34 Physical Object Representations for Perception and Cognition

ILKER YILDIRIM, MAX SIEGEL, AND JOSHUA TENENBAUM

ABSTRACT Theories of perception typically assume that the goal of sensory processing is to output simple categorical labels or low-dimensional quantities, such as the identities and locations of objects in a scene. But humans perceive much more in a scene: we perceive rich and detailed three-dimensional shapes and surfaces, substance properties of objects (such as whether they are light or heavy, rigid or soft, solid or liquid), and relations between objects (such as which objects support, contain, or are attached to other objects). These physical targets of perception support flexible and complex action as the substrate of planning, reasoning, and problem-solving. In this chapter we introduce and argue for a theory of how people perceive, learn, and reason about objects in our sensory environment in terms of what we call *physical object representations* (PORs). We review recent work showing how this explains many human judgments in intuitive physics, provides a basis for object shape perception when traditional visual cues are not available, and, in one domain of high-level vision, suggests a new way to interpret multiple stages of hierarchical processing in the primate brain.

Consider the scenes in figure 34.1*A* and *B*. In each case we see a set of apples in a certain geometric arrangement (figure 34.1*C*, *D*). But we also see so much more: We see fine-grained details of their three-dimensional (3-D) shapes. We infer their physical properties and relationships: which objects are supporting which others and how heavy or light or hard or soft they would feel if we picked them up. We can predict whether the stack would topple if the middle apple on the bottom row were removed, and we can plan how to pick the designated apple without making the rest unstable. We can also "see" that picking the apple in figure 34.1*B* is much easier and can be achieved with just one action using just one hand (as opposed to the two hands or a more complex sequence of actions needed for the stack in figure 34.1*A*). These abilities are present even early in childhood (figure 34.1*E*) and are likely shared with other species, particularly nonhuman primates (figure 34.1*F*). They are general purpose and can be used to think about many different kinds of physical scenarios and judgments: For instance, can you arrange a set of objects into a stable tower using wooden blocks or Lego bricks (as in figure 34.1*E*)? What about using stones or bricks or cups or even apples?

How might we explain these flexible, seemingly effortless judgments? This chapter presents an answer centered at the notion of physical object representations (PORs), a basic system of knowledge that supports perceiving, learning, and reasoning about all the objects in our environment—their shapes, appearances, affordances, substances, and the way they react to forces applied to them. Our goal here is to outline a computational framework for studying the form and content of PORs in the mind and brain. PORs can be considered an interface between perception and cognition, linking what we perceive to how we plan our actions and talk about the world. Despite their fundamental role in perception, many important questions about object representations remain open. What kind of information formats or data structures underlie PORs so as to support the many ways in which humans flexibly and creatively interact with the world? How can properties of objects be inferred from sensory inputs, and how are they represented in neural circuits? How can these representations integrate sense data across vision, touch, and audition?

After introducing the computational ingredients of POR theory from a reverse-engineering perspective, we review recent work that is beginning to answer some of these questions. We focus on three case studies: (1) how PORs can explain human judgments in intuitive physics, across a broad range of physical outcome prediction scenarios; (2) how PORs provide a substrate for physically mediated object shape perception in scenarios where traditional visual cues fail and a natural substrate for multimodal (visual-haptic) perception and crossmodal transfer; and (3) how in one domain of high-level vision—face perception—PORs might be computed by neural circuits, and how thinking in terms of PORs suggests a new way to interpret multiple stages of processing in the primate brain.

Physical Object Representations

How, in engineering terms, can we formalize PORs? There are two main aspects to our proposal. The first is

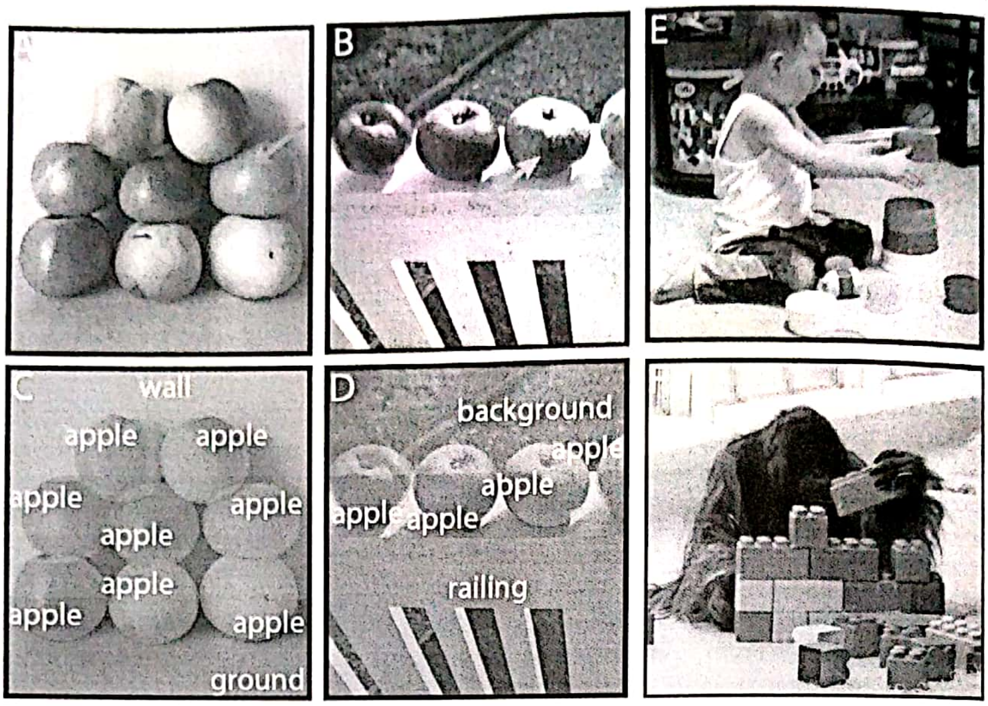

FIGURE 34.1 *A* and *B*, How would you pick up the apples indicated while maintaining a stable arrangement of the other objects? It is easy to see that you will likely need to touch more objects (and probably use two hands) in panel (*A*), while the apple in panel (*B*) can be removed on its own with just one hand. *C* and *D*, What is where? Semantic segmentation maps showing class labels and locations of objects from panels (*A* and *B*). *E*, A child playing with stacking cups. Screenshot from https://www.youtube.com/watch?v=dEnDjyWHN4A. *F*, An orangutan building a tower with large Lego-like blocks. Screenshot from https://www.youtube.com/watch?v=MxRJjzSY_JE&t=21s. (See color plate 37.)

