Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Current Opinion in
Neurobiology

An integrative computational architecture for object-driven cortex

Ilker Yildirim[1,4], Jiajun Wu[1,3], Nancy Kanwisher[1,2,4] and Joshua Tenenbaum[1,2,3,4]

Computational architecture for object-driven cortex

Objects in motion activate multiple cortical regions in every lobe of the human brain. Do these regions represent a collection of independent systems, or is there an overarching functional architecture spanning all of object-driven cortex? Inspired by recent work in artificial intelligence (AI), machine learning, and cognitive science, we consider the hypothesis that these regions can be understood as a coherent network implementing an integrative computational system that unifies the functions needed to perceive, predict, reason about, and plan with physical objects—as in the paradigmatic case of using or making tools. Our proposal draws on a modeling framework that combines multiple AI methods, including causal generative models, hybrid symbolic-continuous planning algorithms, and neural recognition networks, with object-centric, physics-based representations. We review evidence relating specific components of our proposal to the specific regions that comprise object-driven cortex, and lay out future research directions with the goal of building a complete functional and mechanistic account of this system.

Addresses
[1] Center for Brains, Minds, and Machines, MIT, Cambridge, MA 02138, United States
[2] McGovern Institute for Brain Research, MIT, Cambridge, MA 02138, United States
[3] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02138, United States
[4] Department of Brain & Cognitive Science, MIT, Cambridge, MA 02138, United States

Corresponding author: Yildirim, Ilker (ilkery@mit.edu)

Current Opinion in Neurobiology 2019, **55**:73–81

This review comes from a themed issue on **Machine learning, big data, and neuroscience**

Edited by **Jonathan Pillow** and **Maneesh Sahani**

For a complete overview see the Issue and the Editorial

Available online 28th February 2019

https://doi.org/10.1016/j.conb.2019.01.010

Introduction

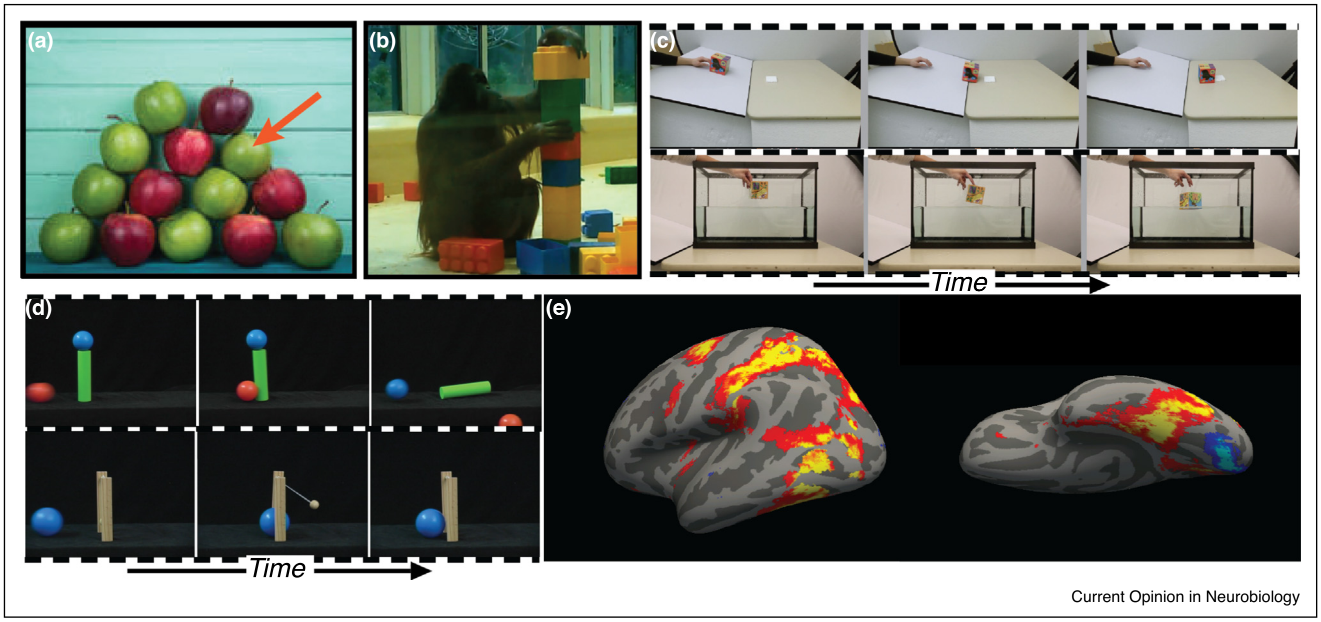
Many everyday activities revolve around objects—seeing, reasoning about, planning with, and manipulating them—in flexible and often creative ways. We see an object's three-dimensional (3D) shape and appearance; we perceive or reason about how it supports or is supported by other objects and surfaces (Figure 1a); when it moves, we track and predict its position and infer its physical properties (e.g. mass) (Figure 1c). These percepts support planning and production of complex motor behaviors (Figure 1b): We reach, grasp, push, pull, pick up, stack, balance, cut, throw, or sit on objects.

Commensurate with the centrality of objects in perception and cognition, large and diverse regions of the human brain are driven by dynamic object stimuli (e.g. Ref. [1]) compared to scrambled versions of the stimuli (Figure 1d, e). These regions include the traditional object-selective occipitotemporal regions (e.g. Ref. [2]), such as the lateral occipital cortex (LOC) and posterior fusiform (pFus), as well as regions in the intraparietal sulcus [3••,4,5,6•] and frontal cortex that show large overlaps with networks implicated in tool use and action planning [3••]. Presumably, these different regions process dynamic objects in different ways and for different functional purposes [7••]. But is there also a unified function that all regions, working together, might subserve?

