THE JOURNAL OF PHILOSOPHY
VOLUME CX, NO. 9, SEPTEMBER 2013

EXPECTING THE WORLD: PERCEPTION, PREDICTION, AND THE ORIGINS OF HUMAN KNOWLEDGE*

Perception, I shall argue, is the successful prediction of the current sensory signal using stored knowledge about the world. This model of perception is increasingly common in cognitive scientific discourse.[1] But it has so far made little impact on philosophical theorizing.[2] Nonetheless, the model has intuitive appeal, is backed by increasing neuroscientific evidence, and has been shown to be

*Thanks to Miguel Eckstein, Mike Gazzaniga, Michael Rescorla, and the faculty and students at the Sage Center for the Study of Mind, University of California, Santa Barbara, where, as a Visiting Fellow in September 2011, I was privileged to road-test some of this material. Thanks too to Markus Werning and the organizers and participants of the 2010 meeting of the European Society for Philosophy and Psychology, held at Ruhr-Universität Bochum, August 2010; to Nihat Ay, Ray Guillery, Bruno Olshausen, Murray Sherman, Fritz Sommer, and the participants at the Perception and Action Workshop, Santa Fe Institute, New Mexico, September 2010; and to Karl Friston, Daniel Dennett, Peter König, Susanna Siegel, Mark Sprevak, Matteo Colombo, Matthew Nudds, and Bill Phillips.

[1] Classic treatments include David Mumford, "On the Computational Architecture of the Neocortex II: The Role of Cortico-Cortical Loops," *Biological Cybernetics*, LXVI, 3 (January 1992): 241–51; Rajesh Rao and Dana Ballard, "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *Nature Neuroscience*, II, 1 (January 1999): 79–87; Tai Sing Lee and Mumford, "Hierarchical Bayesian Inference in the Visual Cortex," *Journal of the Optical Society of America A*, XX, 7 (Jul. 1, 2003): 1434–48; and Karl Friston, "A Theory of Cortical Responses," *Philosophical Transactions of the Royal Society B: Biological Sciences*, CCCLX, 1456 (Apr. 29, 2005): 815–36. For a fairly accessible recent treatment, see Friston, "The Free-Energy Principle: A Unified Brain Theory?," *Nature Reviews: Neuroscience*, XI, 2 (February 2010): 127–38.

[2] Notable exceptions include work by Jakob Hohwy, including "Functional Integration and the Mind," *Synthese*, CLIX, 3 (December 2007): 315–28, and "Attention and Conscious Perception in the Hypothesis Testing Brain," *Frontiers in Psychology*, III, 96 (Apr. 2, 2012): 1–14; as well as work by Rick Grush and by Chris Eliasmith (see for example Grush, "The Emulation Theory of Representation: Motor Control, Imagery, and Perception," *Behavioral and Brain Sciences*, XXVII, 3 (April 2004): 377–96; Eliasmith, "How to Build a Brain: From Function to Implementation," *Synthese*, CLIX, 3 (December 2007): 373–88).

0022-362X/13/1009/469–496

computationally effective. It depicts the perceptual process as involving the Bayesian estimation of distal properties and features. If correct, it explains why perception, although carried out by the brain, cannot help but reach out to a distal world; it shows why that "reaching out" reveals a world that is already structured and to that extent (weakly) "conceptualized"; and it offers a new and powerful tool for thinking about debates concerning the origins and development of abstract knowledge.

Section I outlines the broad shape of the prediction-based model and illustrates the operation of a key component: the use of generative models to construct the sensory signal "from the top down." Section II discusses the crucial role of learning, with a special emphasis on the way prediction-based learning uncovers the deep structuring causes (or "latent variables") that best explain the shape of the current sensory signal. It is these deep structuring causes that we come to recognize as the external world, populated by familiar objects, features, and properties. Section III argues that the resulting model should fundamentally reconfigure debates concerning innate knowledge, revealing more of the hidden richness of statistical learning. Section IV raises and attempts to resolve some puzzles arising from this general picture of prediction-based perceptual contact with the world. Section V then asks what kind of perceptual relation to the world (direct, indirect, neither?) this account implies, and suggests that it may best be captured as "not-indirect perception." There is a short conclusion.

I. PREDICTING THE PRESENT

What happens when, after a brief chat with a colleague, I re-enter my office and visually perceive the hot, steaming, red cup of coffee that I left waiting on my desk? One possibility is that my brain receives a swathe of visual signals (imagine, for simplicity, an array of activated pixels) that specify a number of elementary features such as lines, edges, and color patches. Those elementary features are then progressively accumulated and (where appropriate) bound together, yielding shapes and specifying relations. At some point, these complex shapes and relations activate bodies of stored knowledge, turning the flow of sensation into world-revealing perception: the seeing of coffee, steam, and cup, with the steaming bound to the coffee, the color red to the cup, and so on. Call this the "passive accumulation" model of the perceptual process. Such a model, though here simplistically expressed, corresponds quite accurately to traditional cognitive scientific approaches (for example, by David Hubel and Torsten Wiesel, David Marr, and Irving

Biederman[3]) that depict perception as a cumulative process of "bottom-up" feature detection.

Here is an alternative scenario. As I re-enter my office my brain already commands a complex set of coffee-involving expectations. Glancing at my desk sets off a chain of visual processing in which current bottom-up signals are met by a stream of downward predictions concerning the anticipated states of various neuronal groups along the appropriate visual pathway. In essence, a multilayer downward cascade is attempting to "guess" the present states of all the key neuronal populations responding to the present state of the visual world. There ensues a rapid exchange (a dance between multiple top-down and bottom-up signals) in which incorrect guesses yield error signals that propagate forward and are used to extract better guesses. When top-down guessing adequately accounts for the incoming signal, the visual scene is perceived. As this process unfolds, top-down processing is trying to generate the incoming sensory signal for itself. When and only when this succeeds, and a match is established, do we get to experience (veridically or otherwise[4]) a meaningful visual scene.[5]

[3] David H. Hubel and Torsten N. Wiesel, "Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat," *Journal of Neurophysiology*, XXVIII, 2 (Mar. 1, 1965): 229–89; David Marr, *Vision* (San Francisco: Freeman, 1982); Irving Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review*, XCIV, 2 (April 1987): 115–47.

[4] Thus consider expert observers of, say, sports or chess. Such observers benefit from much richer structures of knowledge supporting their top-down predictions. But their brains may, as a result, sometimes overweight acquired expectations ("priors") relative to the driving sensory signal. Daily life, where we are all expert observers to some degree, provides many examples of such overweighting, for example when we constantly seem to see our familiar but temporarily absent pet in the subtle play of light and shadow.