a working hypothesis about the contents of PORs. We draw on tools developed for video game engines (Gregory, 2014), including graphics (Blender Online Community, 2015) and physics engines (Coumans, 2010; Macklin, Müller, Chentanez, & Kim, 2014) and planning engines from robotics for grasping and other humanoid motions (Miller & Allen, 2004; Todorov, Erez, & Tassa, 2012; Toussaint, 2015). These tools instantiate simplified but algorithmically tractable models of reality that capture our basic knowledge of how objects work and how our bodies interact with them. In these systems, objects are described by just those attributes needed to simulate natural-looking scenes and motion over short timescales (~2 seconds): 3-D geometry, substance or mechanical material properties (e.g., rigidity), optical material properties (e.g., texture), and dynamical properties (e.g., mass). Video game engines provide causal models in the sense that the process by which the data (i.e., natural-looking scenes) are generated has some abstract level of resemblance to its corresponding real-world process in a form efficient enough to support real-time interactive simulation.

Second, we embed these simulation engines within probabilistic generative models. Physical properties of an object are not directly observable in the raw signals arriving at our sensory organs. These properties, including 3-D shape, mass, or support relations, are latent variables that need to be inferred given sense inputs; they are products of perception. Probabilistic modeling provides the mathematical language to rigorously and unambiguously specify the domain and task being studied, and to explain how, given sensory inputs, latent properties and relations in the underlying physical scene can be reliably inferred through some form of approximate Bayesian inference (see Kersten and Schrater [2002] for an in-depth treatment of this perspective). The probabilistic models we build to capture PORs can be seen as a special case of *probabilistic programs*, or generalizations of directed graphical models (Bayesian networks) that define random variables and conditional probability distributions relating variables using more general data structures and algorithms than simply graphs and matrix algebra (see Ghahramani [2015] and Goodman and Tenenbaum [2016] for an introduction).

The POR framework is closely related to *analysis-by-synthesis* (A×S) accounts of perception: the notion that perception is fundamentally about inverting the causal process of image formation (Helmholtz & Southall, 1924; Rock, 1983). In this view, perceptual systems model the causal processes by which natural scenes are constructed, as well as the process by which images are formed from scenes; this is a mechanism for the hypothetical "synthesis" of natural images, in the style of computer graphics, by using a graphics engine. Perception (or "analysis") is then the search for or inference to the best explanation (or plausible explanations) of an observed image in terms of this synthesis, which in the POR framework can be implemented using Bayesian inference.

Most mechanisms for approximating Bayesian inference that have traditionally been proposed in analysis by synthesis (e.g., Markov chain Monte Carlo, or MCMC) seem implausible when considered as an algorithmic account of perception: they are inherently iterative and almost always far too slow relative to the dynamics of perception in the mind or brain. We draw on recent advances in machine learning and probabilistic programming (including deep neural networks, particle filters or sequential importance samplers, data-driven MCMC, approximate Bayesian computation, and hybrids of these methods) to construct efficient and neurally plausible approximate algorithms for the physical inference tasks specified with our probabilistic models.

While our focus in this chapter is perception, the domain of the POR framework is more general. With a causal model of the world (including its state-space structure—i.e., object dynamics and interactions in a physics engine) and a planner based on a body model, the POR framework transforms the physical environment around us into something computable, naturally supporting many aspects of cognition, including reasoning, imagery, and planning for locomotion and object manipulation via simulation-based inference and control algorithms. In this sense, PORs express functionality somewhat analogous to the "emulators" of emulation theory (Grush, 2004), an earlier proposal for an integrated account of perception, imagery, and motor planning that also fits broadly within a Bayesian approach to inference and control. A key difference is the language of representation for state, dynamics, and observation. Emulation theory was formulated using classical ideas from estimation and control, such as the Kalman filter: body and environment state are represented as vectors, dynamics are linear, and observations are linear functions of the state with Gaussian added noise. The computations supported are simpler but much less expressive than in the POR framework, where state is represented with structured object and scene descriptions, dynamics using physics engines, and observation models using graphics engines. PORs can thus explain how cognitive and perceptual processes operate over a much wider range of physical scenarios, varying greatly in complexity and content, although they require more algorithmic machinery to do so.

Intuitive Physical Reasoning

Having overviewed the basic components of PORs, we now turn to recent computational and behavioral work exploring their application in several domains. We begin with intuitive physics, in the context of scene understanding. Recall the introductory example displayed in figure 34.1. The POR framework was first introduced to answer these kinds of questions, in a form similar to how we characterize it here, by Battaglia, Hamrick, and Tenenbaum (2013). They showed that approximate probabilistic inferences over simulations in a game-style physics engine could be used to perform many different tasks in blocks-world type scenes. While physics engines are designed to be deterministic, Battaglia, Hamrick, and Tenenbaum (2013) found that human judgments were best captured using a probabilistic model that combined the deterministic dynamics of the physics engine with probability distributions over the uncertain geometry of objects' initial configurations and/or shapes, their physical attributes (e.g., their masses), and perhaps the nature of the forces at work (e.g., friction or perturbations of the supporting surface).