Here, we present a computational hypothesis for the integrated function of these brain regions, which we collectively refer to as 'object-driven cortex' (Figure 1e). Our proposed architecture integrates the computations involved in seeing an object at an initial glance, tracking it dynamically as it moves, updating estimates of its physical properties based on its motion, reasoning about its likely and possible future behaviors, contingent on forces applied, and planning actions toward it to achieve goals. This hypothesis draws on and extends recent work in the fields of cognitive science, artificial intelligence (AI), and machine learning (ML), bringing together causal generative models, neural networks for efficiently approximating Bayesian inferences in those models, and hybrid task-and-motion planning algorithms to explain how humans understand and interact with physical scenes, and how robots might do the same.

The expanse of activations comprising object-driven cortex overlaps with cortical regions that have been discussed extensively in other theoretical contexts. These include the multiple demand network [8], and cortical systems engaged in numerical cognition [9], object-directed action [3••], logical reasoning [10], and action

Figure 1



(a) We can predict whether the pile would topple if the indicated apple were removed, and readily plan how to pick it up without making the rest unstable. **(b)** Some of these abilities are likely shared across other species, particularly non-human primates. Snapshot is extracted from https://www.youtube.com/watch?v=7GiQkxsje5c. **(c)** In some dynamic scenes, unfolding motion reveals physical object properties (e.g. mass; [49••,67]). **(d)** Example dynamic stimuli used in fMRI experiments (from Ref. [1]). **(e)** Group-level random-effects analysis of the contrast of viewing dynamic objects > scrambled objects ($N = 52$; p-values range from 0.001 to 10^{-7}, red to yellow).

emulation [11•]. Here, we consider a particular end-goal or functionality of this system, that of ‘object cognition’, encompassing the computations underlying how we see, think about, and manipulate objects. This framework may ultimately subsume or reduce to other proposals for functional interpretations of these regions; how our framework relates to prior proposals is an important question and we cannot hope to comprehensively review that literature here. Our goal is simply to take the initial step of articulating a framework for understanding the neural basis of object cognition in precise and interpretable functional terms, which we hope will spur further thinking and empirical work.

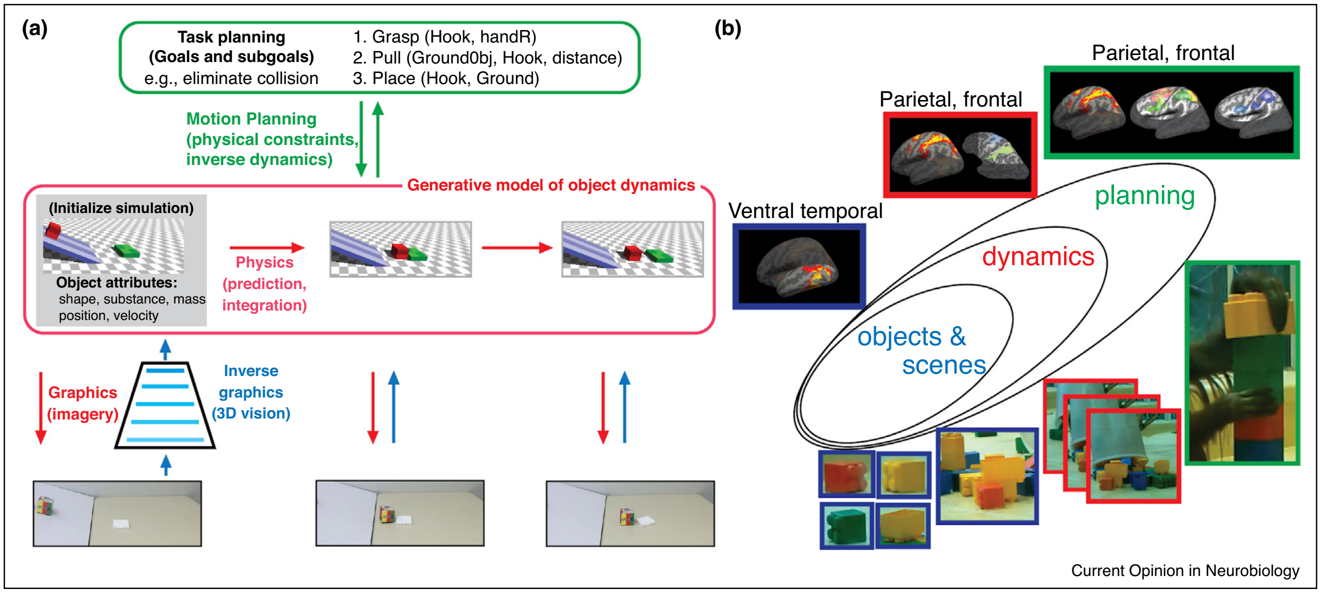
We focus on three main components of our computational architecture — generative models for simulating object dynamics, planning algorithms that use these generative models together with simulatable body models to construct action plans, and recognition models for efficiently perceiving the properties of objects critical to their dynamics — and discuss evidence linking each component to specific regions of object-driven cortex. We conclude with a discussion of future research directions.

