higher-order priors on object shape can be implemented in a hierarchical neural network. For example, a hierarchical neural network based on the neocognitron[24] has been shown to infer occluded contours through feedforward and feedback interactions[171].

When we conceptualize the visual system as performing generative inference[2], amodal completion can be considered to be an emergent phenomenon that results from inference about whole objects from partial input. Here, gating and routing mechanisms that instantiate dynamical assignments during hierarchical, iterative inference are particularly important. Lower-level units that respond to the visible parts of a partially occluded object activate units at the next higher level that represent the hypothesis that the object is present. The likelihood of this hypothesis can be evaluated by feedback connections that predict the presence of the full object at the lower, part level[172]. Such predictions will not match the evidence at the site of occlusion, unless the representation of the occluder explains away the occluded portion[173,174]. Alternatively, a feedback-controlled gating mechanism could restrict the evaluation of the likelihood of the presence of the partially occluded object to the unoccluded portion. With either mechanism, the occluder-induced gating prevents the absence of evidence for the object where it is occluded from being misinterpreted as evidence of absence of the object. This is consistent with the fact that occlusions, but not deletions, induce amodal completion[175].

Representing and tracking multiple objects. When multiple objects need to be represented or tracked by object-based representations, an accounting mechanism may be helpful that ensures a one-to-one mapping between slots and objects. Ensuring a one-to-one mapping prevents interference between features of different objects (the superposition problem; Box 2). This can be implemented by different routing mechanisms. One approach is temporal multiplexing—the separation of different objects in time. Temporal multiplexing can operate at a fine temporal scale, with precise spike synchrony[154] or a shared oscillatory phase[153,156], indicating that two signals belong to the same object. Alternatively, temporal multiplexing can operate at a coarse temporal scale, for example, when covert or overt attention sequentially selects different objects[148,176–178]. As an alternative to temporal multiplexing, a unique frequency[179] can be used to tag an object slot and avoid interference with objects represented by other slots. For any of these tagging mechanisms, an inhibitory mechanism between slots can ensure that each slot is assigned a unique tag. In the framework of predictive coding, one-to-one mappings can dynamically emerge through error representations and explaining away. Tracking of objects across time can be achieved by combining the prior prediction of the object's position with the incoming sensory evidence.

Bridging spatiotemporal gaps. As an object moves, it might become occluded by other objects. When it disappears behind an occluder and reappears on the other side later on, the spatiotemporal gap in the stream of visual evidence may be too large for local mechanisms—such as lateral associative filters—to bridge. The gap induced by a full occlusion of the object also severs the established routing between the sensory signals and the object-based representation. How can an object slot re-establish its correspondence to the sensory evidence after such a gap?

An object could be tracked through occlusion via a model-based temporal filter that continuously simulates its hidden state (including its motion and other property transformations) through the period of full occlusion. At the same time, a mechanism is needed that prevents the visual input from the occluder from interfering with the representation of the hidden object. This can be accomplished by a gating mechanism or by recurrent dynamics that separate sensory and mnemonic contents into different linear subspaces of a neural representation[180]. Correspondence with the sensory stream could be re-established if the object reappears within the margin of error of the simulated position.

A short-term memory mechanism can maintain the hidden object state while the object is occluded. Several mechanisms have been proposed to explain how information is maintained in a network over a limited amount of time[181,182]. The most popular class of model proposes that recurrent dynamics retain information in

attractor states[183–186]. Such mechanisms have been used to model object permanence in infants. The mechanism predicts the disappearance of an object behind an occluder, dynamically maintains the representation of the object while it is invisible and predicts its reappearance[187,188].

Short-term memory is a central requirement not only for object tracking but also for many cognitive tasks. An alternative to active maintenance is activity-silent storage, which could be supported by short-term plasticity of connections. The activity representing the object can be restored after retrieval[189,190]. Recently, both active and activity-silent mechanisms have been shown to dynamically interact in short-term memory depending on task demands[191].

Beyond information storage, short-term memory also needs to support flexible updating of content, retrieval of a subset of the information for ongoing computations and selective deletion[192]. Similar to object tracking, these operations require a gating mechanism[193–195] that can rapidly grant access to a stored memory or protect its content from interference (Fig. 3d). The long short-term memory[152] and related gating mechanisms have been successfully used to address this problem.

Modern deep neural networks as models of human object vision. The neural network mechanisms for untethered object perception described in the previous section were often implemented in small models that could handle only toy tasks. Candidate mechanisms for explaining human vision need to scale to real-world tasks. The breakthroughs with deep CNNs[140,196] and the associated hardware and software advances have provided the technological basis for addressing this challenge[197,198].

Modern deep neural network models are typically constructed by training an architecture on a particular objective using backpropagation. The neural mechanisms emerge from the interplay of the architecture, the optimization objective, the learning rule and the training data. On the one hand, learning is necessary for a complex model to absorb the knowledge and skills needed for successful performance under real-world conditions. A vision model, for example, needs to learn what things look like. On the other hand, the fact that the neural mechanisms emerge through learning renders a trained model with millions of parameters mysterious, motivating post hoc investigations into its mechanism[199]. Modellers do exert control over the mechanisms, but at a more abstract level—by designing the architecture, the optimization objective, the learning rule and the training experiences[200].

It is an open question whether brains can use backpropagation or a related error-driven learning rule[201–206]. Whether or not it is biologically plausible, backpropagation can serve as a tool to set the parameters of models meant to capture the computations underlying perceptual performance. When we use it as such, we forgo any claims as to how the interaction of genes, development and experience produced such solutions in humans.

Modern deep neural networks scale up many of the known neural network mechanisms. Feedforward CNNs have been very successful in tasks such as visual object recognition[140,207]. The architecture of CNNs[23,24] is inspired by the primate visual hierarchy. CNNs capture many aspects of cognitive and neuroscientific theories of pre-attentive parallel visual processing. They integrate information over a hierarchy of spatial or spatiotemporal filters, with filter templates replicated across spatial positions. When trained to recognize object categories, their internal representations are similar to those of the human and non-human primate ventral visual stream[208–212].

To date, the best computer-vision models for object recognition are deep CNNs. However, CNNs lack many of the mechanisms of human object perception. For example, it has been shown that these networks rely more strongly on texture than humans, whose recognition prominently depends on global shape information[25,213,214]. CNNs see the image in terms of summary statistics that pool local image features, which provides a surprisingly powerful mechanism for discriminating between object categories. However, they do not decompose the scene into objects, or objects into their parts, as is required for the model to understand the structure of the scene (artificial intelligence (AI) objective) and to explain human cognitive phenomena, such as amodal completion and object files.

Computer vision must solve many tasks beyond texture-based recognition, such as localization, instance segmentation[215,216] and multiple-object tracking (for example, of pedestrians, sports players, vehicles or animals)[217]. Like the human visual system, these models must localize, individuate, identify and keep track of multiple objects. They use computational strategies that are broadly similar to those in the cognitive literature. For example, object localization models[218] use region-proposal methods, a strategy that is similar to the saliency maps of the visual system[176,219], and sequential instance segmentation and recognition of objects[220,221] (Fig. 3e), which resembles the cognitive theory of sequential individuation and identification[85]. Computer vision also uses global shifts of attention as a form of temporal multiplexing to infer multiple objects[222]. Computer-vision systems often combine learned CNN components with hand-crafted higher-level mechanisms such as physics engines[223], providing interesting hybrid (cognitive and neural) models that could be tested formally as models of human vision. However, it is also important to pursue more organically integrated RNN models that can maintain representations over time, sequentially attend to different portions of the visual input, and individuate, identify and track multiple objects.

Models that are more consistent with human object vision can be developed by introducing constraints at each of Marr's three levels of analysis[57]: the level of biological implementation, the level of representation and algorithm, and the level of the computational objective. We consider these three levels in turn.