In one version of this model (figure 34.2), input images comprised one or more static 2-D views of a tower of blocks in 3-D that might fall over under gravity, and the task was to make various judgments about what would or could happen in the near future. Object shapes and physical properties were assumed to be known, but the model had to estimate the 3-D scene configuration for the blocks. This inference step used A×S with a top-down stochastic search-based (MCMC) procedure: Block positions in 3-D are iteratively and randomly adjusted until the rendered (synthesized) 2-D images approximately match the input images; multiple runs of this procedure yield slightly different outputs, representing samples from an approximate Bayesian posterior distribution on scenes given images. Once these physical object representations are established, they support a wide range of dynamical inferences that go well beyond the purely static content in the perceptual input. How likely is the tower to fall? If it falls, how much of the tower will fall? In which direction will the blocks fall? How far will they fall? If the table supporting the tower were bumped, how many or which of the blocks would fall off the table? If the tower is unstable, what kind of applied force or other action could hold it stable?

To see how these judgments are computed, consider answering the questions: How likely is the tower to fall?

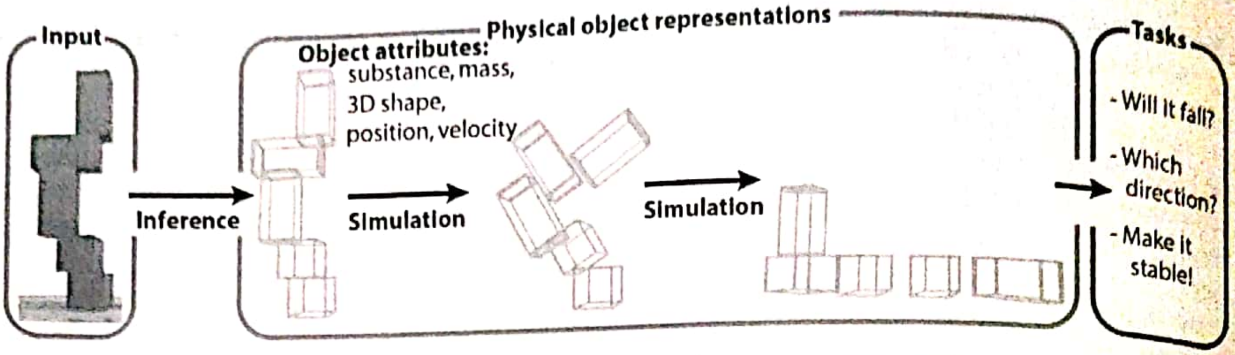

FIGURE 34.2 A schematic of the POR framework applied to intuitive physical reasoning with a tower of wooden blocks. *Left to right,* The input image; inference to recover the 3-D scene and physical properties of objects; physics engine simulation to predict near-future states given the inferred initial configuration; and questions that can be answered and tasks that can be performed based on such simulations.

How much of this tower is likely to fall? One way to make these judgments is to run a small number of forward simulations using a physics engine (implemented, e.g., using Bullet & Coumans, 2010), starting from the sample of configurations returned by the probabilistic 3-D scene inference procedure. These simulations run until all objects stop moving, or some short time limit has elapsed. The distribution of their outcomes represents a sample of the Bayesian posterior predictive distribution on future states, conditioned on the input image and the model's representation of physics. Predictive judgments such as those above can then be calculated by simply querying each sample and aggregating: for example, the model's judgment of "How likely is the tower to fall?" is calculated as the average number of simulations in which the tower fell (relative to the total number of simulations ran); "How much of the tower is likely to fall?" is calculated by averaging the proportion of blocks that fell in each simulation.

Strikingly, Battaglia, Hamrick, and Tenenbaum (2013) found that only a few such posterior samples (they estimated typically three to seven samples per participant, per trial), generated from the highly approximate simulations of video game physics engines under perceptual uncertainty, were sufficient to account for human judgments across a wide range of tasks with high quantitative accuracy. In the last several years, a growing number of behavioral and computational studies have developed approximate probabilistic simulation models of the PORs underlying our everyday physical reasoning abilities. Studies have examined intuitive judgments of mass from how towers do or don't fall (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016); predictions about future motions (Smith, Battaglia, & Vul, 2013b; Smith, Dechter, Tenenbaum, & Vul, 2013a); judgments of multiple physical properties (e.g., friction as well as mass) and latent forces such as magnetism from examining how objects move and collide in planar motion (Ullman, Stuhlmuller, Goodman, & Tenenbaum, 2018; see also the seminal earlier work on probabilistic inference in collisions by Sanborn, Mansinghka, and Griffiths [2013]); and predictions about the behavior of liquids such as water and honey (Bates, Yildirim, Battaglia, & Tenenbaum, 2015; Kubricht et al., 2016), and granular materials such as sand (Kubricht et al., 2017), falling under gravity. Taken together, these studies show how the POR framework provides a broadly applicable, quantitatively testable, and functionally powerful computational substrate for everyday intuitive physical scene understanding.

How might PORs and their associated computations be implemented in neural hardware? As a first step toward addressing this question, a recent functional magnetic resonance imaging (fMRI) study in humans aimed to localize cortical regions involved in many of the intuitive physics judgments discussed above (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). Fischer et al. (2016) found a network of parietal and premotor regions that was differentially activated for physical reasoning tasks in contrast to difficulty-matched nonphysical tasks (such as color judgments, or social predictions) with the same or highly similar stimuli. These regions were consistent across multiple experiments controlling for different task demands and across different visual scenarios. A recent fMRI study in macaques found a similar brain network differentially recruited for analogous physical versus nonphysical stimulus contrasts, in a passive-viewing paradigm (Sliwa & Freiwald, 2017). These networks closely overlap with networks for action planning and tool use in humans (see Gallivan and Culham [2015] for a review) and the mirror neuron system in monkeys that is thought to be involved in action understanding (Rizzolatti & Craighero, 2004), consistent with the proposal that PORs provide a bridge between perception and cognitive functions of action