Physical scene understanding via causal generative models

Scene understanding entails not just recognizing what objects are where, but reasoning about their physical dynamics and relations. We see not only one thing on top of another, but the fact that one is *supporting* the other; this includes whether objects are stably balanced or likely to fall, and if one falls, which way it is likely to fall. If an object does not fall as expected, we may infer it has a different mass or mass distribution than we first thought. What computations support such intuitive physical reasoning abilities?

The first component of our computational architecture addresses this challenge using generative models of physical object representations and their dynamics. Specifically, we have implemented these models in probabilistic extensions of video game engine [12] components, especially graphics engines and physics engines [13–15], which instantiate our basic knowledge of how objects work in simplified but algorithmically efficient simulators. In these systems, objects are described by just those attributes needed to simulate natural-looking motion over short time scales (~2 s): their geometry (shape, size), and the material properties that govern their dynamics (e.g. rigidity, mass, surface friction). Game-engine physics instantiates a causal generative model for object motion in the sense that the mechanisms by which motion trajectories are generated have some abstract resemblance to the corresponding real-world processes — but in a form that is efficient enough to support real-time interactive simulation. A diagram of

Figure 2



Current Opinion in Neurobiology

(a) A schematic of our integrative computational architecture. The architecture consists of three elements: (red) generative models of object dynamics and image formation implemented using physics and graphics engines, (green) planners to compute actions that achieve goals, subject to physical and geometric constraints, and (blue) recognition models for online perception (inverse graphics). The generative model enables not only predictions about the near-term future states of objects but also integration of motion and interactions for dynamic updates to physical object properties such as an object's mass. It can also support a form of visual imagery through its graphics components. The planner, given a goal, enables sequencing of action primitives based on the constraints arising from physics and geometry for complex object manipulation tasks including tool use. Inverse graphics maps individual frames to 3D physical scene descriptions, the core latent variables of the generative model. **(b)** A schematic summary of the mappings between our computational architecture and object-driven cortex. The highlighted regions in the insets above overlap with regions that are involved in object perception (blue; e.g. Ref. [2]), physical reasoning (right panel in red; [26••]), and action planning (middle panel in green; [3••]) and tool use (right panel in green; [3••]).

such a generative model is shown in Figure 2a (red rectangle).

Battaglia *et al.* proposed such a model, which they called an 'intuitive physics engine', as an account of physical scene understanding [16••]. They showed how approximate probabilistic inferences over simulations in a game-style physics engine could explain how people perform a wide variety of tasks in blocks-world scenes, including both familiar tasks (e.g. Will this tower fall? Which way will it fall?), and novel tasks in novel scenarios (e.g. If a table supporting a complex configuration of blocks is bumped, which of these blocks might fall off the table?). Humans can perform these tasks with little or no training, and the ability to do so is a key advantage of generative models over pattern recognition approaches such as neural network classifiers [17]. Subsequent work has shown how the framework extends more broadly across many aspects of intuitive physics, including predictions of future motion for rigidly colliding objects [18,19], predictions about the behavior of liquids (e.g. water, honey) [20,21] and granular materials (e.g. sand) [22,23], and judgments about objects' dynamic properties and interactions from how they move under gravity as well as latent forces such as magnetism [24,25].

Do parts of object-driven cortex contribute to intuitive physical reasoning? Recent imaging work in humans identified a network of parietal and premotor regions that are activated more by these same kinds of physical reasoning tasks (e.g. Where will a tower of blocks fall?) than non-physical tasks (e.g. Are there more blue or yellow blocks in the tower?). These regions overlap substantially with parts of object-driven cortex in parietal and frontal regions [26••]. Further support for a link comes from an fMRI study with macaque monkeys. Sliwa and Freiwald [27•] found that passive viewing of videos of interacting objects compared to still or non-interacting objects selectively activated parietal and pre-motors regions. These results are suggestive of a neural physics engine implemented in this network of regions across parietal and frontal cortex.

Planning with physical and geometric constraints

Why would the brain devote circuitry for predicting object dynamics and interactions, and why would that circuitry overlap regions involved in action planning and tool use? One hypothesis comes from the recent robotics literature, where it has been argued that modeling and exploiting constraints from the geometry and physics of objects is essential for flexible action planning in robots that will interact with objects in human-like ways (e.g. Refs. [28–30]). For example, stacking a tower requires sensing, predicting, and maintaining its stability (Figure 1b), grasping a cup requires knowing and reacting to its weight and the slipperiness of its surface, and reaching for a far object using a hook requires knowing about the constraints imposed by the layout of objects and their geometries and how they interact. These requirements are in addition to the need for a simulatable body model (similar to the forward models proposed in the motor control literature, e.g. Refs. [31–33]) that can be used by embodied agents to foresee and evaluate the consequences of their actions on objects before actually performing them (see Ref. [34] for a discussion in the context of mammalian somatosensory cortex).

Some of the most advanced humanoid robot motion planners, or hybrid task and motion planners, combine physics-engine representations of object dynamics with a simulatable body model or part of a body model (e.g. an articulated hand), and aim to jointly and efficiently solve for effective action sequences subject to the physical constraints of object dynamics and interactions [e.g. Refs. 35••,36,37••]. The use of differentiable physics engines allows these systems to support gradient-based optimization for efficient model-based control. These planners can generate remarkably complex and human-like sequences of action, including improvised use of objects as tools to accomplish non-trivial tasks, such as reaching for a small hook, which can then be used to retrieve a large hook, which can then be used to retrieve an otherwise out-of-reach goal object. Yildirim *et al.* [38•] showed that such a planning framework not only produces physically stable simulated solutions in tower-building scenarios (e.g. re-configure a tower), but also matches human intuitions on how to build the target tower. These results support the idea that reasoning about geometry and physics facilitate planning complex motor actions and using tools.