Constraints from neurobiology. Deep CNNs provide a coarse abstraction of the feedforward computations performed by the human visual system. However, they do not have lateral and top-down recurrent connections and therefore lack the ability to maintain representations over time[164]. RNN models trained on object recognition provide better models of human brain representations and behaviour compared with deep feedforward networks[224–227]. Segmentation, identification and amodal completion of object instances are naturally solved by iterative algorithms that can be implemented in recurrent networks. This may explain why neural networks endowed with recurrence yield better performance in object recognition under challenging conditions such as occlusions[224,228,229]. Biologically inspired gating of lateral connections has been shown to yield more sample-efficient training during tasks such as segmentation[230]. Neurobiology continues to provide rich inspiration for modelling work that will explore the computational benefits of more realistic model units, architectural connectivity and learning rules.

Constraints on representations and algorithms. The space of possible solutions that an RNN may implement for a particular task is large. Object-based representations or generative inference do not automatically emerge through task training. Modellers have therefore endowed their architectures with representational structure thought to reflect aspects of the generative structure of the world. For example, models use neural slots at the latent level for inference in static images and in dynamic tasks[220,222,223,231–233]. Slots are attractive because they are interpretable and provide a strong inductive bias for task-trained models. However, slots may fall short in capturing phenomena such as illusory conjunctions[72] or the capacity limitations of human cognition[124,126], which can manifest in gradual degradation of the fidelity with which objects are represented as the number of objects grows[130–132]. Representing a variable number of

objects in a shared neural population resource[133,234–236] combined with binding mechanisms (Box 2) promises to explain these cognitive phenomena.

Modellers can also constrain the inference algorithm by imposing hierarchical representations[237,238]. Inference in capsule networks[237,239] is based on the idea that the visual input can be segmented into hierarchical groupings of parts. The recurrent inference process decomposes a scene into a hierarchy of parts[10,155,159]. This is accomplished by a routing mechanism that enhances the connectivity between the lower-level capsule and the corresponding higher-level capsule while attenuating connectivity to competing higher-level capsules thereby implementing explaining away. Humans and feedforward neural network models both struggle to recognize objects in visual clutter, a phenomenon that is known as visual crowding[240] (Box 1d). However, human recognition of the central object is undiminished if the visual clutter can be explained away as part of other objects. This uncrowding effect[241] has recently been demonstrated for capsule networks[242], which separate the clutter from the object by representing each in a different capsule.

Discrete relational structures can be expressed in a graph, in which objects and parts are nodes and edges represent relations. Graph neural networks provide a general and powerful class of model that can perform computations on a graph using neural network components[243,244]. A softer way to impose structure is to encourage the emergence of a disentangled representation through a prior on the latent space[158,245]. A key question for current research is how structured representations and computations may be acquired through experience and implemented in biologically plausible neural networks[246].

Constraints on the computational objective. Recent modelling work has moved beyond supervised training objectives, such as mapping images to labels. Rooted in theories of biological reinforcement learning, deep reinforcement learning requires weaker external feedback (just a reward signal), making it more realistic as a model of how an agent might learn through interaction[247,248]. In the absence of any feedback, an agent can use unsupervised learning, aiming to capture statistical dependencies in the sensory data. An agent interested in all regularities, not just those that are useful for a specific task, will learn a generative model of the data and can base inferences on the more comprehensive understanding provided by such a model[1,16,157]. To learn all kinds of regularities, an agent may challenge itself with its own games of prediction. In self-supervised learning, the model learns to predict portions of the data from other portions across time and space (for example, the future from the past and vice versa, the left half from the right half and vice versa)[249]. The ability to learn without any feedback may be essential for acquisition of knowledge that generalizes to new tasks.

Self-supervised learning techniques have reinvigorated the construction of complex generative models of images and videos[250–252]. Although the true generative model of visual data is intractable, these models learn rich compositional structure to meet their training objectives, such as predicting upcoming video frames. Object representations provide a natural way to compress and predict the physical world, rendering compression and prediction promising objectives for unsupervised learning of object representations[197]. Nevertheless, learning object-based representations by self-supervision still appears to require strong structural inductive biases on the generative model[253].

Even for a simplified generative model of real-world visual data, inferring the posterior over the latents is intractable. Most deep generative models amortize the inference into a feedforward recognition model. The human brain most likely uses a balance between amortized inference using a feedforward mechanism and iterative generative inference using a recurrent mechanism[164]. Neural network models with object representations that combine amortized

and generative inference[233,254] may more closely capture the inference dynamics of the human visual system. Discovering good latent representations and approximate inference algorithms will require bringing together the perspectives of engineering, neuroscience and cognitive science.

Towards neural network models with untethered object representations

The cognitive and modelling literatures present the pieces of the puzzle: the cognitive component functions and potential neural mechanisms. Now we have to put the pieces together and build models of how humans see the world as structured into objects under natural conditions. This will require a new scale of collaboration among cognitive scientists and engineers.

Two key components of this endeavour are tasks and benchmarks. A task is a computer-simulated environment that an agent (a human, other animal or computational model) interacts with through an interface of perceptions and actions. Computer-administered tasks give us control of all aspects of the interaction. We can design the task world: its perceptual appearance, the set of actions available and the objectives and rewards.

Tasks lend direction to cognitive science and AI by posing well-defined challenges that provide stepping stones and enable us to measure cognitive performance. In cognitive science, a task carves out what behaviours are under investigation. In AI, a task defines the engineering challenge. If cognitive science and engineering are to provide useful constraints for each other, it will be essential that they engage a shared set of tasks. Tasks should be designed and implemented for use in both human behavioural experiments and neural network modelling[255,256]. To enable the training and testing of models, stimuli and task scenarios should be procedurally generated to enable the production of an infinite number of new experiences.

Tasks form the basis of behavioural benchmarks for models: model evaluation functions that define progress and enable us to select and improve models. We now discuss how new tasks and benchmarks shared among cognitive scientists and engineers can drive progress.

Tasks to train and test untethered object perception. Cognitive scientists and engineers tend to design tasks by different criteria, resulting in little overlap in the tasks used. Engineers have focused on tasks that are relevant to real-world applications, often engaging complex natural stimuli and dynamics[257–259]. Modelling performance under natural conditions is the ultimate goal. However, complex models are slow to train and difficult to understand. Thus, engineers should also engage simplified tasks that focus on particular computational challenges. Cognitive scientists often strive to carve cognition at its joints, guided by assumptions about the mind. This has classically led to tasks stripped down to the essential elements required to expose some cognitive component. Simple controlled tasks promise to isolate the primitives of cognitive function[91,132,260,261], rendering behaviour directly interpretable in terms of cognitive theory (Box 1). However, we must also engage complex and naturalistic tasks to understand how the primitives interact and scale to real-world cognition. Although behaviour in complex tasks is harder to interpret per se, it can be used to adjudicate among explicit computational models. Neural network models therefore relax the constraint for our tasks to isolate cognitive primitives, liberating us to explore more complex naturalistic tasks. Even if our tasks do not carve cognition at its joints, they can usefully focus our investigation on a subset of cognitive phenomena, the computational mechanisms of which are within our reach of understanding.