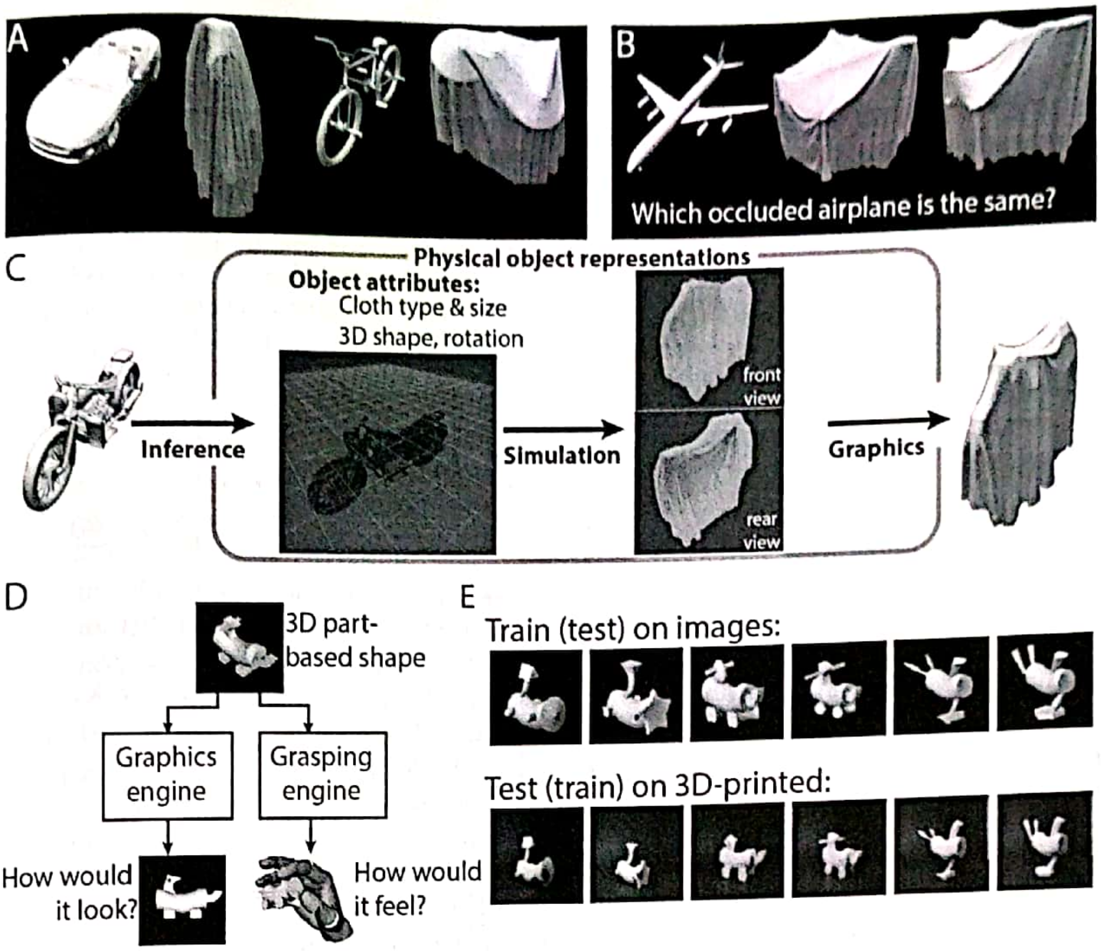

FIGURE 34.3 *A*, Example pairs of unoccluded objects and cloth-occluded matches in different poses. *B*, An example trial from Yildirim, Siegel, and Tenenbaum (2016), where the task is to match the unoccluded object to one of the two occluded objects. *C*, A schematic of the POR framework applied to the object-under-cloth task. *Left to right*, The input image; inference to recover the 3-D shape of the unoccluded object and imagining a cloth positioned above it; physics engine simulation to the predict dropping of the cloth on the object shown at two different angles; and graphics to predict what the resulting scene would look like. *D*, A multisensory causal model combining a graphics engine with a grasp-planning engine. *E*, Example novel objects from Yildirim and Jacobs (2013), rendered visually and photographed after 3-D printing using plastic.

planning, reasoning, and problem solving. Future experimental work using physiological recordings, informed by some of the more neurally grounded models discussed later in this chapter, can now target neural populations in these brain networks in order to elucidate the neural circuits underlying intuitive physics.

Physics-Mediated Object Shape Perception

We now turn to the role of PORs in a more purely perceptual task: perceiving object shape. Vision scientists traditionally study many cues as routes to 3-D shape, such as contours, shading, stereo disparity, or motion. But physics can also be an essential route to shape, especially when these traditional cues are unavailable or insufficient; such cues may be necessary for the correct recovery of a target shape but fail to capture all of the causal processes underlying the appearance of an image. Consider seeing an object that is heavily or even entirely occluded, as when draped by a cloth (figures 34.2*B* and 34.3*A*). It is likely you haven't seen airplanes or bicycles occluded under a cloth before, but it is still relatively easy to pair an unoccluded object with its randomly rotated and occluded counterpart. Of course, shading cues allow you to see the contours of the cloth as an occluding surface. Yet these cues alone do not explain how you perceive the shape of the underlying occluded object, which together with the physical properties of the cloth is the real cause of the shading patterns observed.

Most contemporary approaches to visual object perception emphasize learning to "untangle" or become invariant to sources of variation in the image (DiCarlo & Cox, 2007; Serre, Oliva, & Poggio, 2007). On this account, a processing hierarchy (such as a deep neural network) progressively transforms sensory inputs until reaching an encoding that is diagnostic for a particular object shape or identity and invariant to other factors (Riesenhuber & Poggio, 1999). These approaches can perform very well when trained to ignore a given class of variations, but to achieve optimal performance, they must be trained anew (or at least "fine-tuned") independently for every new kind of invariance. They do not show instantaneous (zero-shot) invariance for new ways an object might appear, such as those arising from an occluding cloth.