[5] This alternative scenario has its roots in the work of Hermann von Helmholtz's vision of perception as a process of unconscious inference. See Helmholtz, *Handbuch der physiologischen optic* (Leipzig, Germany: Voss, 1867); English-language edition *Treatise on Physiological Optics, Volume III: The Perceptions of Vision*, ed. James P. C. Southall (New York: Dover, 1962). The idea, recast in contemporary terms, is that what we experience is the most *probable* worldly state, given what we already know about the world (our priors) and the current sensory stimulation (the evidence, or "likelihood"). Helmholtz's vision made its way into cognitive psychology as the paradigm known as "analysis-by-synthesis" (Ulric Neisser, *Cognitive Psychology* (New York: Appleton-Century-Crofts, 1967); and, for a recent review, Daniel Kersten and Alan L. Yuille, "Vision as Bayesian Inference: Analysis by Synthesis?," *Trends in Cognitive Sciences*, X, 7 (July 2006): 301–08). In the last two decades these broad visions were given effective computational flesh. Key publications include Geoffrey E. Hinton et al., "The 'Wake-Sleep' Algorithm for Unsupervised Neural Networks," *Science*, CCLXVIII, 5214 (May 26, 1995): 1158–61; and Rao and Ballard, "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects." For a useful review, see Hinton, "Learning Multiple Layers of Representation," *Trends in Cognitive Sciences*, XI,

This does not mean, of course, that perceptual experience occurs only after all forward-flowing error is eliminated. Percepts here take shape only when downward predictions match the incoming sensory signal, but this matching is itself a multilevel, piecemeal matter in which rapid perception of the general nature or "gist" of a scene is performed on the basis of first-pass matches established using general expectations and simple (for example, low spatial frequency) cues, perhaps in the manner recently described by Kveraga, Ghuman, and Bar.[6] Richer detail then emerges concurrently with the progressive reduction of residual error signals: a process that may also be mediated by attention, as argued by Harriet Feldman and Karl Friston.[7] Perception, if such models are correct, is a matter of the brain using stored knowledge to predict, in a progressively more refined manner, the patterns of multilayer neuronal response elicited by the current sensory stimulation. Thus described, the process of perception is one in which the brain is highly proactive, busily attempting to predict (at multiple levels) its own current internal states. Call this the "active self-prediction" model of the perceptual process.

To see how this works, it helps briefly to consider a somewhat different, but better specified, problem: that of recognizing spoken words in some natural language. Here too there is a traditional "passive accumulation" model according to which:

> Representations constructed at earlier stages of processing feed immediately higher levels in a feedforward manner…this process proceeds incrementally until access to a "lexical–conceptual" representation has been achieved. In speech recognition…this involves a conversion from acoustic features onto phonetic representations, phonetic representations onto phonological representations, and finally access of the lexical item based on its phonological structure.[8]

---

10 (October 2007): 428–34. Work by David Mumford ("On the Computational Architecture of the Neocortex II") and by Friston ("A Theory of Cortical Responses") makes suggestive contact with neurobiology and is consistent with compelling bodies of work in psychophysics and cognitive psychology showing that perception often conforms to Bayesian principles of optimal reasoning under uncertainty (for a review, see David C. Knill and Alexandre Pouget, "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation," *Trends in Neuroscience*, XXVII, 12 (December 2004): 712–19).

[6] Kestutis Kveraga, Avniel S. Ghuman, and Moshe Bar, "Top-Down Predictions in the Cognitive Brain," *Brain and Cognition*, LXV, 2 (November 2007): 145–68.

[7] Harriet Feldman and Friston, "Attention, Uncertainty, and Free-Energy," *Frontiers in Human Neuroscience*, IV, 215 (Dec. 2, 2010): 1–23.

[8] This description of the standard model is from David Poeppel and Philip J. Monahan, "Feedforward and Feedback in Speech Perception: Revisiting Analysis

The alternative, once again, is to use whatever stored knowledge is available to guide a set of guesses about the shape of the present sound stream, and then to compare those guesses to the incoming signal, using residual errors to decide between competing guesses and (where necessary) to reject one set of guesses and replace it with another. This is "analysis-by-synthesis,"[9] which David Poeppel and Philip Monahan summarize as a three-step process involving:

> (1) the extraction of (necessarily brief and coarse) cues in the input signal to elicit hypotheses, that while coarse, are sufficient to generate plausible guesses about classes of sounds (for example, plosives, fricatives, nasals, and approximants), and that permit subsequent refinement; (2) the actual synthesis of potential sequences consistent with the cues; and (3) a comparison operation between synthesized targets and the input signal delivered from the auditory analysis of the speech.[10]

Once we know enough about the (language-specific) structure of the sound stream and plausible flows of semantic content, we can use that knowledge actively to predict the incoming signal, thus anticipating the current sensory input. The process is kick-started by a few simple, rapidly processed sensory cues, but these immediately recruit a cascade of downwards-sweeping predictions. There ensues a complex dance between top-down predictions and increasingly detailed bottom-up signaling. The goal of the dance is to find the linked set of predictions, spanning multiple temporal and spatial scales, that best accounts for the signal. This is the "winning hypothesis." Along the way, further processing of the incoming signal is itself conditioned by competing hypotheses at many levels. Thus we sample and search the scene in ways determined by the brain's competing guesses. The extensive use of existing knowledge (driving the guessing) has many advantages, enabling us to hear what is said despite noisy surroundings, to adjudicate between alternate possibilities each consistent with the bare sound stream, and so on.[11]

---

by Synthesis," *Language and Cognitive Processes*, XXVI, 7 (2011): 935–51. The quoted passage is from p. 936. Poeppel and Monahan do not, however, endorse that traditional model, and instead argue for the alternative approach described here.

[9] See Kenneth N. Stevens and Morris Halle, "Remarks on Analysis by Synthesis and Distinctive Features," in W. Wathen-Dunn, ed., *Models for the Perception of Speech and Visual Form* (Cambridge: MIT, 1967), pp. 88–102.

[10] Poeppel and Monahan, "Feedforward and Feedback in Speech Perception: Revisiting Analysis by Synthesis," p. 939.

[11] A common (but fortunately misplaced) worry goes like this. Suppose you suddenly and unexpectedly awake in a brand new environment. None of your brain's first shots at predicting the sensory input would work. What then? In such cases the driving

Perceptual content, as delivered by such a process of active self-prediction, is inherently organized and outward-looking. By this I mean that it reveals—and cannot help but reveal—a structured (hence in some weak sense "conceptualized"[12]) external world. In this (admittedly restricted) sense, the world thus revealed is inherently meaningful. It is an external arena populated not by proximal stimulations but by distal, causally interacting items and forces whose joint action best explains the current suite of sensory stimulation. This is just the kind of grip on the world that an intelligent agent must possess if she is to act knowingly. When such an agent sees the world, she sees a structure of distal, interacting causes. That, I suggest, is why we perceive an external world and not "sense data": we must meet the transduced pixels with a top-down cascade of represented, interacting distal causes. The so-called "transparency" of perceptual experience[13]—the fact that, in normal daily perception, we seem to simply see tables, chairs, and bananas rather than to experience the more proximal innervations of our sensory surfaces—falls quite naturally out of such models.