Cognitive scientists and engineers can then benefit from co-opting each other's criteria for a good task. As the former are looking to engage cognition under natural conditions and the
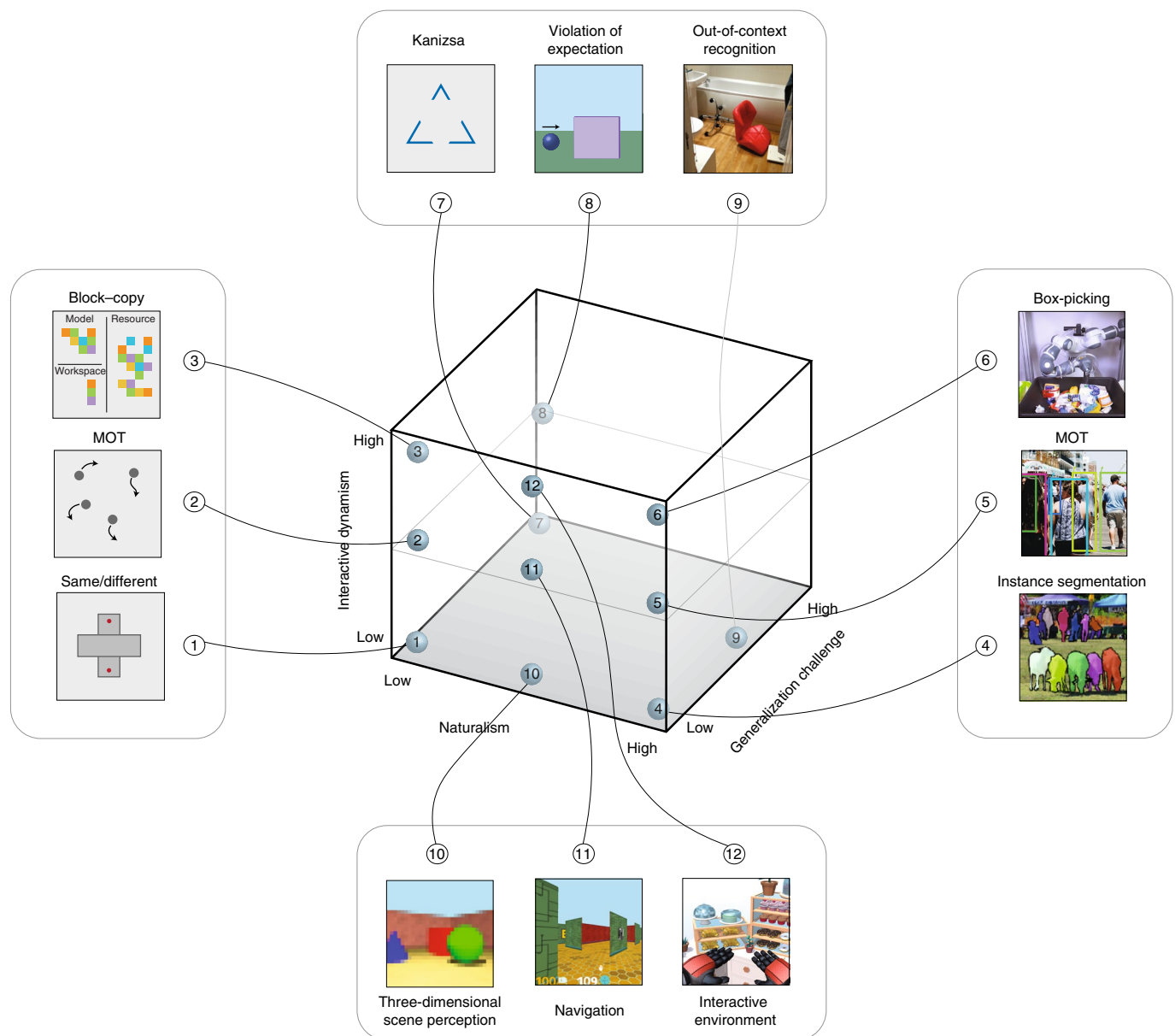
Fig. 4 | Space of tasks for untethered object perception. Three particularly important dimensions of the space of tasks are naturalism, interactive dynamism and generalization challenge. Naturalism (*x* axis): tasks can be rendered naturalistically or abstracted to their essence. Tasks used in cognitive science (1–3) and machine learning (4–6) tend to concentrate at opposing poles of the naturalism axis. Computer-simulated environments and virtual reality enable us to bridge this gap (10 (ref. [285]) and 11, DeepMind Lab[286]; 12, A2I-THOR[287]). Interactive dynamism (*y* axis): this axis summarizes the degree of dynamism of the stimuli (for example, video versus static image) and responses (for example, motion trajectory versus button press) and the degree of interactivity (that is, the rate and balance of sensory and motor information flow). Static stimuli as in grouping (1) and segmentation (4) tasks[288], dynamic stimuli as in multiple-object tracking tasks (2 and 5)[289], interactive tasks as in the block-copy tasks (3)[266] or box-picking tasks (6, a robot arm has to pick objects from a box with objects)[290] are indicated. Generalization challenge (*z* axis): tasks can be loosely ordered by the degree to which stimuli are representative of situations encountered during training, be it evolution and learning for the human visual system or the training set used to optimize a neural network model. Tasks that confront the system with untypical (that is, out-of-training-distribution) situations (7–8; 9, Objectnet[268]) have high generalization demands and can help to reveal the inductive biases of the visual system[29]. MOT, multiple-object tracking.

latter are seeking to discover the computational components missing from current AI models, both fields should engage the whole spectrum of tasks, from simple toy tasks to natural dynamic tasks. This strengthens the motivation to collaborate across disciplines on a shared set of tasks.

Cognitive tasks such as segmentation, visual search, multiple-object tracking, physics prediction or goal-oriented manipulation are good starting points because they focus on plausible cognitive primitives. The world in each of these tasks is a scene composed of persistent objects that can occlude each other and may obey some approximation to Newtonian physics. Here we propose to push tasks toward greater complexity along three particularly important axes: naturalism, interactive dynamism and generalization challenge (Fig. 4).

Naturalism. Naturalism refers to the degree to which the simulated task world resembles the real world. Although abstract stimuli are useful for adjudicating among simple models[262], the ultimate goal

is to explain perception under natural conditions[263]. A synthesis of these two complementary approaches is provided by methods that optimize stimuli to adjudicate among complex models[29], yielding synthetic stimuli that reflect the natural image statistics the models have learned. Similarly, for object-based vision, tasks should achieve various degrees of naturalism while enabling us to adjudicate among models that implement alternative computational theories. We can develop these tasks towards greater naturalism by replacing abstract shapes with photos or 3D models of objects. Incorporating different object categories into these tasks enables us to study the domain specialization of the mechanisms of object perception. For example, tracking of humans and inanimate objects may rely on separate replications of these mechanisms (independent slots) that bring in particular prior knowledge about humans, animals and inanimate objects.

Interactive dynamism. Object representations support continuous interaction with a dynamic world (Fig. 1). Perception operates at multiple time scales, supporting higher cognitive functions, including memory, prediction and planning. We therefore need tasks that probe performance in dynamic and interactive settings. Cognitive science originally investigated untethered object perception with tasks in which a predefined set of static stimuli presented on separate trials elicited a button-press response[100,264,265] (Fig. 1). However, more dynamic tasks such as multiple-object tracking[85,91] and interactive tasks such as reproducing an arrangement of blocks[266] (Fig. 1) have also been developed. In non-interactive tasks, the initial state is controlled by the experimenter in each of a sequence of trials, rendering behavioural responses easier to analyse and more directly interpretable. However, when our theories have been expressed in computational models, we can also use interactive dynamic tasks to adjudicate among theories. In fact, interactive dynamic tasks will often have a higher bit rate of recorded behaviour, promising greater constraints on theory, in addition to enabling us to understand how agents engage dynamic, interactive environments. Tasks can be pushed from simple toy tasks towards greater interactive dynamism by giving the objects dynamic trajectories and recording responses such as mouse-pointer or eye movements continuously.

Generalization challenge. New experiences require generalization and are often particularly revealing of the computational mechanism and inductive bias used by a perceptual system. By probing a model with parameters of the task-generative world that differ from the training distribution, we can generate generalization tests that reveal a model's inductive bias[26,267]. To probe untethered object representations, we can present humans and models with new objects (for example, procedurally generated 3D models) or with known objects in new poses or contexts[253,268] and study whether task performance generalizes. Tracked objects may change their appearance and shape across time[269], which may be difficult for models that track by appearance, but easy for humans who primarily track objects based on spatiotemporal properties[20,74–76]. We may also use Gestalt stimuli that elicit grouping in humans (for example, point light displays of biological motion[270]). We may push our notion of generalization even further to scenarios in which there may be no objectively correct response. For example, there is no objectively correct inference to perceive either one or two distinct objects during the tunnel effect[76]. However, humans perceive a single object when the spatiotemporal dynamics are consistent with the motion of a single object, revealing the implicit prior assumption that objects are more likely to change than to vanish and appear. Cognitive scientists have probed human perceptual inductive biases with hand-designed stimuli and controlled tasks. These form the basis for generative models of stimuli and tasks that will enable us to comprehensively test and compare generalization behaviour in humans and machines.