The POR framework provides a different approach in which the goal is not learning invariances but explaining variation in the image with respect to the causal process generating images from 3-D physical scenes (e.g., Mumford, 1997; Yuille & Kersten, 2006). For the object-under-cloth task, this process can be captured by composing (1) a physics engine simulating how cloth drapes over 3-D rigid shapes, (2) a graphics engine simulating how images look from the resulting scenes (occluded or unoccluded), and (3) a probabilistic inference engine. The inference engine inverts the graphics process to recover 3-D shapes from unoccluded images and then imagines likely images under different ways these shapes could be rotated and draped under cloth (figure 34.3*C*). Yildirim, Siegel, and Tenenbaum (2016) presented preliminary evidence that such a mechanism fits human judgments in a match-to-sample task, akin to figure 34.3*B*, across four difficulty levels. In contrast, a deep neural network trained for invariant object recognition, but not specifically for scenes involving cloth-based occlusion, could fit the easiest human judgments but failed to generalize above chance for the harder judgments. These results illustrate a key advantage of the POR framework: the ability to generalize to novel settings not by requiring further training but by combining or composing existing causal models.

The POR framework supports combining causal models not only across multiple visual cues but also across sensory modalities. This is because the contents of PORs are not specific to vision or any single modality but instead capture the physical properties of objects that are the root causes of sense data in every modality, via appropriate modality-specific "rendering" engines (such as a graphics engine in vision). Embedded in a framework for probabilistic inference to invert these renderers, PORs provide a basis for perceiving shape from any form of sense data, as well as for multisensory integration and cross-modal perception. Consider the POR-based model shown in figure 34.3*D*: Starting from a probabilistic generative model over part-based body shapes in 3-D, the multisensory causal model combines a visual graphics engine that generates the 2-D appearance of each shape viewed in a given pose with a touch or haptic rendering engine, based on a kinematic grasp planner, that generates the way a shape feels in the hand given a certain grasp trajectory. Bayesian inference then allows the model to estimate a 3-D shape that explains inputs from either visual or haptic channels, or both, as well as to automatically and without further training transfer that shape from objects first encountered in one modality (e.g., visually) to recognize how they would be perceived in another modality (e.g., haptically). Yildirim and Jacobs (2013) found that this model accounted for the performance of human participants in a visual-haptic crossmodal categorization task (example stimuli are shown in figure 34.3*E*). These results were extended to a visual-haptic shape similarity judgment task (Erdogan, Yildirim, & Jacobs, 2015).

The idea that shared neural representations support object perception across multiple sensory modalities is consistent with a number of fMRI studies (e.g., Amedi, Jacobson, Hendler, Malach, & Zohary, 2002; James et al., 2002; Lacey, Tal, Amedi, & Sathian, 2009; Lee Masson, Bulthé, Op de Beeck, & Wallraven, 2016; Tal & Amedi, 2009). The POR framework provides explicit hypotheses as to what the format of such multisensory neural representations might be. Erdogan, Chen, Garcea, Mahon, and Jacobs (2016) used fMRI to test one such hypothesis introduced in their earlier computational work (Erdogan, Yildirim, & Jacobs, 2015). In addition to finding that visual and haptic exploration of novel objects gave rise to similar patterns of neural activity in the lateral occipital cortex (LOC), they also found that this activity could be crossmodally decoded to the part-based 3-D object structure mentioned above (Erdogan, Yildirim, & Jacobs, 2015). This activity may be a result of visual imagery as opposed to haptic processing; however, other work suggests that imagery only minimally activates LOC (Amedi, Malach, Hendler, Peled, & Zohary, 2001; James et al., 2002). Further experimental work along these lines, aiming to quantitatively test specific POR models and ideally extending into physiological recordings from neural populations, could lead to a more precise understanding of the neurocomputational basis of multisensory perception and crossmodal transfer.

Reverse-Engineering Ventral Visual Stream Computations Using Physical Object Representations

We now turn to discussing how the POR framework can illuminate aspects of the neural circuits underlying perception. Even though traditional A×S methods can recover PORs from sense inputs, these algorithms (based on top-down, iterated stochastic search) do not