If the brain adopts similar mechanisms for flexible action planning, that could explain why the network of physical reasoning regions described in the previous section appears to closely overlap with regions involved in motor planning and tool use in humans [3••], and with the mirror neuron network in monkeys that is thought to be involved in action understanding [39]. These parietal and premotor regions might implement a planning system based on simulatable body models and object models, encoding physical and geometric constraints in something like the form of a physics engine, as in analogous robotics systems. In AI, the same physics engine-based systems that support these object-directed action plans, such as Mujoco [29], can be used (and frequently are used) for efficient approximate simulation of complex multi-object interactions even in the absence of any body model or action planning task; the same could be true of the human brain.

This proposed architecture for how physical object representations, simulations, and action planning are integrated in the parietal and premotor regions of object-driven cortex is consistent with the similar notion of an 'emulated action system' that has been proposed as a functional account of a different but overlapping network, the dorsal frontoparietal network [11•]. The specific models we propose, however, are intended to offer a concrete computational framework that articulates the functional components required for action planning and how they might interact with each other, in the more general context of perceiving, planning, and thinking about the actual or possible motions of objects.

Perception and dynamic belief updates with recognition models

A key observation [26••] is that passive viewing of objects in motion not only activates the traditional visual and ventral pathway regions but also strongly drives activity in physical reasoning regions in parietal and premotor cortex (see also Figure 1e). This finding suggests that when presented with structured dynamic visual input, the brain not only constructs rich 3D scenes of objects and surfaces, but also, akin to the construction of object files (e.g. Refs. [40,41]), automatically and in an online fashion tracks and updates objects' physical properties based on how they move and interact. How can the brain so efficiently estimate and update rich physical representations of objects during online perception?

From a Bayesian viewpoint, physical object properties, including 3D shape, size, mass, friction, stiffness or other parameters required for physical reasoning, are latent variables in a probabilistic generative model that need to be inferred and dynamically updated given changing sensory inputs [25,26••,42]. The most familiar mechanisms for performing these Bayesian inferences in complex structured generative models are approximate ones, based on sequential, stochastic sampling methods such as Markov Chain Monte Carlo (or MCMC). These methods can work given enough time, but they seem implausible as algorithmic accounts of perception in the brain: they are inherently iterative and almost always far too slow relative to the dynamics of perception.

Recently, researchers have begun to answer these challenges of efficient scene perception and dynamic belief updates by building recognition models that exploit the causal (conditional independence) structure within a model of the generative process. These recognition models can be constructed using modern neural networks such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and are trained to directly estimate maximum a posteriori (MAP) values for the latent variables in generative models in a data-driven and efficient manner. Many of these architectures focus on perceiving the geometry of objects or scenes (e.g. Refs.

[43–47]), but some integrate 3D geometry with physical dynamics.

In one recent example, Wu *et al.* [48••] built such a system, referred to as 'visual de-animation', that uses a cascade of CNNs for inverse graphics, including segmenting input images to individual objects from an initial video frame (a form of attention), and mapping each segmented object to its full set of physical object properties (Figure 2a; inverse graphics). The system uses motion information across consecutive frames to train its CNN based on an efficient inference procedure: given a dataset of unlabeled videos, the system infers an initial scene configuration, extrapolates its motion over time using a physics engine, and renders individual predicted frames using a graphics engine, with the goal of minimizing reconstruction errors with respect to the input video. Wu *et al.* showed that once trained, the network can be used in prediction and reasoning tasks across both simulated and real-world scenarios with a variable number of objects, solely from visual input and in real time. The model can support predictions such as near-future configurations of billiard tables, or can be used to plan interventions such as applying a force to stabilize an unstable block tower.

The visual de-animation system and its predecessors (e.g. Galileo [49••]) assume that the underlying intrinsic physical object properties (e.g. shape, mass, and friction) are fixed, and do not address dynamic belief updating behavior. To answer this challenge, Yildirim *et al.* [50••] built a recurrent recognition network based on the overall conditional independence structure in the underlying generative model (Figure 2a). Through a combination of supervised and unsupervised training, this recognition network learns to implement approximate Bayesian estimates about the values of key physical variables conditioned on dynamic input stimuli (e.g. videos), by compiling inference [51] in the generative model to a set of neuron-like computations (a cascade of CNNs and RNNs). The model dynamically updates latent physical object properties with each incoming frame of input, and also learns to attend to the relevant regions in the image (e.g. collision regions when two objects are about to collide). The model accurately captures human belief updating patterns in a relative-mass judgment task, and corresponds more closely to human judgments than an ideal observer model, suggesting it might also capture some of the dynamic cognitive processes underlying performance in the task.