To make good on these claims, we will shortly need to consider a vital further ingredient. That ingredient is learning, and it will occupy

---

sensory signal will generate very specific error messages (specific mismatches with your first attempts to predict it) that rapidly lead to the activation of multiple other models you already command ("gothic table to your left, ornate dragon sculpture to your right") so as to quash the error by "explaining away" the signal. In a very real sense forward-flowing prediction error here plays the role more traditionally assigned to the sensory input itself (see Feldman and Friston, "Attention, Uncertainty, and Free-Energy"). It can do this because the "error signal" in a trained-up predictive coding scheme is highly informative and carries detailed information about the mismatched content itself. For further discussion, see Andy Clark, "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences*, XXXVI, 3 (May 2013): 181–204.

[12] It is a weak sense because there is no guarantee that the potential thinking of beings deploying this strategy will form the kind of closed set (encompassing all possible combinations of the component grasps) required by the so-called Generality Constraint (Gareth Evans, *The Varieties of Reference* (Oxford: University Press, 1982)). I am persuaded by Peter Carruthers's argument ("Invertebrate Concepts Confront the Generality Constraint (and Win)," in Robert W. Lurz, ed., *The Philosophy of Animal Minds* (New York: Cambridge, 2009), pp. 89–107) rejecting the Generality Constraint as a necessary condition for either the having of thoughts or the possession of concepts. But those persuaded by the constraint may safely recast my claim as the putatively weaker assertion that beings deploying the prediction-based perceptual strategy are thereby placed in some form of cognitive contact with a structured external realm represented as populated by distinct, causally interacting items and entities.

[13] See, for example, George Edward Moore's "The Refutation of Idealism," reprinted in *Philosophical Studies* (London: Routledge and Kegan Paul, 1903/1922); and Gilbert Harman's 1990 paper, "The Intrinsic Quality of Experience," in James Tomberlin, ed., *Philosophical Perspectives 4* (Atascadero, CA: Ridgeview, 1990).

us in the next section. Before doing so, it will be useful to develop one last illustration of the central notion of meaningful, structured, prediction-based perception. The illustration, suggested to me by Daniel Dennett,[14] involves a device imagined (and subsequently built by the software engineer Steve Barney) as a means of preventing geology students from cheating at their assignments. The students could cheat by simply copying, from public sources, stratigraphic images that the assignment really required them to understand. A stratigraphy drawing—literally, the drawing of layers—is a cross-sectional depiction of rock formations and layerings that aims to reveal the way complex geological structures result from temporally sequenced combinations of interacting causes. Successful copying or tracing of such a drawing is, however, a poor indicator of a student's true geological grasp. To combat the problem, Dennett imagined a device that was later prototyped and dubbed SLICE.

SLICE ran on an original IBM PC and was essentially a drawing program whose action was not unlike that of the Etch-a-Sketch device many of us played with as children, except that this device controlled the drawing in a much more complex and interesting fashion. SLICE was equipped with a number of virtual "knobs," and each knob controlled the unfolding of a representation of a basic geological process—for example, one knob would deposit layers of sediment, another would erode, another would intrude lava, another would control fracture, another fold, and so on. The form of the assignment is then as follows: the student is given a stratigraphic drawing and is tasked with recreating the picture using the device. The only way to achieve this is by matching the drawing—by twiddling the right knobs, in the right order. Tracing and simple copying are not options on the device. Instead, the student must find the correct knobs and deploy them with the right intensities (each works like a kind of volume control) in the appropriate sequence for yielding the designated outcome. If a student could do that, Dennett reasoned, then she really did understand quite a lot about how hidden geological causes (like sedimentation, erosion, lava flow, and fracture) conspire to generate physical outcomes as captured by different stratigraphy drawings. The successful student would have to command a "generative model," enabling her to construct various geological outcomes from their component parts, as they interact in space and time.

We can take this further by requiring the student to command a *probabilistic* generative model. For a single presented picture there

[14] The illustration is used with Dennett's kind permission.

will often be a number of different ways of combining the various knob-twiddlings to yield it. Some of these will represent far more likely events and event combinations than others. To get full marks, then, the student should deploy the device so as to unearth the set of events (the set of "hidden geological causes") that are the *most likely* to have brought about the observed outcome. More advanced tests might involve explicitly ruling out the most common sets of causes, thus forcing the student to find an alternative way of bringing about that state (forcing her to find the *next most likely* set of causes, and so on). SLICE thus allows the user to deploy what she knows about geological causes (sedimentation, erosion, and so on) and how they interact to self-generate a stratigraphic image: one that, taking prior knowledge and any additional constraints into account, best accounts for the image given in the homework.

To complete the analogy, we need to remove the student user from the loop. SLICE* is just like SLICE, except that SLICE* incorporates its own internal model of how hidden geological causes combine to bring about the outcomes depicted by the stratigraphic images. When SLICE* is shown such an image, it automatically seeks to match that sensory input using a top-down generative cascade that puts together the most likely set of causes whose interaction would yield (and hence explain) the present input.

This is exactly the trick that the brain uses, if the models I am considering are on track, to make sense of the sensory signal received from the world. We perceive the world by activating (using a knowledge-driven top-down cascade) representations of the set of interacting external causes that make the sensory data most likely. The process settles when the interaction of a (now? previously?) linked set of hypothesized causes delivers, from the top down, patterns of neural activation that match (and thus "explain away") those resulting from the driving ("bottom-up") sensory signal. Otherwise put, the brain's job is to account for the sensory signal by finding a way to generate, in a kind of rolling present, that incoming signal for itself. To do this, the brain must find the structure in the signal. But the structure in the sensory signal is mostly determined by the structure in the world (making sure that is the case is pretty much the job description of a sensory transducer). So the best way to anticipate/match the incoming signal is to discover and deploy internal resources that amount to a kind of virtual-reality generator that models the distal elements and their typical modes of interaction (simplistically, if it generates "car" and "sudden braking" it might also generate "smoke from tires"). An agent perceives when

the virtual-reality generator can use its resources to capture (match, cancel out) the structure of the incoming signal. That implies putting together the right set of interacting distal causes, just as in the SLICE* example.

This process has an immediate, and compelling, Bayesian gloss. The system settles on the hypothesis (or better, the consistent linked set of hypotheses) that maximizes the posterior probability of the observed sensory data. It is this winning hypothesis that determines what we perceive. We see the scene, and not (for example) our retinal or neural activations, because perception just is a process of explaining away the sensory signal by finding the most likely set of interacting distal causes: the ones that, given prior probabilities and present evidence, best predict the incoming signal.

Notice, finally, that the real-world perceptual matching task targets not a single static outcome (as in SLICE*) but rather an evolving real-world scene. Matching the dynamic incoming signal here requires knowing how the elements of the scene will evolve and interact across multiple spatial and temporal scales. To actively recreate the incoming signal from the external scene with that kind of temporal and spatial acuity is, I suggest, to understand a lot about the world immediately. To perceive the world in this way is to deploy knowledge not just about how the sensory signal should be right now, but about how it will probably change and evolve over time. It is only by means of such longer-term and larger-scale knowledge that we can robustly match the incoming signal, moment to moment, with apt expectations (predictions). But to know that (to know how the present sensory signal is likely to change and evolve over time) just *is* to understand a lot about how the world is, and the kinds of entities and events that populate it. Creatures deploying this strategy, when they see the grass twitch in just that certain way, are already expecting to see the tasty prey emerge, and already expecting to feel the sensations of their own muscles tensing to pounce. An animal, or machine, that has *that kind of grip* on its world is already deep into the business of understanding that world.