Benchmarks to evaluate models. Tasks form the basis for defining behavioural benchmarks for models. A benchmark is an evaluation function that enables us to select and improve models and to define progress. Engineering has relied on overall task-performance benchmarks[257]. However, a benchmark can also be defined to measure how close a model comes to emulating human patterns of success and failure across different stimuli and contexts[29,256,271–275]. For dynamic interactive tasks, each behavioural episode of a human or model generates a unique trajectory of stimuli and responses. A major challenge is to define useful summary statistics that enable comparisons among humans and models.

Summary statistics can be based on patterns of responses or the performance in a task, such as multiple-object tracking, physical reasoning[276], physical scene understanding[277–280], goal-directed manipulation of objects[276,281] or navigation[282]. A qualitative description such as 'performs mental physics simulation' or 'can do object tracking' provides only a coarse characterization of a cognitive process. Benchmarks should be based on summary statistics that provide rich quantitative signatures of behaviour (for example, tracking performance as a function of the number of objects to be tracked and other context variables), revealing how humans differ from models[278,281]. Psychophysics and cognitive psychology have developed an arsenal of ingenious methods to probe object perception in humans (Box 1), providing much inspiration for the development of benchmarks measuring the behavioural similarity between models and humans[267,274].

Conclusion

Perceiving the world around us in terms of objects provides a powerful inductive bias that links perception to symbolic cognition and action, and forms the basis of our causal understanding of the physical world. Object percepts form through a constructive process of interaction among stages of representation. Deep neural network models have begun to capture components of the process by which object percepts emerge, including grouping, segmentation and tracking. They do not yet capture the interplay between these components and the powerful abstract inductive biases of human vision. A common set of tasks and benchmarks will help cognitive scientists and engineers to join forces. For our models to achieve human-level performance, we will need to be interested not only in the successes, but also in the detailed patterns of failure that characterize human vision.

References

1. Von Helmholtz, H. *Handbuch der Physiologischen Optik* (Voss, 1867).
2. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
3. Pearl, J. *Causality* (Cambridge Univ. Press, 2009).
4. Piaget, J. *The Construction of Reality in the Child* (Basic Books, 1954).
5. Adelson, E. H. On seeing stuff: the perception of materials by humans and machines. In *Human Vision and Electronic Imaging VI* (eds. Rogowitz, B. E. & Pappas, T. N.) vol. 4299 1–12 (SPIE, 2001).
6. Clowes, M. B. On seeing things. *Artif. Int.* **2**, 79–116 (1971).
7. Julesz, B. Experiments in the visual perception of texture. *Sci. Am.* **232**, 34–43 (1975).
8. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Ann. Rev. Neurosci.* **24**, 1193–1216 (2001).
9. Rosenholtz, R., Li, Y. & Nakano, L. Measuring visual clutter. *J. Vis.* **7**, 17–17 (2007).
10. Hoffman, D. D. & Richards, W. A. Parts of recognition. *Cognition* **18**, 65–96 (1984).
11. Michotte, A. et al. *Les Complements Amodaux des Structures Perceptives* (Institut de psychologie de l'Université de Louvain, 1964).
12. Rensink, R. A. The dynamic representation of scenes. *Visual Cogn.* **7**, 17–42 (2000).
13. Gregory, R. L. Perceptions as hypotheses. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **290**, 181–197 (1980).

14. Rock, I. *Indirect Perception* (The MIT Press, 1997).
15. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
16. Friston, K. J. A theory of cortical responses. *Phil. Trans. R. Soc. B Biol. Sci.* **360**, 815–836 (2005).
17. van Steenkiste, S., Greff, K. & Schmidhuber, J. A perspective on objects and systematic generalization in model-based RL. Preprint at http://arxiv.org/abs/1906.01035 (2019).
18. Greff, K., van Steenkiste, S. & Schmidhuber, J. On the binding problem in artificial neural networks. Preprint at http://arxiv.org/abs/2012.05208 (2020).
19. Spelke, E. S. Principles of object perception. *Cogn. Sci.* **14**, 29–56 (1990).
20. Scholl, B. J. Object persistence in philosophy and psychology. *Mind Lang.* **22**, 563–591 (2007).
21. Sarkka, S. *Bayesian Filtering and Smoothing* (Cambridge Univ. Press, 2013).
22. Deneve, S., Duhamel, J.-R. & Pouget, A. Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of Kalman filters. *J. Neurosci.* **27**, 5744–5756 (2007).
23. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
24. Fukushima, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
25. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. 7th International Conference on Learning Representations* (OpenReview.net, 2019).
26. Kansky, K. et al. Schema networks: zero-shot transfer with a generative causal model of intuitive physics. In *Proc. 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) vol. 70 1809–1818 (PMLR, 2017).
27. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
28. Yildirim, I., Wu, J., Kanwisher, N. & Tenenbaum, J. An integrative computational architecture for object-driven cortex. *Curr. Opin. Neurobiol.* **55**, 73–81 (2019).
29. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl Acad. Sci.* **117**, 29330–29337 (2020).
30. Gibson, J. J. *The Ecological Approach to Visual Perception: Classic Edition* (Houghton Mifflin, 1979).
31. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
32. Treisman, A. The binding problem. *Curr. Opin. Neurobiol.* **6**, 171–178 (1996).
33. von der Malsburg, C. The Correlation Theory of Brain Function. in *Models of Neural Networks: Temporal Aspects of Coding and Information Processing in Biological Systems* (eds. Domany, E., van Hemmen, J. L. & Schulten, K.) 95–119 (Springer, 1981).
34. Duncan, J. Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.* **113**, 501–517 (1984).
35. Neisser, U. *Cognitive Psychology* (Appleton-Century-Crofts, 1967).
36. Treisman, A. Features and objects in visual processing. *Sci. Am.* **255**, 114–125 (1986).
37. Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1993).
38. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).
39. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **28**, 229–289 (1965).
40. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
41. Roelfsema, P. R. Cortical algorithms for perceptual grouping. *Ann. Rev. Neurosci.* **29**, 203–227 (2006).
42. Field, D. J., Hayes, A. & Hess, R. F. Contour integration by the human visual system: evidence for a local 'association field'. *Vis. Res.* **33**, 173–193 (1993).
43. Geisler, W. S. Visual perception and the statistical properties of natural scenes. *Ann. Rev. Psychol.* **59**, 167–192 (2008).
44. Bosking, W. H., Zhang, Y., Schofield, B. & Fitzpatrick, D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* **17**, 2112–2127 (1997).
45. Koffka, K. *Principles of Gestalt Psychology* (Harcourt, Brace, 1935).
46. Rock, I. & Palmer, S. The legacy of gestalt psychology. *Sci. Am.* **263**, 84–91 (1990).
47. Wertheimer, M. Untersuchungen zur lehre von der gestalt. *Psychol. Forsch.* **4**, 301–350 (1923).
48. Nakayama, K. & Shimojo, S. Experiencing and perceiving visual surfaces. *Science* **257**, 1357–1363 (1992).
49. Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N. & Wattenberg, M. An intuitive model of perceptual grouping for HCI design. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 1331–1340 (ACM, 2009).
50. Li, Z. A neural model of contour integration in the primary visual cortex. *Neural Comput.* **10**, 903–940 (1998).
51. Yen, S.-C. & Finkel, L. H. Extraction of perceptually salient contours by striate cortical networks. *Vis. Res.* **38**, 719–741 (1998).
52. Roelfsema, P. R., Lamme, V. A. & Spekreijse, H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature* **395**, 376–381 (1998).
53. Nakayama, K. & Silverman, G. H. Serial and parallel processing of visual feature conjunctions. *Nature* **320**, 264–265 (1986).
54. Alais, D., Blake, R. & Lee, S.-H. Visual features that vary together over time group together over space. *Nat. Neurosci.* **1**, 160–164 (1998).
55. Vecera, S. P. & Farah, M. J. Is visual image segmentation a bottom-up or an interactive process? *Percept. Psychophys.* **59**, 1280–1296 (1997).
56. Sekuler, A. & Palmer, S. Perception of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol. Gen.* **121**, 95–111 (1992).
57. Marr, D., Ullman, S. & Poggio, T. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information* (W. H. Freeman, 1982) .
58. Michotte, A. & Burke, L. Une nouvelle enigme dans la psychologie de la perception: le'donne amodal'dans l'experience sensorielle. In *Proc. XIII Congrés Internationale de Psychologie* 179–180 (1951).
59. Komatsu, H. The neural mechanisms of perceptual filling-in. *Nat. Rev. Neurosci.* **7**, 220–231 (2006).
60. Shore, D. I. & Enns, J. T. Shape completion time depends on the size of the occluded region. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 980–998 (1997).
61. He, Z. J. & Nakayama, K. Surfaces versus features in visual search. *Nature* **359**, 231–233 (1992).
62. Rensink, R. A. & Enns, J. T. Early completion of occluded objects. *Vis. Res.* **38**, 2489–2505 (1998).
63. Kellman, P. J. & Shipley, T. F. A theory of visual interpolation in object perception. *Cogn. Psychol.* **23**, 141–221 (1991).
64. Tse, P. U. Volume completion. *Cogn. Psychol.* **39**, 37–68 (1999).
65. Buffart, H., Leeuwenberg, E. & Restle, F. Coding theory of visual pattern completion. *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 241–274 (1981).
66. Weigelt, S., Singer, W. & Muckli, L. Separate cortical stages in amodal completion revealed by functional magnetic resonance adaptation. *BMC Neurosci.* **8**, 70 (2007).
67. Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J. & van Lier, R. Neuroimaging findings on amodal completion: a review. *i-Perception* **10**, 2041669519840047 (2019).
68. Mooney, C. M. Age in the development of closure ability in children. *Can. J. Psychol.* **11**, 219–226 (1957).
69. Snodgrass, J. G. & Feenan, K. Priming effects in picture fragment completion: support for the perceptual closure hypothesis. *J. Exp. Psychol. Gen.* **119**, 276–296 (1990).
70. Treisman, A. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
71. Pylyshyn, Z. W. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav. Brain Sci.* **22**, 341–365 (1999).
72. Wolfe, J. M. & Cave, K. R. The psychophysical evidence for a binding problem in human vision. *Neuron* **24**, 11–17 (1999).
73. Ullman, S. The interpretation of structure from motion. *Proc. R. Soc. Lon. B Biol. Sci.* **203**, 405–426 (1979).
74. Flombaum, J. I., Scholl, B. J. & Santos, L. R. Spatiotemporal priority as a fundamental principle of object persistence. In *The Origins of Object Knowledge* (eds. Hood, B. M. & Santos, L. R.) 135–164 (Oxford University Press, 2009).
75. Mitroff, S. R. & Alvarez, G. A. Space and time, not surface features, guide object persistence. *Psychon. Bull. Rev.* **14**, 1199–1204 (2007).
76. Burke, L. On the tunnel effect. *Quart. J. Exp. Psychol.* **4**, 121–138 (1952).
77. Flombaum, J. I. & Scholl, B. J. A temporal same-object advantage in the tunnel effect: facilitated change detection for persisting objects. *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 840–853 (2006).
78. Hollingworth, A. & Franconeri, S. L. Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition* **113**, 150–166 (2009).
79. Moore, C. M., Stephens, T. & Hein, E. Features, as well as space and time, guide object persistence. *Psychon. Bull. Rev.* **17**, 731–736 (2010).
80. Papenmeier, F., Meyerhoff, H. S., Jahn, G. & Huff, M. Tracking by location and features: object correspondence across spatiotemporal discontinuities during multiple object tracking. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 159–171 (2014).
81. Liberman, A., Zhang, K. & Whitney, D. Serial dependence promotes object stability during occlusion. *J. Vis.* **16**, 16 (2016).