The inverse graphics component of these recognition models (that is, the cascade of CNNs transforming images to 3D scenes) implements a functionality that most naturally maps to the ventral pathway computations including parts of the visual cortex and ventral temporal cortex. Abstract scene information such as identity, shape, and position of objects becomes more explicit through this processing hierarchy of the ventral pathway, particularly in its middle and later stages [2,52,53]. But ventral processing is only the first stage in object cognition, and in the typical dynamics of object-driven cortex. In the recognition models discussed here, physical properties of objects are fed to the physics engine for integration of information across time (e.g. updating beliefs about an object's mass), future prediction, and reasoning (e.g. computing the force to apply to keep a tower stable). In line with this computational pipeline, recent brain imaging work suggests that abstract object information such as shape is 'uploaded' from ventral pathway to regions in the parietal cortex [4,5,6•] where it may adaptively support aspects of cognition and action [54]. If, in addition to shape, visually computed representations of objects' dynamic physical properties such as their mass are uploaded from ventral stream to an intuitive physics engine in parietal and premotor cortex, then we should expect to see representations of these properties in those regions. Schwettmann *et al.* [55••] recently found exactly that: Object mass can be decoded from the parietal and frontal physical reasoning regions [26••] in a manner invariant to the specifics of how objects are visually presented. For example, an object's mass could be decoded from the brain's response to viewing it splash into a bowl of water after training on viewing that same object falling onto a pillow, or vice versa.

Discussion

Here we have proposed a reverse-engineering account of the functions of object-driven cortex, including its components in the ventral pathway and parietal/pre-motor regions, and how these components interact in dynamic object perception and in making plans directed toward objects (Figure 2b). At its core, our proposal is a hypothesis that the targets of perception are not just object shapes or action affordances, but physical object representations that are the key elements of causal generative models — models of how objects move and interact, and how we can move and interact with them to achieve our goals. These representations are engaged and updated automatically, in a bottom-up fashion using recognition networks that are driven through visual inputs. These representations natively support thinking about relations, motions, and interactions of objects; and they facilitate planning complex sequences of actions toward objects and tool use. Neural data consistent with our hypothesis include the overlap of object-driven cortex, regions involved in thinking about the physics of objects [26••], and regions involved in object-directed action [3••], and the characteristics of how visual information propagates from ventral to dorsal streams [4,54], allowing physical variables such as mass to be decoded from parietal and frontal regions based on activity arising from passive viewing [55••].

We should be clear about what we are not claiming in advancing this hypothesis. We do not mean to suggest that object perception, dynamic prediction, and action planning are not distinct computations, or are not implemented in distinct, potentially modular brain systems. Much evidence suggests that they are distinct in these ways. And yet from a functional point of view, these different computational components must work together to support flexible everyday engagement with objects. They must, in some sense, also form a functionally integrated system, likely with some shared representational substrate. Here we have tried to lay out what that integrated system could look like architecturally, how it could work computationally, drawing on recent advances in AI and machine learning, and how these computations might be implemented in a network of brain regions which are all engaged automatically when seeing physical objects in motion.

Having laid out this proposal, many questions arise. On the modeling side, the most urgent questions revolve around building neurally plausible versions of richly structured generative models, such as physics engines, graphics engines and body planning models. Recent developments in machine learning and perception suggest several possibilities, based on deep learning systems trained to emulate structured generative models (e.g. Refs. [56,57,58•,59]). These neural networks provide partial hypotheses for how graphics and physics might be implemented in neural circuits; they are surely incomplete, at best, and much more work is needed here. Crucially, while these networks learn distributed representations of force dynamics, they all invoke discrete, symbolic representations of objects and their interactions (like nodes and edges in a graph), just as in conventional physics engines or cognitive architectures based on object files [40,41]. Whether and how such graph-like representations are implemented in the brain are questions of great interest.

Relating our proposal to conventional models of visual perception is another priority. Our architecture naturally supports a range of functions that are difficult to account for if we treat object perception as primarily the computations of a feedforward network in the ventral stream [60,61]; these include mental imagery, top–down context effects, and multisensory/crossmodal perception. Mental imagery can be implemented in a generative model that couples a physics engine to planning algorithms that support amodal reasoning and to a graphics engine that produces visual imagery, as in the visual de-animation model [48••]. In addition, aspects of multisensory perception and crossmodal transfer can be modeled by composing causal generative models for multiple sensory modalities that share the same underlying latent variables — those represented in the physics engine. Most of these extensions of our framework have been implemented computationally in some form, and received some behavioral support [62–64], but it is an open question whether or how these computations might be instantiated in object-driven cortex.

Another important goal is to explore further how the computational architecture presented here connects to existing theoretical accounts of the parietal–frontal regions and their interactions [3,8–10]. At a basic level, our framework can provide several of the building blocks needed by these systems for their more mathematical and computational formulations. For example, in the context of the multiple demand network [8], it is not clear how in functional terms subtasks could be flexibly assembled in neural circuits. Our framework suggests a means to solve one instance of this challenge, in the form of sequencing subgoals for tool use and complex object manipulation. We hope that further articulation and study of our framework could simultaneously advance a mechanistic account of the multiple demand network.

Perhaps the most important goal for future research will be to empirically test and refine predictions of our hypothesis. What exactly is represented in each region, and when? Beyond representing the shape of an object [2] and its grasp points [3••], and its mass [55••], does object-driven cortex represent other dynamically relevant physical properties, such as friction, rigidity or elasticity? How are forces that one object exerts on another, or stable relations such as support and containment, represented in neural circuits? Which aspects of physical representations are computed rapidly and automatically, suggesting feedforward mechanisms, and which require more conscious, controlled processing? Are causal generative representations constructed automatically, or only when relevant to the task at hand? What exactly is the division of computational labor (if any) across the regions comprising object cortex, and can this division of labor be understood within the framework proposed here? Although some of these questions can be addressed using fMRI and EEG/MEG in humans, future experimental work using electrophysiological recordings, informed by some of the more neurally grounded models discussed above, can target neural populations in object-driven cortex in greater detail to elucidate neural circuits of object cognition at more fine-grained functional and anatomical resolutions. These can include an understanding of functions and circuits of how object files are created and updated in neural populations, how a body model is implemented and simulated, and how these two systems interact with each other.