II. INHERITING (THE DYNAMICS OF) THE EARTH

To complete this picture we need to address an obvious and important question. Where does all that knowledge (the knowledge that powers the active self-predictions that underlie perception) come from in the first place? It is an attractive feature of the story on offer that it is the *very same process* (that of attempting to predict the current sensory input) that underlies

both experience-driven learning[15] and online response. Moreover, it is a natural consequence of this process that the learner uncovers (when all is working correctly) the weave of interacting distal causes that characterizes the environment in which learning occurs. In this way (more on which in section III), prediction-based learning brings into view a structured external world, built of persisting (though often temporally evolving) objects, properties, and complex nested causal relations. The upshot, according to neuroscientists Stefan Kiebel, Jean Daunizeau, and Karl Friston, is that "the recognition system 'inherits' the dynamics of the environment and can predict its sensory products accurately."[16]

How might such learning occur? One possibility, of course, is that it simply does not, and that the bulk of the required knowledge is innate, gradually installed in the shape and functioning of our neural circuits over many millennia. For some time, it seemed as if this was the only plausible explanation of our ability to know the world on the basis of what seemed like slim and underdetermining sensory pickings.[17] Connectionist models of learning raised important doubts about such arguments, showing that it was possible to learn quite a lot from the statistically rich bodies of sensory data that we actually encountered.[18] But standard connectionist approaches were hampered in two ways. The first was the need to provide sufficient amounts of pre-categorized training data to drive the most powerful forms of learning, namely, those relying upon the so-called "back-propagation of error."[19] This is learning in which the current output (typically, some kind of categorization of the input) is compared to the correct output and connection weights are slowly adjusted to bring future response more and more into line. The second problem was the difficulty of training such networks in multilayer[20] forms, since this required distributing the

[15] In this context, "experience-driven" just means driven by the incoming streams of sensory information. Such incoming information need not (though it may) be consciously experienced.

[16] Stefan J. Kiebel, Jean Daunizeau, and Friston, "Perception and Hierarchical Dynamics," *Frontiers in Neuroinformatics*, III, 20 (2009): 1–9, at p. 7.

[17] For a review, see Steven Pinker's *How the Mind Works* (New York: Norton, 1997).

[18] See for example Jeffrey L. Elman, "Connectionist Models of Cognitive Development: Where Next?," *Trends in Cognitive Sciences*, IX, 3 (March 2005): 111–17.

[19] See David E. Rumelhart, Hinton, and Robert J. Williams, "Learning Internal Representations by Error Propagation," in Rumelhart, James L. McClelland, and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (Cambridge: MIT, 1986), pp. 318–62.

[20] By this I mean, in forms that multiplied the layers of so-called "hidden units" intervening between input and output. For a nice discussion of these difficulties, see Hinton, "Learning Multiple Layers of Representation." For an application, see

response to the error signal in hard-to-determine ways across all the layers. Yet such multilayer forms, as we shall see in more detail in section III, are the key to learning about our kind of world: a world that is highly structured, displaying regularity and pattern at many spatial and temporal scales, and populated by a wide variety of interacting and complexly nested causes.

It is here that active self-prediction and hierarchical learning mark a real advance over previous work. This work builds upon crucial advances in machine learning that began with work on the aptly named "Helmholtz Machine,"[21] which was an early example of a multilayer architecture trainable without reliance upon pre-classified examples. Instead, the Helmholtz Machine "self-organized" by attempting to generate the training data for itself using top-down connections. That is to say, instead of starting with the task of classifying the data, it had first to learn how to generate the incoming data for itself. (This corresponds to the imaginary version of SLICE, SLICE*, in which the machine itself must learn, from the incoming pixel patterns, the set of combinable geological causes that would give rise to those very patterns.)

This can seem an impossible task, since performing the generation requires the very knowledge (about how the training items might be systematically generated by complex interacting causes) that the system is hoping to acquire. For example, to generate the phonetic structures proper to some public language you would need already to know a lot about the various speech sounds and how they are (in that language) articulated and combined.[22] In other words, a system could learn to perform the classification task (taking sound streams as input and delivering a phonetic parse as output) if it already commanded a generative model of phonetically structured speech in the language. Conversely, it could learn to perform the generation task if a recognition model (one supporting

---

Hinton, "To Recognize Shapes, First Learn to Generate Images," in P. Cisek, T. Drew, and J. Kalaska, eds., *Computational Neuroscience: Theoretical Insights into Brain Function* (Boston: Elsevier, 2007), pp. 535–48.

[21] Peter Dayan et al., "The Helmholtz Machine," *Neural Computation*, VII, 5 (September 1995): 889–904; and Dayan and Hinton, "Varieties of Helmholtz Machine," *Neural Networks*, IX, 8 (November 1996): 1385–403. See also Hinton and Richard S. Zemel, "Autoencoders, Minimum Description Length and Helmholtz Free Energy," in Jack D. Cowan, Gerald Tesauro, and J. Alspector, eds., *Advances in Neural Information Processing Systems 6* (San Mateo, CA: Morgan Kaufmann, 1994).

[22] For a nice account defending the prediction-based model described in the text, see Poeppel and Monahan, "Feedforward and Feedback in Speech Perception: Revisiting Analysis by Synthesis."

the classification task) was already in place. The problem was solved, in principle at least, by the development of algorithms (such as the so-called "wake-sleep algorithm"[23]) that, starting with random weight assignments, used each task to bootstrap the other.[24]

The Helmholtz Machine was an early version of what is now known as a "deep architecture,"[25] where these are multilayer networks capable of powerful forms of data-driven learning. Recent advances in learning using such deep architectures lead Hinton to assert that:

> The limitations of backpropagation learning can now be overcome by using multilayer neural networks that contain top-down connections and training them to generate sensory data rather than to classify it.[26]

In hierarchical predictive coding schemes[27] this kind of process is implemented by the ongoing attempt to use top-down connections to predict the incoming signal. The sensory signal is met by a top-down "guess" constructed using multiple layers of downward influence, and any mismatch is passed forward as an error signal. There are several worked-out examples of this in the literature, and I discuss some of them in more detail elsewhere.[28] An early paper by Rajesh Rao and Dana Ballard provides the classic proof-of-concept, in which prediction-based learning targets image patches drawn from natural scenes.[29] In this work a multilayer artificial neural

[23] This was a computationally tractable approximation to "maximum-likelihood learning" as used in the expectation-maximization (EM) algorithm. See Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, xxxix, 1 (1977): 1–38. It allowed the system to learn both the recognition and the generation models by training both sets of weights (starting from small random assignments) in an alternating fashion.