82. Fischer, C. et al. Context information supports serial dependence of multiple visual objects across memory episodes. *Nat. Commun.* **11**, 1932 (2020).

83. Irwin, D. E. Memory for position and identity across eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.* **18**, 307–317 (1992).

84. Richard, A. M., Luck, S. J. & Hollingworth, A. Establishing object correspondence across eye movements: flexible use of spatiotemporal and surface feature information. *Cognition* **109**, 66–88 (2008).

85. Kahneman, D., Treisman, A. & Gibbs, B. J. The reviewing of object-files: object specific integration of information. *Cogn. Psychol.* **24**, 174–219 (1992).

86. Pylyshyn, Z. W. The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. *Cognition* **32**, 65–97 (1989).

87. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).

88. Cavanagh, P. & Alvarez, G. A. Tracking multiple targets with multifocal attention. *Trends Cogn. Sci.* **9**, 349–354 (2005).

89. Bahcall, D. O. & Kowler, E. Attentional interference at small spatial separations. *Vis. Res.* **39**, 71–86 (1999).

90. Franconeri, S. L., Alvarez, G. A. & Cavanagh, P. Flexible cognitive resources: competitive content maps for attention and memory. *Trends Cogn. Sci.* **17**, 134–141 (2013).

91. Pylyshyn, Z. W. & Storm, R. W. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vis.* **3**, 179–197 (1988).

92. Intriligator, J. & Cavanagh, P. The spatial resolution of visual attention. *Cognit. Psychol.* **43**, 171–216 (2001).

93. Scholl, B. J. & Pylyshyn, Z. W. Tracking multiple items through occlusion: clues to visual objecthood. *Cognit. Psychol.* **38**, 259–290 (1999).

94. Yantis, S. Multielement visual tracking: attention and perceptual organization. *Cognit. Psychol.* **24**, 295–340 (1992).

95. Vul, E., Alvarez, G., Tenenbaum, J. B. & Black, M. J. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Proc. Advances in Neural Information Processing Systems* 22 (eds Bengio, Y. et al.) 1955–1963 (Curran Associates, 2009).

96. Alvarez, G. A. & Franconeri, S. L. How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *J. Vis.* **7**, 14 (2007).

97. Flombaum, J. I., Scholl, B. J. & Pylyshyn, Z. W. Attentional resources in visual tracking through occlusion: the high-beams effect. *Cognition* **107**, 904–931 (2008).

98. Vecera, S. P. & Farah, M. J. Does visual attention select objects or locations? *J. Exp. Psychol. Gen.* **123**, 146–160 (1994).

99. Chen, Z. Object-based attention: a tutorial review. *Atten. Percept. Psychophys.* **74**, 784–802 (2012).

100. Egly, R., Driver, J. & Rafal, R. D. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *J. Exp. Psychol. Gen.* **123**, 161–177 (1994).

101. Houtkamp, R., Spekreijse, H. & Roelfsema, P. R. A gradual spread of attention. *Percept. Psychophys.* **65**, 1136–1144 (2003).

102. Jeurissen, D., Self, M. W. & Roelfsema, P. R. Serial grouping of 2D-image regions with object-based attention in humans. *eLife* **5**, e14320 (2016).

103. Moore, C. M., Yantis, S. & Vaughan, B. Object-based visual selection: evidence from perceptual completion. *Psychol. Sci.* **9**, 104–110 (1998).

104. Peters, B., Kaiser, J., Rahm, B. & Bledowski, C. Activity in human visual and parietal cortex reveals object-based attention in working memory. *J. Neurosci.* **35**, 3360–3369 (2015).

105. Peters, B., Kaiser, J., Rahm, B. & Bledowski, C. Object-based attention prioritizes working memory contents at a theta rhythm. *J. Exp. Psychol. Gen.* (2020).

106. Baillargeon, R. Object permanence in 3 1/2- and 4 1/2-month-old infants. *Dev. Psychol.* **23**, 655–664 (1987).

107. Baillargeon, R., Spelke, E. S. & Wasserman, S. Object permanence in five-month-old infants. *Cognition* **20**, 191–208 (1985).

108. Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K. Origins of knowledge. *Psychol. Rev.* **99**, 605–632 (1992).

109. Wilcox, T. Object individuation: infants' use of shape, size, pattern, and color. *Cognition* **72**, 125–166 (1999).