Finally, if object-driven cortex indeed constitutes a computationally integrated network, then we would expect structural connections between the cortical regions comprising this network. While we do not know of a detailed analysis of the long-range structural connections between these specific object-preferring regions, prior evidence suggests the existence of connections between object-processing regions of the ventral temporal and parietal lobes [65], and extensive structural connections are known to connect parietal and frontal regions in primates (e.g. Ref. [66]).

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank David Beeler for help with data analysis and preparing figures. This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216; ONR MURI N00014-13-1-0333; a grant from Toyota Research Institute; and a grant from Mitsubishi MELCO; and National Institutes of HealthDP1HD091947 to NK.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Julian JB, Fedorenko E, Webster J, Kanwisher N: **An algorithmic method for functionally defining regions of interest in the ventral visual pathway**. *Neuroimage* 2012, **60**:2357-2364.

2. Grill-Spector K, Kourtzi Z, Kanwisher N: **The lateral occipital complex and its role in object recognition**. *Vis Res* 2001, **41**:1409-1422.

3. •• Gallivan JP, Culham JC: **Neural coding within human brain areas involved in actions**. *Curr Opin Neurobiol* 2015, **33**:141-149.

A review of the human brain areas involved in actions and how activity across these regions might support goal-directed behavior.

4. Xu Y: **A tale of two visual systems: invariant and adaptive visual information representations in the primate brain**. *Ann Rev Vis Sci* 2018, **4**:311-336.

5. Vaziri-Pashkam M, Taylor J, Xu Y: **Spatial frequency tolerant visual object representations in the human ventral and dorsal visual processing pathways**. *J Cogn Neurosci* 2018:1-14.

6. • Vaziri-Pashkam M, Xu Y: **An information-driven 2-pathway characterization of occipitotemporal and posterior parietal visual object representations**. *Cereb Cortex* 2018. https://academic.oup.com/cercor/advance-article/doi/10.1093/cercor/bhy080/4969359.

Using a data-driven approach across 10 fMRI experiments, they found posterior parietal regions contain rich non-spatial visual information, similar to the ventral pathway object-selective regions. However, multi-dimensional scaling analysis suggested that posterior and ventral object representations were distinct from each other—multi-voxel patterns of these two pathways clustered in a non-overlapping manner.

7. •• Goodale MA, Humphrey GK: **The objects of action and perception**. *Cognition* 1998, **67**:181-207.

Seminal work analyzing and contrasting functional roles of vision in perceiving versus directing actions towards objects.

8. Duncan J: **The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour**. *Trends Cogn Sci* 2010, **14**:172-179.

9. Dehaene S, Molko N, Cohen L, Wilson AJ: **Arithmetic and the brain**. *Curr Opin Neurobiol* 2004, **14**:218-224.

10. Goel V: **Anatomy of deductive reasoning**. *Trends Cogn Sci* 2007, **11**:435-441.

11. • Ptak R, Schnider A, Fellrath J: **The dorsal frontoparietal network: a core system for emulated action**. *Trends Cogn Sci* 2017, **21**:589-599.

This study characterizes function of the dorsal frontoparietal network as a system for ‘emulated action’, a form of simulation for planning motor behavior. Ptak*et al.* consider emulated action to be a core system that through development grows to support other mental processes more broadly such as attention and cognitive control.

12. Gregory J: *Game Engine Architecture*. AK Peters/CRC Press; 2014.

13. Blender Online Community: *Blender – A 3D Modelling and Rendering Package [Computer Software Manual]*. . Retrieved from Amsterdam: Blender Institute; 2015 In: *http://www.blender.org*.

14. Coumans E: *Bullet Physics Engine*. Open Source Software; 2010 In: *http://bulletphysics.Org*.

15. Macklin M, Müller M, Chentanez N, Kim TY: **Unified particle physics for real-time applications**. *ACM Trans Graph (TOG)* 2014, **33**:153.

16. •• Battaglia PW, Hamrick JB, Tenenbaum JB: **Simulation as an engine of physical scene understanding**. *Proc Natl Acad Sci U S A* 2013 http://dx.doi.org/10.1073/pnas.1306572110.

This paper introduced the idea of an ‘intuitive physics engine’, which akin to a video game physics engine, allows us to approximately but efficiently simulate how the world of objects in our sensory surroundings might unfold. Behavioral experiments across various tasks suggest that human error patterns were best captured using a probabilistic extension of these game-style physics simulation systems.

17. Lerer A, Gross S, Fergus R: **Learning physical intuition of block towers by example**. *Proceedings of the 33rd International Conference on International Conference on Machine Learning* 2016, **48**:430-438.

18. Smith KA, Battaglia P, Vul E: **Consistent physics underlying ballistic motion prediction**. *Annual Meeting of the Cognitive Science Society* 2013, **vol 35**.

19. Smith KA, Dechter E, Tenenbaum JB, Vul E: **Physical predictions over time**. *Annual Meeting of the Cognitive Science Society* 2013, **vol 35**.