[24] Radford M. Neal and Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants," in Michael I. Jordan, ed., *Learning in Graphical Models* (Dordrecht: Kluwer, 1998), pp. 355–68; Hinton et al., "The 'Wake-Sleep' Algorithm for Unsupervised Neural Networks."

[25] For a review of learning using these deep architectures, see Yoshua Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, ii, 1 (Nov. 15, 2009): 1–127.

[26] Hinton, "Learning Multiple Layers of Representation," p. 428.

[27] See note 1 for some of the classic treatments. In addition, and with a broader and perhaps more philosophical slant, see Friston and Klaas E. Stephan, "Free-Energy and the Brain," *Synthese*, clix, 3 (December 2007): 417–58. See also Janneke F. M. Jehee and Ballard, "Predictive Feedback Can Account for Biphasic Responses in the Lateral Geniculate Nucleus," *PLoS Computational Biology*, v, 5 (May 2009): e1000373. For a review, see Yanping Huang and Rao, "Predictive Coding," *Wiley Interdisciplinary Reviews: Cognitive Science*, ii, 5 (September/October 2011): 580–93.

[28] Clark, "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science."

[29] Rao and Ballard, "Predictive Coding in the Visual Cortex."

network, with no preset task apart from that of using the downward flow to match inputs with successful predictions, developed a nested structure of units with simple-cell-like receptive fields, while capturing a variety of important, empirically observed receptive field effects.[30] The system learnt to use units in one layer to track simple features such as oriented bars and edges, while units in the next layer tracked more complex patterns involving the spatial combination of such features (repeating bars, or stripes, for example).

These were early, limited, and relatively low-level results, but the learning model itself has proven rich and widely applicable.[31] It assumes only that the environment generates sensory signals by means of nested interacting causes and that the task of the perceptual system is to invert this structure by learning and applying a hierarchical generative model so as to predict the unfolding sensory stream. Learning routines of this kind have recently been successfully applied in many domains, including speech perception, reading, and recognizing the actions of oneself and of other agents.[32] This is not surprising, since the underlying rationale is also quite general. If you want to predict the way some set of sensory signals will change and evolve over time, a good thing to do is to learn how those sensory signals are determined by interacting external causes, for the flow of sensation is predictable just to the extent that there is spatial and temporal pattern in that flow. But such pattern is a function of the properties and features of external

[30] Especially so-called "non-classical receptive field" effects—see also Rao and Terrence Sejnowksi, "Predictive Coding, Cortical Feedback, and Spike-Timing Dependent Cortical Plasticity," in Rao, Bruno A. Olshausen, and Michael S. Lewicki, eds., *Probabilistic Models of the Brain: Perception and Neural Function* (Cambridge: MIT, 2002), pp. 297–316.

[31] Recent applications include Jakob Hohwy, Andreas Roepstorff, and Karl Friston's model of binocular rivalry ("Predictive Coding Explains Binocular Rivalry: An Epistemological Review," *Cognition*, CVIII, 3 (September 2008): 687–701), and Tobias Egner, James Michael Monti, and Christopher Summerfield's compelling experimental exploration of predictive coding as a model of ventral-stream responses to faces and to non-face stimuli ("Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream," *Journal of Neuroscience*, XXX, 49 (Dec. 8, 2010): 16601–08). Related, though in some important ways computationally distinct, demonstrations include the benchmark work in machine learning due to Geoffrey Hinton and colleagues—see Hinton, "Learning to Represent Visual Input," *Philosophical Transactions of the Royal Society B: Biological Sciences*, CCCLXV, 1537 (Jan. 12, 2010): 177–84.

[32] For these examples see (respectively) Poeppel and Monahan, "Feedforward and Feedback in Speech Perception: Revisiting Analysis by Synthesis"; Cathy Price and Joe Devlin, "The Interactive Account of Ventral Occipitotemporal Contributions to Reading," *Trends in Cognitive Sciences*, XV, 6 (June 2011): 246–53; Friston, Jérémie Mattout, and James M. Kilner, "Action Understanding and Active Inference," *Biological Cybernetics*, CIV, 1–2 (February 2011): 137–60.

objects, and of their interactions with each other and with the agent.[33] Thus the pattern of sensory stimulation reaching the eye from, say, an observed football game is a function of the lighting conditions, the players, the observer, and their respective motions. It is also a function of a variety of more abstract interacting features and forces, including the patterns of offense and defense characteristic of each team, the current state of play (there are strategic alterations when one team is far behind, for example), and so on. The beauty of the various waves of work in computational neuroscience and machine learning just described is that they begin to show how to learn about such complex stacks of interacting causes without requiring extensive prior knowledge. This should fundamentally reconfigure our thinking about both the debate between nativism and empiricism, and about the nature and possibility of a view of perception as "carving nature at the joints," as we shall now see.

III. LAYERS UPON LAYERS: LEARNING FROM THE TOP DOWN

The prediction task (which we may gloss as "guess the sensory input") allows a system to learn without the provision of pre-classified training data. This strategy is most potent, however, when it is applied in multilayer, hierarchical settings. Recent advances in machine learning, as described above, provide an existence proof of the possibility of successful learning in such multilayer settings.[34] In addition, work in computational and cognitive neuroscience makes plausible proposals concerning the neural implementation of multilayer, prediction-driven learning.[35] Taken together, these results have the potential fundamentally to alter the landscape of debates concerning innate knowledge and the possibility of strong, rational, and world-revealing forms of learning.

[33] For simplicity, I shall not here pursue the important contribution made by the active embodied agent, but this contribution may be treated in just the same way. A creature's body and self-generated motions are additional hidden causes of sensory variation, and their nature and properties may be unearthed using the same learning routines. See Friston et al., "Action and Behavior: A Free-Energy Formulation," *Biological Cybernetics*, CII, 3 (March 2010): 227–60.

[34] Important computational differences separate the various bodies of work appealed to in section II. These differences mostly concern the precise ways in which top-down expectations and bottom-up sensory signals are combined, both in learning and during online response. Although significant, these differences may safely be ignored for present purposes.

[35] See the review by Huang and Rao, "Predictive Coding"; and the important body of work by Friston and collaborators, usefully summarized in Friston's "The Free-Energy Principle: A Rough Guide to the Brain?," *Trends in Cognitive Sciences*, XIII, 7 (July 2009): 293–301.

The central plank of this reconfiguration is the capacity of these forms of learning to underwrite the development of so-called Hierarchical Bayesian Models (HBMs).[36] A Hierarchical Bayesian Model is one in which multiple layers of processing are interanimated in an especially potent way, with each layer attempting to account for the patterns of activation (encoding some probability distribution of variables) at the level below. This, of course, is precisely the architecture mandated, at the so-called "process level,"[37] by hierarchical predictive coding. When such a system is up and running, mini-hypotheses at all these multiple levels settle into the mutually consistent set that best accounts for the incoming sensory signal, taking into account what the system has learnt and the present sensory evidence (including the system's best estimation of the reliability of that evidence[38]). During prediction-driven learning, this implements a powerful multilayer form of Bayesian inference in which each layer is trying to build knowledge structures that will enable it to generate the patterns of activity occurring at the level below. In Bayesian terms, each layer is learning "priors"[39] on the level below. This whole multilayer process is driven by the incoming sensory signal and implements the strategy known as "empirical Bayes,"[40] in which a system acquires its own priors from the data as learning

[36] See Joshua B. Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science*, CCCXXXI, 6022 (Mar. 11, 2011): 1279–85; and Charles Kemp, Amy Perfors, and Tenenbaum, "Learning Overhypotheses with Hierarchical Bayesian Models," *Developmental Science*, X, 3 (May 2007): 307–21.