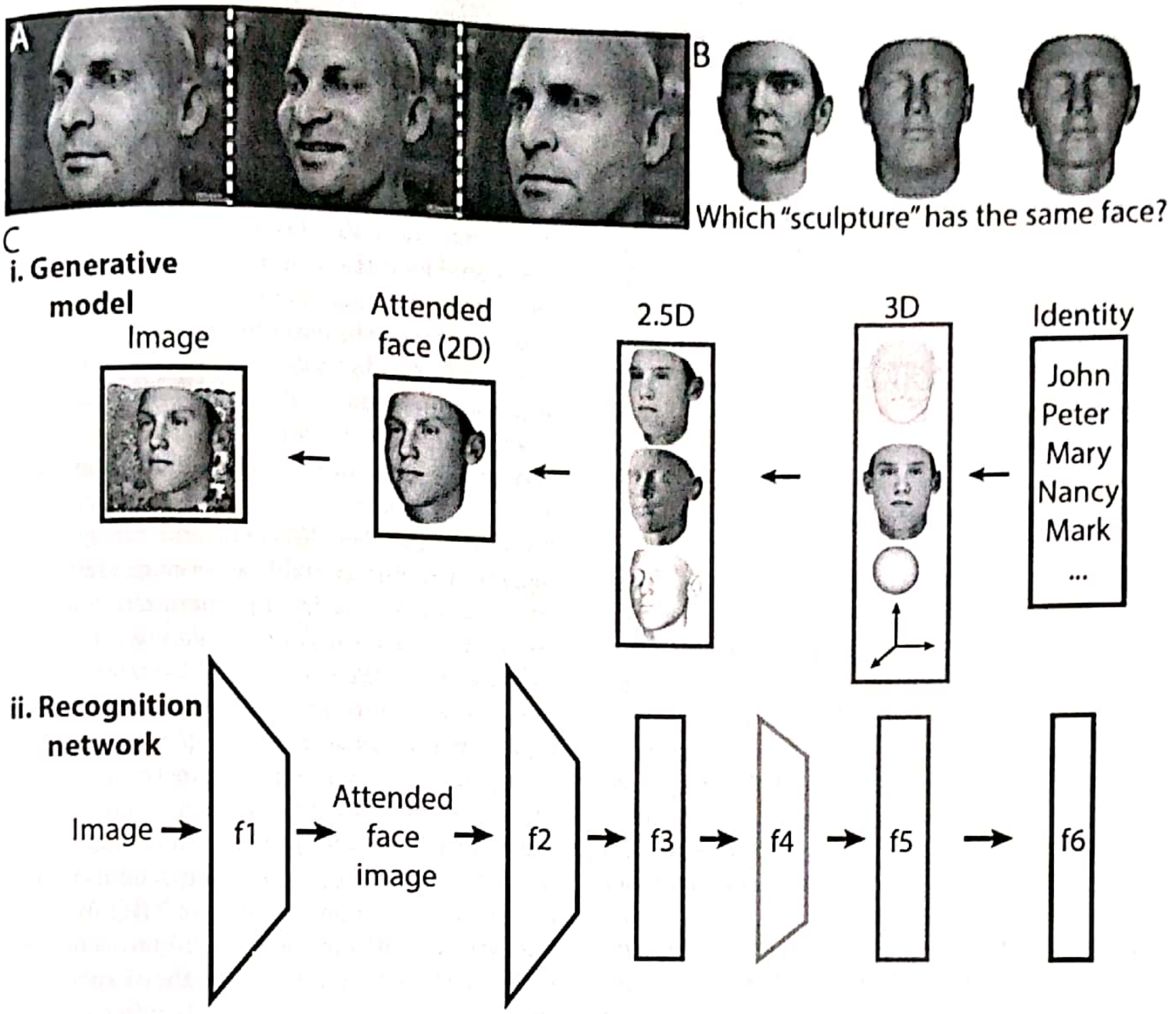

FIGURE 34.4 *A*, Samples from a modern 3-D graphics model of a human face, yielding near photorealistic images (Credit: NVIDIA and University of Southern California Institute for Creative Technologies). Across the three images of this face, in addition to knowing that identity is preserved, we can also appreciate the details of the face's 3-D shape and texture, the subtleties of expression, that vary or remain constant across images. *B*, Despite their unfamiliarity, most observers can match the identity of the naturalistic face on the left to one of the textureless faces ("sculptures"), which must rely on a sense of 3-D shape. *C*, Schematic of the efficient A×S approach, including a probabilistic generative model of face image formation (*panel i*) and the recognition network (*panel ii*). Layers f1 through f6 indicate the different components of the recognition network. Trapezoids show single or multiple layers of transformations where a layer can consist of convolution, normalization, and a nonlinear activation function. Yildirim et al. (2019) found that transformations across the model layers f3, f4, and f5 closely captured the transformations observed in the neural data from ML/MF (middle lateral and middle fundus areas) to AL (anterior lateral area) to AM (anterior medial area; Freiwald & Tsao, 2010). (See color plate 38.)

readily map onto neural computation. Many authors have thus preferred feedforward network models, most recently deep convolutional neural networks (CNNs), which are both more directly relatable to neural circuit-level mechanisms and more consistent with the fast bottom-up processing observed in perception. However, CNNs, typically trained for invariant object recognition or "untangling," do not explicitly address the question of how vision recovers the causal structure of scene and image formation. Therefore, neither traditional approaches to A×S nor modern CNNs really answer the challenge: How do our brains compute rich descriptions of scenes, with detailed 3-D shapes and surface appearances, in much less than a second?

A new class of computational models aim to combine the best aspects of these two approaches by using CNNs or recurrent networks to map images to their underlying scene descriptions, thereby accomplishing otherwise computationally costly inference in one or a few bottom-up passes on the image (Eslami et al., 2018; George et al., 2017; Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015; Yildirim, Kulkarni, Freiwald, &

Tenenbaum, 2015). Yildirim, Belledonne, Freiwald, and Tenenbaum (2019) developed one such approach using the POR framework and tested it as a computational theory of multiple stages of processing in the ventral visual stream, a hierarchy of processing stages in the visual brain (Conway, 2018). This model consists of two parts: a generative model based on a multistage 3-D graphics program for image synthesis (figure 34.4C) and a recognition model based on a CNN that approximately inverts the generative model, stage by stage (figure 34.4C). The recognition network is different from conventional CNNs for vision in two ways. First, it is trained to produce the inputs to a graphics engine, the latent or unobservable variables of the probabilistic model, instead of predicting class labels such as face identities. And second, it is trained in a self-supervised fashion, with inputs and targets internally synthesized by the probabilistic graphics component; no externally generated labels are needed. This approach differs from other recent efficient A×S approaches (Eslami et al., 2018; Kulkarni et al., 2015) and their earlier counterparts (Dayan, Hinton, Neal, & Zemel, 1995) in that it is based on a probabilistic graphics engine (instead of learning an unstructured generative model via a generic function approximator) and therefore more closely captures the causal structure of how 3-D scenes give rise to images.

Yildirim, Belledonne, Freiwald, and Tenenbaum (2019) tested their approach in one domain of high-level perception, the perception of faces. Faces give rise to a rich sense of 3-D shape in addition to percepts of a discrete individual's identity (see figure 34.4A, B), and face perception has been extensively studied in both psychology and neurophysiology, thus providing a rich source of data and constraints for modeling. The sense of a face's 3-D shape also crosses between visual and haptic modes of perception (Dopjans, Wallraven, & Bulthoff, 2009), as in the examples discussed above.