20. Bates C, Battaglia P, Yildirim I, Tenenbaum JB: **Humans predict liquid dynamics using probabilistic simulation**. *Annual Meeting of the Cognitive Science Society*. 2015.

21. Bates CJ, Yildirim I, Tenenbaum JB, Battaglia P: *Modeling Human Intuitions about Liquid Flow with Particle-Based Simulation*. . arXiv preprint arXiv:1809.01524 2018.

22. Kubricht J, Jiang C, Zhu Y, Zhu SC, Terzopoulos D, Lu H: **Probabilistic simulation predicts human performance on viscous fluid-pouring problem**. *Annual Meeting of the Cognitive Science Society*. 2016:1805-1810.

23. Kubricht J, Zhu Y, Jiang C, Terzopoulos D, Zhu SC, Lu H: **Consistent probabilistic simulation underlying human judgment in substance dynamics**. *Annual Meeting of the Cognitive Science Society*. 2017:700-705.

24. Ullman TD, Stuhlmüller A, Goodman ND, Tenenbaum JB: **Learning physical parameters from dynamic scenes**. *Cogn Psychol* 2018, **104**:57-82.

25. Hamrick JB, Battaglia PW, Griffiths TL, Tenenbaum JB: **Inferring mass in complex scenes by mental simulation**. *Cognition* 2016, **157**:61-76.

26. •• Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N: **Functional neuroanatomy of intuitive physical inference**. *Proc Natl Acad Sci U S A* 2016, **113**:E5072-E5081.

This study identified brain regions involved in intuitive physical reasoning, finding that several parietal and pre-motor regions selectively activate in a contrast of physical reasoning tasks versus non-physical tasks. They also observed that these regions were activated just from passive viewing of dynamic object stimuli.

27. • Sliwa J, Freiwald WA: **A dedicated network for social interaction processing in the primate brain**. *Science* 2017, **356**:745-749.

In addition to characterizing a network of macaque brain regions for processing social interactions, they also found that a network of parietal and pre-motor regions were selectively activated in a contrast of interacting versus non-interacting objects. These regions showed significant overlap with the macaque mirror-neuron system.

28. Miller AT, Allen PK: **Graspit! a versatile simulator for robotic grasping**. *IEEE Robot Autom Mag* 2004, **11**:110-122.

29. Todorov E, Erez T, Tassa Y: **Mujoco: a physics engine for model-based control**. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012:5026-5033.

30. Toussaint M: **Logic-geometric programming: an optimization-based approach to combined task and motion planning**. *International Joint Conference on Artificial Intelligence*. 2015:1930-1936.

31. Jordan MI, Rumelhart DE: **Forward models: supervised learning with a distal teacher**. *Cogn Sci* 1992, **16**:307-354.

32. Wolpert DM, Flanagan JR: **Forward models**. *The Oxford Companion to Consciousness*. Oxford: Oxford University Press; 2009.

33. Wolpert DM, Kawato M: **Multiple paired forward and inverse models for motor control**. *Neural Netw* 1998, **11**:1317-1329.

34. Brecht M: **The body model theory of somatosensory cortex**. *Neuron* 2017, **94**:985-992.

35. Toussaint M, Allen K, Smith K, Tenenbaum JB: **Differentiable**
•• **physics and stable modes for tool-use and manipulation planning**. *Robotics: Science and Systems*. 2018.
This paper introduces a symbolic-continuous system that integrates physical constraints into task and motion planning. This robotic planner can sequence small subgoals, including spontaneous use of available objects as tools to accomplish complex tasks.

36. Mordatch I, Todorov E, Popović Z: **Discovery of complex behaviors through contact-invariant optimization**. *ACM Trans Graph (TOG)* 2012, **31**:43.

37. Todorov E: **Goal directed dynamics**. *IEEE International*
•• *Conference on Robotics and Automation (ICRA)* 2018:2994-3000.
Goal-directed dynamics is an efficient real-time control framework that solves for coupled control and physics constraints in the same computation, while simultaneously allowing for a high-level interface through cost functions.

38. Yildirim I, Gerstenberg T, Saeed B, Toussaint M, Tenenbaum JB:
• **Physical problem solving: joint planning with symbolic, geometric, and dynamic constraints**. *39th Annual Meeting of the Cognitive Science Society*. 2017.
This paper found that human intuitions about constructing block towers can be well-described using a planner that augments a symbolic-geometric solver with a physics engine.

39. Rizzolatti G, Craighero L: **The mirror-neuron system**. *Annu Rev Neurosci* 2004, **27**:169-192.

40. Treisman A: **Perceiving and re-perceiving objects**. *Am Psychol* 1992, **47**:862.

41. Pylyshyn Z: **The role of location indexes in spatial perception: a sketch of the FINST spatial-index model**. *Cognition* 1989, **32 (1)**:65-97.

42. Kersten D, Schrater PR: **Pattern inference theory: a probabilistic approach to vision**. In *Perception and the Physical World*. Edited by Mausfeld R, Heyer D. Chichester: John Wiley & Sons, Ltd.; 2002.

43. George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, Marthi B, Lou X, Meng Z, Liu Y, Wang H *et al.*: **A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs**. *Science* 2017, **358**:eaag2612.

44. Eslami SA, Rezende DJ, Besse F, Viola F, Morcos AS, Garnelo M, Ruderman A, Rusu AA, Danihelka I, Gregor K *et al.*: **Neural scene representation and rendering**. *Science* 2018, **360**:1204-1210.