110. Rosander, K. & von Hofsten, C. Infants' emerging ability to represent occluded object motion. *Cognition* **91**, 1–22 (2004).

111. Moore, M. K., Borton, R. & Darby, B. L. Visual tracking in young infants: evidence for object identity or object permanence? *J. Exp. Child Psychol.* **25**, 183–198 (1978).

112. Freyd, J. J. & Finke, R. A. Representational momentum. *J. Exp. Psychol. Learn. Mem. Cognit.* **10**, 126–132 (1984).

113. Benguigui, N., Ripoll, H. & Broderick, M. P. Time-to-contact estimation of accelerated stimuli is based on first-order information. *J. Exp. Psychol. Hum. Percept. Perform.* **29**, 1083–1101 (2003).

114. Rosenbaum, D. A. Perception and extrapolation of velocity and acceleration. *J. Exp. Psychol. Hum. Percept. Perform.* **1**, 395–403 (1975).

115. Franconeri, S. L., Pylyshyn, Z. W. & Scholl, B. J. A simple proximity heuristic allows tracking of multiple objects through occlusion. *Atten. Percept. Psychophys.* **74**, 691–702 (2012).

116. Matin, E. Saccadic suppression: a review and an analysis. *Psychol. Bull.* **81**, 899–917 (1974).

117. Henderson, J. M. Two representational systems in dynamic visual identification. *J. Exp. Psychol. Gen.* **123**, 410–426 (1994).

118. Bahrami, B. Object property encoding and change blindness in multiple object tracking. *Visual Cogn.* **10**, 949–963 (2003).

119. Pylyshyn, Z. Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cogn.* **11**, 801–822 (2004).

120. Horowitz, T. S. et al. Tracking unique objects. *Percept. Psychophys.* **69**, 172–184 (2007).

121. Fougnie, D. & Marois, R. Distinct capacity limits for attention and working memory: evidence from attentive tracking and visual working memory paradigms. *Psychol. Sci.* **17**, 526–534 (2006).

122. Hollingworth, A. & Rasmussen, I. P. Binding objects to locations: the relationship between object files and visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **36**, 543–564 (2010).

123. Awh, E., Barton, B. & Vogel, E. K. Visual working memory represents a fixed number of items regardless of complexity. *Psychol. Sci.* **18**, 622–628 (2007).

124. Cowan, N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001).

125. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–284 (1997).

126. Miller, G. A. The magical number seven. *Psychol. Rev.* **63**, 81–97 (1956).

127. Bays, P. M., Wu, E. Y. & Husain, M. Storage and binding of object features in visual working memory. *Neuropsychologia* **49**, 1622–1631 (2011).

128. Fougnie, D. & Alvarez, G. A. Object features fail independently in visual working memory: evidence for a probabilistic feature-store model. *J. Vis.* **11**, 3 (2011).

129. Brady, T. F., Konkle, T. & Alvarez, G. A. A review of visual memory capacity: beyond individual items and toward structured representations. *J. Vis.* **11**, 4 (2011).

130. Alvarez, G. A. & Cavanagh, P. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol. Sci.* **15**, 106–111 (2004).

131. Bays, P. M. & Husain, M. Dynamic shifts of limited working memory resources in human vision. *Science* **321**, 851–854 (2008).

132. Wilken, P. & Ma, W. J. A detection theory account of change detection. *J. Vis.* **4**, 1120–1135 (2004).

133. Oberauer, K. & Lin, H.-Y. An interference model of visual working memory. *Psychol. Rev.* **124**, 21–59 (2017).

134. Bouchacourt, F. & Buschman, T. J. A flexible model of working memory. *Neuron* **103**, 147–160 (2019).

135. Baddeley, A. D. & Hitch, G. Working Memory. In *Psychology of Learning and Motivation* (ed. Bower, G. H.) vol. 8 47–89 (Academic, 1974).

136. Cowan, N. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychol. Bull.* **104**, 163–191 (1988).

137. Miyake, A. & Shah, P. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (Cambridge Univ. Press, 1999).

138. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **5**, 115–133 (1943).

139. O'Reilly, R. C. & Munakata, Y. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, 2000).

140. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems 25* (eds. Pereira, F. et al.) (Curran Associates, 2012).

141. Rosenblatt, F. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms* Technical Report (Cornell Aeronautical Lab, 1961).

142. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

143. Ivakhnenko, A. G. Polynomial theory of complex systems. In *Proc. IEEE transactions on Systems, Man, and Cybernetics* 364–378 (IEEE, 1971).

144. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory* (J. Wiley, Chapman & Hall, 1949).

145. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).

146. Zemel, R. S. & Mozer, M. C. Localist attractor networks. *Neural Comput.* **13**, 1045–1064 (2001).

147. Schmidhuber, J. Learning to control fast-weight memories: an alternative to dynamic recurrent networks. *Neural Comput.* **4**, 131–139 (1992).

148. Olshausen, B. A., Anderson, C. H. & Essen, D. V. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719 (1993).

149. Anderson, C. H. & Van Essen, D. C. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Natl Acad. Sci. USA* **84**, 6297–6301 (1987).

150. Burak, Y., Rokni, U., Meister, M. & Sompolinsky, H. Bayesian model of dynamic image stabilization in the visual system. *Proc. Natl Acad. Sci. USA* **107**, 19525–19530 (2010).

151. Salinas, E. & Thier, P. Gain modulation: a major computational principle of the central nervous system. *Neuron* **27**, 15–21 (2000).

152. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735–1780 (1997).

153. Reichert, D. P. & Serre, T. Neuronal synchrony in complex-valued deep networks. In *Proc. 2nd International Conference on Learning Representations* (2014).

154. Gray, C. M. & Singer, W. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl Acad. Sci. USA* **86**, 1698–1702 (1989).

155. Hummel, J. E. & Biederman, I. Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* **99**, 480–517 (1992).

156. Fries, P. Rhythms for cognition: communication through coherence. *Neuron* **88**, 220–235 (2015).

157. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

158. Higgins, I. et al. Towards a definition of disentangled representations. Preprint at http://arxiv.org/abs/1812.02230 (2018).

159. Feldman, J. What is a visual object? *Trends Cogn. Sci.* **7**, 252–256 (2003).

160. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).

161. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *JOSA A* **20**, 1434–1448 (2003).

162. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The Helmholtz machine. *Neural Comput.* **7**, 889–904 (1995).

163. Stuhlmüller, A., Taylor, J. & Goodman, N. Learning stochastic inverses. In *Proc. Advances in Neural Information Processing Systems 26* (eds Burges, C. J. et al.) 3048–3056 (Curran Associates, 2013).

164. Bergen, R. S. van. & Kriegeskorte, N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opinion Neurobiol.* **65**, 176–193 (2020).

165. von der Heydt, R., Friedman, H. S. & Zhou, H. Searching for the neural mechanisms of color filling-in. In *Filling-in: From perceptual completion to cortical reorganization* (eds. Pessoa, L. & De Weerd, P.) 106–127 (Oxford Univ. Press, 2003).

166. Kogo, N. & Wagemans, J. The 'side' matters: how configurality is reflected in completion. *Cogn. Neurosci.* **4**, 31–45 (2013).

167. Craft, E., Schütze, H., Niebur, E. & von der Heydt, R. A neural model of figure-ground organization. *J. Neurophysiol.* **97**, 4310–4326 (2007).

168. Grossberg, S. & Mingolla, E. Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol. Rev.* **92**, 173–211 (1985).

169. Mingolla, E., Ross, W. & Grossberg, S. A neural network for enhancing boundaries and surfaces in synthetic aperture radar images. *Neural Netw.* **12**, 499–511 (1999).

170. Zhaoping, L. Border ownership from intracortical interactions in visual area V2. *Neuron* **47**, 143–153 (2005).

171. Fukushima, K. Neural network model for completing occluded contours. *Neural Netw.* **23**, 528–540 (2010).

172. Tu, Z. & Zhu, S.-C. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Anal.* **24**, 657–673 (2002).

173. Fukushima, K. Restoring partly occluded patterns: a neural network model. *Neural Netw.* **18**, 33–43 (2005).

174. Lücke, J., Turner, R., Sahani, M. & Henniges, M. Occlusive components analysis. In *Proc. Advances in Neural Information Processing Systems 22* (eds Bengio, Y. et al.) 1069–1077 (Curran Associates, 2009).