Yildirim, Belledonne, Freiwald, and Tenenbaum (2019) compared two broad classes of hypotheses for how we perceive the 3-D shape of a face and how these computations are implemented in the primate ventral stream: (1) the efficient A×S hypothesis implemented in their recognition network, which posits that the targets of ventral stream processing are latent variables in a probabilistic causal model of image formation, and (2) the untangling hypothesis implemented in standard deep CNNs for face recognition, which posits that the target of ventral stream processing is an embedding space optimized for discriminating among facial identities. Their recognition network implementing the A×S hypothesis recapitulated transformations across multiple stages of processing in inferio temporal (IT) cortex from middle lateral and middle fundus areas (ML/MF) to anterior lateral area (AL) to anterior medial area (AM)—the three sites in the monkey face patch system—with respect to the similarity structure of the population-level activity in each stage (Freiwald & Tsao, 2010). Both in the neural data and in the model, these similarity structures progressed from view-based to mirror-symmetric to view-invariant representations. Alternative models, including a number implementing the untangling hypothesis, did not capture these transformations. The efficient A×S model also accurately matched human error patterns in psychophysical experiments, including experiments designed to determine how flexibly humans can attend to either the shape or texture components of a face stimulus (figure 34.4B). Finally, the recognition model suggested an interpretable account of some intermediate representations in this hierarchy: in particular, population-level similarity structure of middle face patches (ML/MF) can be well accounted for by the similarity structure arising from intermediate surface representations, such as intrinsic images (normal maps or depth maps for surface geometry and albedos for surface color) or a 2.5-D sketch.

The efficient A×S approach thus offers a potential resolution to the issue of interpretability in systems neuroscience (Yamins & DiCarlo, 2016). In addition to assessing accounts of the brain in terms of how much variance in neural firing rates they explain, the efficient A×S approach suggests that computational neuroscientists could aim for "semi-interpretable" models of perception where the recognition network as a whole can be understood as inverting a causal generative model, and subpopulations of neurons in particular stages of the recognition network (such as ML/MF and AM) can be understood as inverting distinct, identifiable stages in the generative model, explicitly representing hypotheses about the corresponding aspects of scene structure encoded in those generative model stages. Other populations of neurons (such as AL) might be better explained as implementing valuable hidden-layer nonlinear transforms between more interpretable parts of the system.

Conclusion and Future Directions

We believe that there is promising, if preliminary, evidence for the centrality of PORs in the mind and brain. The strongest aspect of this proposal so far is theoretical: PORs offer a solution to problems both old (e.g., multimodal perception) and new (e.g., the cloth-draping task presented above), perceptual phenomena that are difficult to explain with alternative accounts in either cognitive neuroscience or artificial intelligence. There remain, however, significant challenges. Empirical work has only begun to test strong predictions of the POR framework; far more behavioral and physiological data are needed. As we have noted, PORs provide a rich foundation for structuring perception and behavior, but this comes with a heavy computational burden. The efficient A×S approach is one possible way the brain might handle this complexity, but again more study is needed, especially relating the dynamics of processing in these models to the dynamics of neural computation. Further theoretical work is also required to explore the origins of PORs: how an organism comes to possess an object-based causal model of the world around it.

The POR framework also offers new research directions for studying aspects of complex behavior production and object manipulation. An important advantage of the POR framework is that causal models of the world allow for flexible action planning, reasoning, and intelligent object manipulation. To illustrate, we revisit the grasping engine shown in figure 34.3*D* in its broader context. This grasping engine implements a planner based on a simulatable body model (similar to forward models typically invoked in models of motor control; Jordan & Rumelhart, 1992; Wolpert & Flanagan, 2009; Wolpert & Kawato, 1998). Such a model allows embodied agents to evaluate the consequences of their actions by simulating them internally before (or without ever) actually performing them. Many organisms likely use this approach—for example, performing simulations for making a judgment about the action "Can I jump?" Brecht (2017) suggested that the microcircuits in the mammal somatosensory cortex implement a simulatable body model that can be used for action planning and decision-making. The POR framework provides a toolkit to capture these computations in engineering terms using existing simulation engines (e.g., see Yildirim, Gerstenberg, Saeed, Toussaint, and Tenenbaum [2017] for a proof-of-concept implementation in the context of complex object manipulation).

Perhaps the most important open question is also the most challenging: How could simulations with richly structured generative models, such as graphics engines, physics engines, and body models, be implemented in neural mechanisms? Recent developments in machine learning and perception suggest intriguing possibilities based on deep learning systems that are trained to emulate a structured generative model in an artificial neural network architecture. Deep networks that emulate graphics engines were mentioned above; while they do not yet come close to the full functionality of traditional graphics engines, their performance in narrow domains can be surprisingly impressive and continues to improve. In intuitive physics, hybrids of discrete symbolic and distributed representations, such as neural physics engines (Chang, Ullman, Torralba, & Tenenbaum, 2016), interaction networks (Battaglia, Pascanu, Lai, & Rezende, 2016) and other graph networks (Battaglia et al., 2018), and hierarchical relation networks (Mrowca et al., 2018), have received much attention lately. These systems assume discrete symbolic representations for each object and its relation to other objects and vector representations for the rules of physical interactions between objects; this allows the dynamics of object motion and interaction (e.g., collisions) to be learned efficiently end-to-end from simulated data. Artificial neural networks such as these can be considered partial hypotheses for how graphics and physics might be implemented in biological neural circuits; they are almost surely wrong or at best incomplete, but they suggest a way forward. Further work is needed to test these models empirically and to develop their capacities; currently, they are very limited in the scope of physics they can learn (e.g., a limited class of rigid body interactions, such as billiard balls colliding on a table). Nevertheless, with these advances and building on the example of the efficient A×S approach and other research linking artificial neural networks to neural representations in the brain, we see promise in linking the POR framework to neural computation in perception and well beyond.

Acknowledgments

We thank Amir A. Soltani and Mario Belledonne for their help with the figures. We thank James Traer, Max Kleiman-Weiner, and our section editor Josh McDermott for their feedback on earlier versions of this chapter. This work was supported by the Center for Brains, Minds and Machines (CBMM) and funded by National Science Foundation STC award CCF-1231216; the Office of Naval Research Multidisciplinary University Research Initiatives grant N00014-13-1-0333; a grant from the Toyota Research Institute; and a grant from the Mitsubishi Electric Corporation.