45. Lin HW, Tegmark M, Rolnick D: **Why does deep and cheap learning work so well?** *J Stat Phys* 2017, **168**:1223-1247.

46. Wu J, Wang Y, Xue T, Sun X, Freeman B, Tenenbaum J: **Marrnet: 3d shape reconstruction via 2.5 d sketches**. *Advances in Neural Information Processing Systems*. 2017:540-550.

47. Yildirim I, Freiwald W, Tenenbaum J: **Efficient inverse graphics in biological face processing**. *bioRxiv* 2018:282798.

48. Wu J, Lu E, Kohli P, Freeman WT, Tenenbaum JB: **Learning to see**
•• **physics via visual de-animation**. *Advances in Neural Information Processing Systems*. 2017:153-164.
This study proposed a paradigm for understanding physical scenes via 'de-animation'—recovering the underlying physical world states that, when paired with a physics engine and a graphics engine, explain raw visual observations.

49. Wu J, Yildirim I, Lim JJ, Freeman WT, Tenenbaum JB: **Galileo:**
•• **perceiving physical object properties by integrating a physics engine with deep learning**. *Advances in Neural Information Processing Systems*. 2015:127-135.
This paper proposed a computational model that learns to infer physical object properties from raw videos, integrating deep recognition networks with a physics engine. The model's predictions align well with humans' across various behavioral studies.

50. Yildirim I, Smith KA, Belledonne M, Wu J, Tenenbaum JB:
•• **Neurocomputational modeling of human physical scene understanding**. *2nd Conference on Cognitive Computational Neuroscience*. 2018.
This paper proposes a family of models including a recurrent recognition network that explains human physical scene understanding, focusing on 'dynamic updates'—how humans update their beliefs over time when watching a video of a physical event unfold.

51. Le TA, Baydin AG, Wood F: **Inference compilation and universal probabilistic programming**. *Artificial Intelligence and Statistics*. 2017:1338-1348.

52. Hong H, Yamins DL, Majaj NJ, DiCarlo JJ: **Explicit information for category-orthogonal object properties increases along the ventral stream**. *Nat Neurosc* 2016, **19**:613.

53. Conway BR: **The organization and operation of inferior temporal cortex**. *Annual Review of Vision Science*. 2018.

54. Xu Y: **The posterior parietal cortex in adaptive visual processing**. *Trends Neurosci* 2018.

55. Schwettmann SE, Tenenbaum JB, Kanwisher N: **Evidence for an**
•• **intuitive physics engine in the human brain**. *2nd Conference on Cognitive Computational Neuroscience*. 2018.
An fMRI study showing mass of an object can be decoded from brain's responses to passive viewing of object motion. A data-driven searchlight procedure yielded that regions where mass information was decodable above chance closely match the intuitive physics regions from Ref. [25].

56. Chang MB, Ullman T, Torralba A, Tenenbaum JB: **A compositional object-based approach to learning physical dynamics**. *International Conference on Learning Representations*. 2017.

57. Battaglia P, Pascanu R, Lai M, Rezende DJ: **Interaction networks for learning about objects, relations and physics**. *Advances in Neural Information Processing Systems*. 2016:4502-4510.

58. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A,
• Zambaldi V, Malinowski M, Tacchetti A, Raposo D, Santoro A, Faulkner R *et al.*: *Relational Inductive Biases, Deep Learning, and Graph Networks*. . arXiv preprint arXiv:1806.01261 2018.
This review paper discusses the motivation, recent development, and potential applications of graph networks. One of these applications is learning generative models of object dynamics and interaction rules using neuron-like units.

59. Mrowca D, Zhuang C, Wang E, Haber N, Fei-Fei L, Tenenbaum JB, Yamins DL: **Flexible neural representation for physics prediction**. *Advances in Neural Information Processing Systems*. 2018.

60. DiCarlo JJ, Zoccolan D, Rust NC: **How does the brain solve visual object recognition?** *Neuron* 2012, **73**:415-434.

61. Serre T, Oliva A, Poggio T: **A feedforward architecture accounts for rapid categorization**. *Proc Natl Acad Sci* 2007, **104**:6424-6429.

62. Yildirim I, Jacobs RA: **Transfer of object category knowledge across visual and haptic modalities: experimental and computational studies**. *Cognition* 2013, **126**:135-148.

63. Yildirim I, Janner M, Belledonne M, Wallraven C, Freiwald WA, Tenenbaum JB: **Causal and compositional generative models in online perception**. *39th Annual Conference of the Cognitive Science Society*. 2017.

64. Erdogan G, Yildirim I, Jacobs RA: **From sensory signals to modality-independent conceptual representations: a**

probabilistic language of thought approach. *PLoS Comput Biol* 2015, **11**:e1004610.

65. Yeatman JD, Weiner KS, Pestilli F, Rokem A, Mezer A, Wandell BA: **The vertical occipital fasciculus: a century of controversy resolved by in vivo measurements**. *Proc Natl Acad Sci* 2014, **111**:E5214-E5223.

66. Parlatini V, Radua J, Dell'Acqua F, Leslie A, Simmons A, Murphy DG, Catani M, de Schotten M T: **Functional segregation and integration within fronto-parietal networks**. *Neuroimage* 2017, **146**:367-375.

67. Wu J, Lim JJ, Zhang H, Tenenbaum JB, Freeman WT: **Physics 101: learning physical object properties from unlabeled videos**. *British Machine Vision Conference* 2016, **2**:7.