175. Johnson, J. S. & Olshausen, B. A. The recognition of partially visible natural objects in the presence and absence of their occluders. *Vis. Res.* **45**, 3262–3276 (2005).

176. Koch, C. & Ullman, S. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience* (ed. Vaina, L. M.) 115–141 (Springer, 1987).

177. Tsotsos, J. K. et al. Modeling visual-attention via selective tuning. *Artif. Int.* **78**, 507–545 (1995).

178. Walther, D. & Koch, C. Modeling attention to salient proto-objects. *Neural Netw.* **19**, 1395–1407 (2006).

179. Kazanovich, Y. & Borisyuk, R. An oscillatory neural model of multiple object tracking. *Neural Comput.* **18**, 1413–1440 (2006).

180. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* https://doi.org/10.1038/s41593-021-00821-9 (2021).

181. Barak, O. & Tsodyks, M. Working models of working memory. *Curr. Opin. Neurobiol.* **25**, 20–24 (2014).

182. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. Neurocomputational models of working memory. *Nat. Neurosci.* **3**, 1184–1191 (2000).

183. Compte, A. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* **10**, 910–923 (2000).

184. Wang, X.-J. Synaptic reverberations underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).

185. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neurosci.* **17**, 431–439 (2014).

186. Zenke, F., Agnes, E. J. & Gerstner, W. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nat. Commun.* **6**, 6922 (2015).

187. Mareschal, D., Plunkett, K. & Harris, P. A computational and neuropsychological account of object-oriented behaviours in infancy. *Developmental Science* **2**, 306–317 (1999).

188. Munakata, Y., Mcclelland, J. L., Johnson, M. H. & Siegler, R. S. Rethinking infant knowledge: toward an adaptive process account of successes and failures in object permanence tasks. *Psychol. Rev.* **104**, 686–713 (1997).

189. Mi, Y., Katkov, M. & Tsodyks, M. Synaptic correlates of working memory capacity. *Neuron* **93**, 323–330 (2017).

190. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).

191. Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* **22**, 1159–1167 (2019).

192. Chatham, C. H. & Badre, D. Multiple gates on working memory. *Curr. Opin. Behav. Sci.* **1**, 23–31 (2015).

193. Frank, M. J., Loughry, B. & O'Reilly, R. C. Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive* **1**, 137–160 (2001).

194. Gruber, A. J., Dayan, P., Gutkin, B. S. & Solla, S. A. Dopamine modulation in the basal ganglia locks the gate to working memory. *J. Comput. Neurosci.* **20**, 153–166 (2006).

195. O'Reilly, R. C. Biologically based computational models of high-level cognition. *Science* **314**, 91–94 (2006).

196. Ciresan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. In *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition* 3642–3649 (IEEE, 2012).

197. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).

198. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

199. Zhou, B., Bau, D., Oliva, A. & Torralba, A. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 2131–2145 (2019).

200. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

201. Körding, K. P. & König, P. Supervised and unsupervised learning with two sites of synaptic integration. *J Comput Neurosci* **11**, 207–215 (2001).

202. Guerguiev, J., Lillicrap, T. P. & Richards, B. A. Towards deep learning with segregated dendrites. *eLife* **6**, e22901 (2017).

203. Scellier, B. & Bengio, Y. Equilibrium propagation: bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* **11**, 24 (2017).

204. Roelfsema, P. E. & van Ooyen, A. Attention-gated reinforcement learning of internal representations for classification. *Neural Comput.* **17**, 2176–2214 (2005).

205. Crick, F. The recent excitement about neural networks. *Nature* **337**, 129–132 (1989).

206. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).

207. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence* 4278–4284 (2017).

208. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).

209. Güçlü, U. & Gerven, M. A. J. V. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).

210. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).

211. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Ann. Rev. Vis. Sci.* **1**, 417–446 (2015).

212. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

213. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).

214. Brendel, W. & Bethge, M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *Proc. 7th International Conference on Learning Representations* (OpenReview.net, 2019).

215. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision 2017* 2980–2988 (IEEE, 2017).

216. Pinheiro, P. O., Collobert, R. & Dollar, P. Learning to segment object candidates. In *Proc. Advances in Neural Information Processing Systems 28* (eds Cortes, C. et al.) 1990–1998 (Curran Associates, 2015).

217. Luo, W. et al. Multiple object tracking: a literature review. *Artificial Intelligence* **293**, 103448 (2021).

218. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 580–587 (IEEE, 2014).

219. Bisley, J. W. & Goldberg, M. E. Attention, intention, and priority in the parietal lobe. *Ann. Rev. Neurosci.* **33**, 1–21 (2010).

220. Burgess, C. P. et al. MONet: unsupervised scene decomposition and representation. Preprint at http://arxiv.org/abs/1901.11390 (2019).

221. Locatello, F. et al. Object-centric learning with slot attention. In *Proc. Advances in Neural Information Processing Systems 33* (eds Larochelle, H. et al.) 11525–11538 (Curran Associates, 2020).

222. Eslami, S. M. A. et al. Attend, infer, repeat: fast scene understanding with generative models. In *Proc. Advances in Neural Information Processing Systems 29* (eds Lee, D. et al.) (Curran Associates, 2016).

223. Wu, J., Lu, E., Kohli, P., Freeman, B. & Tenenbaum, J. Learning to see physics via visual de-animation. In *Proc. Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) (Curran Associates, 2017).

224. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychology* **8**, 1551 (2017).

225. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Proc. Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) (Curran Associates, 2019).

226. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl Acad. Sci. USA* **116**, 21854–21863 (2019).

227. Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Comput. Biol.* **16**, e1008215 (2020).

228. O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B. & Jilk, D. J. Recurrent processing during object recognition. *Front. Psychol.* **4**, 124 (2013).

229. Wyatte, D., Jilk, D. J. & O'Reilly, R. C. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.* **5**, 674 (2014).

230. Linsley, D., Kim, J. & Serre, T. Sample-efficient image segmentation through recurrence. Preprint at https://arxiv.org/abs/1811.11356v3 (2018).

231. Engelcke, M., Kosiorek, A. R., Jones, O. P. & Posner, I. GENESIS: generative scene inference and sampling witho object-centric latent representations. In *Proc. 8th International Conference on Learning Representations* (OpenReview.net, 2020).

232. Steenkiste, S. van, Chang, M., Greff, K. & Schmidhuber, J. Relational neural expectation maximization: unsupervised discovery of objects and their interactions. In *Proc. 6th International Conference on Learning Representations* (OpenReview.net, 2018).

233. Greff, K. et al. Multi-object representation learning with iterative variational inference. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) 2424–2433 (PLMR, 2019).

234. Swan, G. & Wyble, B. The binding pool: a model of shared neural resources for distinct items in visual working memory. *Atten. Percept. Psychophys.* **76**, 2136–2157 (2014).

235. Schneegans, S. & Bays, P. M. Neural architecture for feature binding in visual working memory. *J. Neurosci.* **37**, 3913–3925 (2017).

236. Matthey, L., Bays, P. M. & Dayan, P. A probabilistic palimpsest model of visual short-term memory. *PLoS Comput. Biol.* **11**, e1004003 (2015).

237. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. In *Proc. Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) (Curran Associates, 2017).

238. Xu, Z. et al. Unsupervised discovery of parts, structure, and dynamics. In *Proc. 7th International Conference on Learning Representations* (OpenReview.net, 2019).

239. Kosiorek, A., Sabour, S., Teh, Y. W. & Hinton, G. E. Stacked capsule autoencoders. In *Proc. Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) (Curran Associates, 2019).

240. Pelli, D. G. & Tillman, K. A. The uncrowded window of object recognition. *Nat. Neurosci.* **11**, 1129–1135 (2008).

241. Sayim, B., Westheimer, G. & Herzog, M. Gestalt factors modulate basic spatial vision. *Psychol. Sci.* **21**, 641–644 (2010).

242. Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M. & Herzog, M. H. Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology* **16**, 1–19 (2020).

243. Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. Preprint at http://arxiv.org/abs/1806.01261 (2018).

244. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **20**, 61–80 (2009).

245. Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F. & Niebles, J. C. Learning to decompose and disentangle representations for video prediction. In *Proc. Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) 515–524 (Curran Associates, 2018).

246. Whittington, J. C. R. et al. The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263 (2020).

247. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT press, 2018).

248. Botvinick, M. et al. Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).

249. LeCun, Y. The power and limits of deep learning. *Res. Technol. Manage.* **61**, 22–27 (2018).

250. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *Proc. 2nd International Conference on Learning Representations* (eds. Bengio, Y. & LeCun, Y.) (openreview.net, 2014).

251. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. 31st International Conference on Machine Learning 32* (eds. Xing, E. P. & Jebara, T.) 1278–1286 (PMLR, 2014).

252. Goodfellow, I. et al. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems 27* (eds. Ghahramani, Z. et al.) (Curran Associates, 2014).

253. Weis, M. A. et al. Unmasking the inductive biases of unsupervised object representations for video sequences. Preprint at http://arxiv.org/abs/2006.07034 (2020).

254. Veerapaneni, R. et al. Entity abstraction in visual model-based reinforcement learning. In *Proc. Conference on Robot Learning 100* (eds. Kaelbling, L. P. et al.) 1439–1456 (PMLR, 2020).

255. Watters, N., Tenenbaum, J. & Jazayeri, M. Modular object-oriented games: a task framework for reinforcement learning, psychology, and neuroscience. Preprint at http://arxiv.org/abs/2102.12616 (2021).

256. Leibo, J. Z. et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. Preprint at http://arxiv.org/abs/1801.08116 (2018).

257. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).

258. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012).

259. Sullivan, J., Mei, M., Perfors, A., Wojcik, E. & Frank, M. C. SAYCam: a large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind* **5**, 20–29 (2021).

260. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).

261. Green, D. M., & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, 1966).

262. Rust, N. C. & Movshon, J. A. In praise of artifice. *Nat. Neurosci.* **8**, 1647–1650 (2005).

263. Wu, M. C.-K., David, S. V. & Gallant, J. L. Complete functional characterization of sensory neurons by system identification. *Ann. Rev. Neurosci.* **29**, 477–505 (2006).

264. Ullman, S. Visual routines. *Cognition* **18**, 97–159 (1984).

265. Jolicoeur, P., Ullman, S. & Mackay, M. Curve tracing: a possible basic operation in the perception of spatial relations. *Mem. Cogn.* **14**, 129–140 (1986).

266. Ballard, D. H., Hayhoe, M. M., Pook, P. K. & Rao, R. P. N. Deictic codes for the embodiment of cognition. *Behav. Brain Sci.* **20**, 723–742 (1997).

267. Geirhos, R. et al. Generalisation in humans and deep neural networks. In *Proc. Advances in Neural Information Processing Systems 31* (eds. Bengio, S. et al.) (Curran Associates, 2018).

268. Barbu, A. et al. ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proc. Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) (Curran Associates, 2019).

269. Blaser, E., Pylyshyn, Z. W. & Holcombe, A. O. Tracking an object through feature space. *Nature* **408**, 196–199 (2000).

270. Johansson, G. Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211 (1973).

271. Schrimpf, M. et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).

272. Judd, T., Durand, F. & Torralba, A. *A Benchmark of Computational Models of Saliency to Predict Human Fixations* Technical Report (MIT, 2012).

273. Kümmerer, M., Wallis, T. S. A., Gatys, L. A. & Bethge, M. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision* 4799–4808 (IEEE, 2017).

274. Ma, W. J. & Peters, B. A neural network walks into a lab: towards using deep nets as models for human behavior. Preprint at http://arxiv.org/abs/2005.02181 (2020).

275. Peterson, J., Battleday, R., Griffiths, T. & Russakovsky, O. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 9616–9625 (IEEE, 2019).

276. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L. & Girshick, R. PHYRE: a new benchmark for physical reasoning. In *Proc. Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) (Curran Associates, 2019).

277. Yi, K. et al. CLEVRER: collision events for video representation and reasoning. In *Proc. 8th International Conference on Learning Representations* (OpenReview.net, 2020).

278. Riochet, R. et al. IntPhys: a framework and benchmark for visual intuitive physics reasoning. *CoRR*, abs/1803.07616 (2018).

279. Baradel, F., Neverova, N., Mille, J., Mori, G. & Wolf, C. CoPhy: counterfactual learning of physical dynamics. In *Proc. 8th International Conference on Learning Representations* (OpenReview.net, 2020).

280. Girdhar, R. & Ramanan, D. CATER: A diagnostic dataset for compositional actions & temporal reasoning. In *Proc. 8th International Conference on Learning Representations* (OpenReview.net, 2020).

281. Allen, K. R., Smith, K. A. & Tenenbaum, J. B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl Acad. Sci. USA* **117**, 29302–29310 (2020).

282. Beyret, B. et al. The Animal-AI environment: training and testing animal-like artificial cognition. Preprint at http://arxiv.org/abs/1909.07483 (2019).

283. Kanizsa, G. Margini quasi-percettivi in campi con stimolazione omogenea. *Riv. Psicol.* **49**, 7–30 (1955).

284. Kanizsa, G. Amodale ergänzung und 'erwartungsfehler' des gestaltpsychologen. *Psychol. Forsch.* **33**, 325–344 (1970).

285. Eslami, S. M. A. et al. Neural scene representation and rendering. *Science* **360**, 1204–1210 (2018).

286. Beattie, C. et al. {DeepMind} {Lab}. Preprint at http://arxiv.org/abs/1612.03801 (2016).

287. Kolve, E. et al. AI2-THOR: an interactive 3D environment for visual AI. Preprint at https://arxiv.org/abs/1712.05474 (2017).

288. Lin, T.-Y. et al. Microsoft COCO: common objects in context. In *Computer Vision—ECCV 2014, Lecture Notes in Computer Science* (eds Fleet, D. et al.) 740–755 (Springer, 2014).

289. Milan, A., Leal-Taixe, L., Reid, I., Roth, S. & Schindler, K. MOT16: a benchmark for multi-object tracking. Preprint at http://arxiv.org/abs/1603.00831 (2016).

290. Mahler, J. et al. Learning ambidextrous robot grasping policies. *Sci. Robot.* **4**, eaau4984 (2019).

291. Pitkow, X. Exact feature probabilities in images with occlusion. *J. Vis.* **10**, 42 (2010).

292. O'Reilly, R. C., Busby, R. S. & Soto, R. Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In *The unity of consciousness: Binding, integration, and dissociation* (ed. Cleeremans, A.) 168–190 (Oxford Univ. Press, 2003).

293. Hummel, J. E. et al. A solution to the binding problem for compositional connectionism. In *AAAI Fall Symposium - Technical Report* (eds. Levy, S. D. & Gayler, R.) vol. FS-04-03 31–34 (AAAI Press, 2004).

294. Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. Distributed Representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (eds. Rumelhart, D. E. & McClelland, J. L.) 77–109 (MIT Press, 1987).

295. Treisman, A. Solutions to the binding problem: progress through controversy and convergence. *Neuron* **24**, 105–125 (1999).

296. Ballard, D. H., Hinton, G. E. & Sejnowski, T. J. Parallel visual computation. *Nature* **306**, 21–26 (1983).

297. Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Int.* **46**, 159–216 (1990).

298. Feldman, J. A. Dynamic connections in neural networks. *Biol. Cybernet.* **46**, 27–39 (1982).

299. Von Der Malsburg, C. Am I thinking assemblies? *Brain Theory* 161–176 (1986)

300. Reynolds, J. H. & Desimone, R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24**, 19–29 (1999).

301. Shadlen, M. N. & Movshon, J. A. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* **24**, 67–77 (1999).

Acknowledgements

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Benjamin Peters or Nikolaus Kriegeskorte.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.