[37] The process level here corresponds to what Marr, in *Vision*, described as the level of the algorithm.

[38] Such assessments of reliability, sometimes referred to as the computed "precision" of the sensory signal, are plausible processing correlates for at least some (and perhaps all) varieties of attention. For discussion, see Feldman and Friston, "Attention, Uncertainty, and Free-Energy"; and Hohwy, "Attention and Conscious Perception in the Hypothesis Testing Brain."

[39] Priors are just prior probabilities, and they can take many forms. In the works cited, they mostly take the form of "probability density functions" or PDFs. Such a function assigns a distribution of probabilities across an uncountably large population, relative to which the observed data are treated as a random sample. In systems that learn hierarchical generative models to explain sensory inputs, probability density functions encode each level's knowledge about the level below. Considered in the most general terms, the role of such PDFs is to enable the system to compute the posterior density, where this names the likelihood of some candidate cause, given the stored knowledge and the current input. For an accessible introduction, written with philosophers in mind, see Michael Rescorla, "Bayesian Perceptual Psychology," to appear in Mohan Matthen, ed., *The Oxford Handbook of the Philosophy of Perception* (New York: Oxford, in press).

[40] See Herbert E. Robbins, "An Empirical Bayes Approach to Statistics," in Jerzy Neyman, ed., *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Berkeley: California UP, 1956), pp. 157–63.

proceeds. It does this by using its best current model—at one level—as the source of the priors for the level below, engaging in a bootstrap-type process of "iterative estimation"[41] that allows priors and models to co-evolve across the multiple linked layers of processing.

Such multilayer learning has an additional benefit, in that it lends itself very naturally to the combination of data-driven statistical learning with the kinds of systematically productive knowledge representation long insisted upon by the opponents of early work in connectionism and artificial neural networks.[42] To see this in microcosm, we need only reflect that SLICE* (as described in section 1 above) effectively embodies a productive and systematic body of knowledge concerning geological causes, for it can produce the full set of geological outcomes allowed by the possible combinations and recombinations of hidden causes represented in its generative model. By combining the use of multilayer generative models with powerful forms of statistical learning (indeed, using that learning to induce those very models) we secure many of the benefits of both early connectionist ("associationist") and more classical ("rule-based") approaches. Moreover, there is no need to fix on any single form of knowledge representation. Instead, each layer is free to use whatever form of representation best enables it to predict and (thus) account for the activity at the level below. In many cases, what seems to emerge are structured, productive bodies of knowledge that are nonetheless acquired on the basis of multistage learning driven by the statistical regularities visible in the raw training data.[43] Early learning here induces overarching expectations (for example, very

[41] See Dempster, Laird, and Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm"; and Neal and Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants."

[42] The classic critique is that of Jerry A. Fodor and Zenon W. Pylyshyn ("Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition*, XXVIII, 1–2 (March 1988): 3–71), but related points were made by more ecumenical theorists, such as Paul Smolensky ("On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, XI, 1 (March 1988): 1–23), whose later work on optimality theory and harmonic grammar (Smolensky and Géraldine Legendre, *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, 2 vols. (Cambridge: MIT, 2006)) likewise accommodates both generative structure and statistical learning.

For further discussion of this important issue, see Morten H. Christiansen and Nick Chater, "Constituency and Recursion in Language," in Michael A. Arbib, ed., *The Handbook of Brain Theory and Neural Networks* (Cambridge: MIT, 2003), pp. 267–71.

[43] For an excellent discussion of this attractive feature of hierarchical Bayesian approaches, see Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction." Caution is still required, however, since the mere fact that multiple forms of knowledge representation *can* coexist within such models does not show us, in any detail, how such various forms may effectively be combined in unified problem-solving episodes.

broad expectations concerning what kinds of things matter most for successful categorization within a given domain). Such broad expectations then constrain later learning, reducing the hypothesis space and enabling effective learning of specific cases.

Using such routines, HBMs have recently been shown capable of learning the deep organizing principles for many domains, on the basis of essentially raw data. Such systems have learnt, for example, about the so-called "shape bias" according to which items that fall into the same object category (like cranes, balls, and toasters) tend to have the same shape: a bias that does not apply to substance categories such as gold, chocolate, or jelly.[44] They have also learnt about the kind of grammar (context-free or regular) that will best account for the patterns in a corpus of child-directed speech,[45] about the correct parsing into words of an unsegmented speech stream,[46] and generally about the shape of causal relations in many different domains (for example, diseases cause symptoms, and not vice versa).[47] Recent work has also shown how brand new categories, defined by new causal schemas, can be spawned when assimilation to an existing category would require an overly complex—hence effectively "ad hoc"—mapping.[48] More recently still, such approaches have been shown to be capable of rapidly learning highly abstract domain-general principles (such as a general understanding of causality) by pooling evidence obtained across a wide range of cases.[49] Taken together, this work demonstrates the unexpected power of learning using HBMs.[50] Such approaches allow systems to infer the high-level structure specific to a domain, and even the high-level structures governing multiple domains, by exposing a multilevel learning system to raw data.

[44] Kemp, Perfors, and Tenenbaum, "Learning Overhypotheses with Hierarchical Bayesian Models."

[45] Perfors, Tenenbaum, and Terry Regier, "Poverty of the Stimulus? A Rational Approach," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum, 2006).

[46] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson, "A Bayesian Framework for Word Segmentation: Exploring the Effects of Context," *Cognition*, CXII, 1 (July 2009): 21–54.

[47] Vikash Mansinghka et al., "Structured Priors for Structure Learning," in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* (Arlington, VA: AUAI, 2006), pp. 324–31.

[48] Griffiths et al., "Categorization as Nonparametric Bayesian Density Estimation," in Chater and Mike Oaksford, eds., *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (New York: Oxford, 2008), pp. 303–28.

[49] Noah Goodman, Tomer D. Ullman, and Tenenbaum, "Learning a Theory of Causality," *Psychological Review*, CXVIII, 1 (January 2011): 110–19.

[50] For a thorough review, see Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction."