REFERENCES

Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex, 12*(11), 1202–1212.

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience, 4*(3), 324.

Bates, C., Battaglia, P., Yildirim, I., & Tenenbaum, J. B. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 172–177.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... Gulcehre, C. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv*. Retrieved from 1806.01261.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences, 110* (45), 18327–18332.

Battaglia, P., Pascanu, R., Lai, M., & Rezende, D. J. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing systems*, 4502–4510. Curran Associates, Inc.

Blender Online Community. (2015). Blender—a 3D modelling and rendering package [Computer software manual]. Amsterdam: Blender Institute. http://www.blender.org.

Brecht, M. (2017). The body model theory of somatosensory cortex. *Neuron, 94*(5), 985–992.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *arXiv*. Retrieved from 1612.00341.

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science, 4, 381–402.*

Coumans, E. (2010). Bullet physics engine. [Open-source software]. http://bulletphysics.org.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7(5), 889–904.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341.

Dopjans, L., Wallraven, C., & Bulthoff, H. H. (2009). Crossmodal transfer in visual and haptic face recognition. *IEEE Transactions on Haptics, 2*(4), 236–240.

Erdogan, G., Chen, Q., Garcea, F. E., Mahon, B. Z., & Jacobs, R. A. (2016). Multisensory part-based representations of objects in human lateral occipital cortex. *Journal of Cognitive Neuroscience, 28*(6), 869–881.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Computational Biology, 11*(11), e1004610.

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., ... Reichert, D. P. (2018). Neural scene representation and rendering. *Science, 360*(6394), 1204–1210.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences, 113*(34), E5072–E5081.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science, 330*(6005), 845–851.

Gallivan, J. P., & Culham, J. C. (2015). Neural coding within human brain areas involved in actions. *Current Opinion in Neurobiology, 33*, 141–149.

George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., ... Lavin, A. (2017). A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science, 358*(6368), eaag2612.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452.

Goodman, N. D., Tenenbaum, J. B., & The ProbMods Contributors. (2016). *Probabilistic models of cognition* (2nd ed.). Retrieved September 1, 2018, from https://probmods.org.

Gregory, J. (2014). *Game engine architecture*. Boca Raton, FL: CRC Press.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*(3), 377–396.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition, 157*, 61–76.

Helmholtz, H. V., & Southall, J. P. C. (1924). *Helmholtz's treatise on physiological optics*. Rochester, NY: Optical Society of America.

James, T. W., Humphrey, G. K., Gati, J. S., Servos, P., Menon, R. S., & Goodale, M. A. (2002). Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia, 40*(10), 1706–1714.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*(3), 307–354.

Kersten, D., & Schrater, P. R. (2002). Pattern inference theory: A probabilistic approach to vision. In R. Mausfeld & D. Heyer (Eds.), *Perception and the physical world*, 191–228. Chichester, UK: John Wiley & Sons.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences, 21*(10), 749–759.

Kubricht, J., Jiang, C., Zhu, Y., Zhu, S. C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 1805–1810.

Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S. C., & Lu, H. (2017). Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 700–705.

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4390–4399.

Lacey, S., Tal, N., Amedi, A., & Sathian, K. (2009). A putative model of multisensory object representation. *Brain Topography, 21*(3–4), 269–274.

Le, T. A., Baydin, A. G., & Wood, F. (2016). Inference compilation and universal probabilistic programming. *arXiv*. Retrieved from 1610.09900.

Lee Masson, H., Bulthé, J., Op de Beeck, H. P., & Wallraven, C. (2016). Visual and haptic shape processing in the human brain: Unisensory processing, multisensory convergence, and top-down influences. *Cerebral Cortex, 26*(8), 3402–3412.

Macklin, M., Müller, M., Chentanez, N., & Kim, T. Y. (2014). Unified particle physics for real-time applications. *ACM Transactions on Graphics, 33*(4), 153.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Miller, A. T., & Allen, P. K. (2004). Graspit! A versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine, 11*(4), 110–122.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L., Tenenbaum, J. B., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. *arXiv*. Retrieved from 1806.08047.

Mumford, D. (1996). Pattern theory: A unifying perspective. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference*, 25–62. Cambridge: Cambridge University Press.

Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. In C. Casanova & M. Ptito (Eds.), *Progress in brain research* (Vol. 134, pp. 427–445). New York: Elsevier.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review, 120*(2), 411.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104*(15), 6424–6429.

Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science, 356*(6339), 745–749.

Smith, K. A., Battaglia, P., & Vul, E. (2013b). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 426–3431.

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013a). Physical predictions over time. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 1342–1347.

Tal, N., & Amedi, A. (2009). Multisensory visual-tactile object related network in humans: insights gained using a novel crossmodal adaptation approach. *Experimental Brain Research, 198*(2), 165–182.

Todorov, E., Erez, T., & Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. Vilamoura.

Toussaint, M. (2015). Logic-geometric programming: An optimization-based approach to combined task and motion planning. *International Joint Conferences on Artificial Intelligence*, 1930–1936.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology, 104*, 57–82.

Wolpert, D. M., & Flanagan, J. R. (2009). Forward models. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford Companion to Consciousness*, 294–296. New York: Oxford University Press.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*(7–8), 1317–1329.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. *Advances in neural information processing systems*, 127–135. Curran Associates, Inc.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2019). Efficient inverse graphics in biological face processing. *bioRxiv*, 282798v2.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv*. Retrieved from 1707.08212.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition, 126*(2), 135–148.

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2751–2756.

Yildirim, I., Siegel, M., & Tenenbaum, J. (2016). Perceiving fully occluded objects via physical simulation. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 1265–1270.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.