An important point to notice is that HBMs here allow the learner to acquire the schematic relations characteristic of a domain before "filling in" the details concerning individual exemplars. For example, as noted by Kemp, Perfors, and Tenenbaum:

> a hierarchical Bayesian model of grammar induction may be able to explain how a child becomes confident about some property of a grammar even though most of the individual sentences that support this conclusion are poorly understood.[51]

Similarly, the shape bias for objects may be learnt before learning the names of any of the individual objects. The bias emerges early as the best high-level schema, and once in place it enables rapid learning about specific exemplars falling into that group. This is possible in cases where "a child has access to a large number of…noisy observations [such that] any individual observation may be difficult to interpret, but taken together they may provide strong support for a general conclusion."[52] Thus, the authors continue, one might have sufficient evidence to suggest that visual objects tend to be "cohesive, bounded, and rigid"[53] before forming any ideas about individual concrete objects such as balls, discs, stuffed toys, and so on.

This is, of course, precisely the kind of "top-down" early-acting learning pattern that is easily mistaken as evidence of the influence of innate knowledge about the world. The mistake is natural since the high-level knowledge is tailored to the domain and allows subsequent learning to proceed much more easily and fluently than might otherwise be expected. But instead of thus relying on rich bodies of innate knowledge, HBM-style learners induce such abstract structuring knowledge from the data. The central trick, as we just saw, is to use the data itself in a kind of multistage manner. First the data are used to learn priors that encode expectations concerning the large-scale shape of the domain (what Tenenbaum et al. call the "form of structure" within the domain[54]). Suitably scaffolded by this structure of large-scale (relatively abstract) expectations, learning about more detailed regularities becomes possible. In this way, HBMs actively unearth the abstract structural expectations that enable them to use raw data to learn more and more finely grained models (supporting more finely grained sets of expectations).

[51] Kemp, Perfors, and Tenenbaum, "Learning Overhypotheses with Hierarchical Bayesian Models," p. 318.

[52] *Ibid.*

[53] Elizabeth S. Spelke, "Principles of Object Perception," *Cognitive Science*, XIV, 1 (January–March 1990): 29–56.

[54] Tenenbaum et al., "How to Grow a Mind: Statistics, Structure, and Abstraction."

Such systems are able to induce their own so-called "hyperpriors" from the data. Hyperpriors (here used interchangeably with the "overhypotheses" of Kemp et al.[55]) are essentially "priors upon priors" embodying systemic expectations concerning very abstract (at times almost Kantian) features of the world. For example, one highly abstract hyperprior might demand that each set of multimodal sensory inputs have a single best explanation. This would enforce a single peak for the probabilistic distributions consequent upon sensory stimulation, so that we always see the world as being in one determinate state or another, rather than (say) as a superposition of equiprobable states. Such a hugely abstract hyperprior might be a good candidate for innate specification. But it might equally well be left to early learning, since the need to use sensory input to drive actions, and the physical impossibility of acting in two very different ways at once, could conceivably (as Karl Friston, in personal communication, has suggested to me) drive an HBM to extract even this as a general principle governing inference.

HBMs (and the various process models, including hierarchical predictive coding, that might implement them) thus absolve the Bayesian theorist of the apparent sin of needing to set the right priors in advance of successful learning. Instead, in the manner of empirical Bayes, a multilayer system can learn its own priors from the data. This also delivers maximal flexibility. For although it is now easy to build abstract domain structure which reflects knowledge (in the form of various hyperpriors) into the system, it is also possible for the system to acquire such knowledge, and to acquire it in advance of the more detailed learning that it both streamlines and makes possible. Innate knowledge thus conceived remains "developmentally open" in that it can be smoothed and refined, or even completely undone, by data-driven learning using the same multilayer process.[56]

Of course, as King Lear famously commented, "nothing will come of nothing," and, as hinted above, even the most slim-line learning system must start with some set of biases.[57] Nonetheless, these

[55] Kemp, Perfors, and Tenenbaum, "Learning Overhypotheses with Hierarchical Bayesian Models."

[56] For some nice discussion, see Brian J. Scholl, "Innateness and (Bayesian) Visual Perception: Reconciling Nativism and Development," in Carruthers, Stephen Laurence, and Stephen Stich, eds., *The Innate Mind: Structure and Contents* (New York: Oxford, 2005), pp. 34–52.

[57] For example, a system might start with a set of so-called "perceptual input analyzers" (Susan Carey, *The Origin of Concepts* (New York: Oxford, 2009)) whose effect is to make a few input features more salient for learning. For discussion of the combined effects of HBM learning and such simple biases, see Goodman, Ullman, and Tenenbaum, "Learning a Theory of Causality."

multilayer Bayesian systems have proven capable of acquiring abstract, domain-specific principles without building in the kinds of knowledge (for example, about the importance of shape for learning about material objects) that subsequently account for the ease and efficacy of learning in different domains. Such systems acquire, from the raw data, knowledge of the kinds of abstract organizing principle that *then* allow them to make systematic sense of that very data. This is a very neat trick, indeed, and a suitable antidote to fears of losing the world behind a "veil of perception."

IV. PUZZLES FOR HBMS AND THE PREDICTIVE MIND

The hierachical Bayesian story, as it might be implemented using the distinctive resources of the active self-prediction approach to perception and learning, offers a novel account of our perceptual contact with the world. It makes perceiving dependent upon use of structured internal models capable of generating the sensory signal "from the top down," and it shows (arguably for the first time) how to combine statistical learning with the kinds of systematically productive knowledge representation long insisted upon by the opponents of earlier work in connectionism and artificial neural networks.

Despite these attractions, both the appeal to HBMs and the active self-prediction account of learning and perceptual inference leave many questions unanswered, and they each raise further puzzles in their own right. Some of these are addressed in other treatments.[58] I shall restrict my comments to two key issues arising directly from the overarching picture of perception as the top-down deployment of a multilayer generative model. The first issue concerns what might be termed "agentive perceptual content." The question here is how to bridge the apparent gap between the sub-personal accounts on offer and the kinds of perceptual content that seem to characterize the mental life of a human agent. The second issue comes into focus once we have made some prima facie progress with the first, and concerns the fundamental nature of our implied perceptual contact with the world.

[58] For example, the predictive processing model seems to imply the co-emergence of the faculties of perception, understanding, and imagination. Questions therefore arise concerning the correct way to reconstruct the phenomenal and epistemological differences between those faculties. Further questions concern the relation between these approaches and work that stresses instead the fragmentary, "quick-and-dirty" nature of much evolved problem solving. For some discussion, see my "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." Finally, there are a variety of questions that arise from recent attempts to extend the predictive coding story so as to encompass action and the observation of action. See for example Friston, Mattout, and Kilner, "Action Understanding and Active Inference."

Concerning agentive perceptual content, consider the following case. Geoffrey Hinton and Vinod Nair describe a benchmark machine-learning system for handwritten digit recognition.[59] The system's task is to classify handwritten digits (1s, 2s, 3s, and so on). That is to say, the system aims to take highly variable handwritten digits as inputs, and output the correct numerical classification. This is a difficult task. Considered at the pixel level, a badly written 2 will often have more in common (pixel-by-pixel) with a 3 than with the other 2s. Yet we humans manage the task surprisingly well. The reason we do so may be because our visual systems deploy a powerful generative model based not on raw pixel patterns but on the motor programs that would generate those patterns. Hinton and Nair constructed an artificial neural network that first learns to generate different handwritten digits using a (simplified) simulated motor routine and then uses that knowledge (knowledge about the way different digits are constructed by the simplified motor routine) for the classification task. The system thus classifies the handwritten digits (including many that were never shown in training) by in effect asking itself which of a set of known "canonical" motor routines (one for 1s, one for 2s, one for 3s, and so on) would have been most likely, under mild deformations, to yield a target scribble. In a large proportion of the difficult cases, this solves the problem, showing that the surface (pixel-level) similarity between a badly written 2 and a 3 is merely skin-deep. From the pixel-level input, the motor routine that would have been needed to generate the deviant 2 can be inferred. When that motor routine is a degraded version of the one associated with 2-production, and not the one associated with 3-production, correct classification is achieved. It seems plausible that humans perform this task using just this kind of knowledge. Such a procedure would explain why we are so often able to succeed despite large surface differences in fonts, handwriting, size of strokes, and so on.

It might be suspected (as it was by Mark Sprevak, in personal communication) that the motor routines required to produce these different styles of 2s will themselves be so different as to raise similar puzzles concerning how they are grouped together. Fortunately, this is not the case. From the training data the network infers a single canonical motor program as the "hidden variable" underlying all the handwritten 2s. It is then able to self-generate digit forms and

[59] Hinton and Vinod Nair, "Inferring Motor Programs from Images of Handwritten Digits," in Yair Weiss, Bernhard Schölkopf, and John Platt, eds., *Advances in Neural Information Processing Systems 18* (Cambridge: MIT, 2006), pp. 515–22. See also Hinton, "To Recognize Shapes, First Learn to Generate Images."

can recognize novel variations (new ways of writing recognizable 2s) by exploring the effects of small variations to the outputs of that program. Digits are then classified according to which canonical motor program yields some observed deviant form with least alteration.

But a puzzle now arises. For the internal states that explain the successful classification behavior here concern motor programs for writing digits, yet the content of the percept itself, at least as I have described it, is not motoric at all. What perception here delivers to the agent is a world parsed into differently scribbled versions of the various digits. Notice that we should not see this as ("merely") a question about conscious content versus the contents of sub-personal enabling states. The worry is deeper and more general than that. How, given the apparatus on offer, are we to render intelligible the fact that such an agent represents a world populated not by retinal stimulations or by motor programs but by specific distal entities (in this case, various handwritten digits)? This problem, as Tyler Burge rightly insists, would remain even were the creature in question unable to enjoy conscious perceptual experience at all.[60] What is required is a compelling reason to describe this sensory routine as representing handwritten digits (rather than something more proximal).

Recall that the network, in classifying the handwritten digits, deployed knowledge of motoric hidden causes (acting as the "latent variables" in a generative model). The very fact that the system *needed* to unearth such latent variables to perform the visual classification task provides the solution to the puzzle. For the need for latent variables arises only when proximal stimulations are *not* directly groupable in the right way. Working directly with the pixel-level stimulations an agent would, we saw, be forced to misclassify many perfectly recognizable 2s as 3s, and so on. It is only in virtue of the processing "detour" that infers motoric hidden causes that the network can parse its visual inputs into the correct digit classes, tracking handwritten one-ness, two-ness, and three-ness through the obscuring fog of handwriting variation. A good explanation of what the system is doing thus cannot depict the system as responding merely to the proximal (pixel-level) visual stimulations that impinge on its surfaces.

The moral here is perfectly general. Whenever systems use generative models and latent variables to enable them to parse the environment in the ways demanded by their needs to act and classify (or in this case, in the ways artificially imposed by the theorist who set up the task) we must explain their responses in part by reference

[60] Tyler Burge, *Origins of Objectivity* (New York: Oxford, 2010).

to the distal entities or attributes that are being tracked, rather than by reference to more proximal sensory stimulations (such as the pixel-level patternings) alone. The cases we have described thus satisfy the key requirement that perception, properly so called, occurs only when there is "explanatory need to attribute representation of distal attributes, as distinguished from registration of proximal stimulation."[61] Finally, the conscious cognizer's personal-level perspective on her own behavior should surely follow suit. For the agent, what matters are the action-salient states of the distal environment (including her own body and other agents). In this case, the action-salient environment is one populated by various visually presented handwritten digits. The agent knows that environment in ways served and made possible (in this case) by an unexpectedly motoric generative model. But it is an environment parsed according to her agent-level needs and purposes.

V. "NOT-INDIRECT" PERCEPTION

How should we categorize the model of perception just sketched? The model, as we saw, invokes a top-down cascade of processing to predict (hence "explain away") the current sensory input. Such a conception, it has sometimes been suggested, depicts perception as a process of "controlled hallucination"[62] in which the brain tries to guess what is out there, using stored knowledge recruited and then nuanced by the prediction error signals reporting residual mismatches between best guesses and sensory input. Thus described, the process may seem to invite easy classification among so-called "indirect realist" views of perception such as that of Frank Jackson.[63] To be sure, the accounts on offer unequivocally depict perception as an inferential process in the manner of Helmholtz,[64] and it is presumably with that dimension in mind that Jakob Hohwy recently commented that:

> One important and, probably, unfashionable thing that this theory tells us about the mind is that perception is indirect.…What we perceive is the brain's best hypothesis, as embodied in a high-level generative model, about the causes in the outer world.[65]

[61] *Ibid.*, p. 422.

[62] The phrase "perception as controlled hallucination" is variously attributed to the neuroscientist Ramesh Jain and to the machine-learning theorist Max B. Clowes (see his "On Seeing Things," *Artificial Intelligence*, II, 1 (Spring 1971): 79–116). I think I may have first heard it from Rodolfo Llinás.

[63] Frank Jackson, *Perception: A Representative Theory* (Cambridge, UK: University Press, 1977).

[64] Helmholtz, *Treatise on Physiological Optics, Volume III: The Perceptions of Vision.* See also Irvin Rock, *Indirect Perception* (Cambridge: MIT, 1997).

[65] Hohwy, "Functional Integration and the Mind," p. 323.

As it stands, however, this cannot be quite the right way to capture (or at least to express) the indirectness at issue here. There is no sense, even assuming the prediction-driven account is accepted, in which *what* we perceive is the brain's best hypothesis. Instead, what we perceive is the world, as (hopefully) revealed by the best hypothesis. Nor is there any sense in which the objects of perception are here being treated as anything like Moorean "sense data,"[66] where these are conceived as proxies intervening between the perceiver and the world. The internal representations at issue function *within* us, and are not encountered *by* us. Instead, they make it possible for us to encounter the world. Moreover, they enable us to do so under the ecologically common conditions of noise, uncertainty, and ambiguity. Brains that work like this are statistical wizards able to lock on to many of the causal chains (some of them, as we saw, highly abstract in nature) that actually give rise to the sensory data. The result, as one leading theorist puts it, is that:

> The hierarchical structure of the real world literally comes to be "reflected" by the hierarchical architectures trying to minimize prediction error, not just at the level of sensory input but at all levels of the hierarchy.[67]

The "real world" here is, of course, not (at least in the first instance) the world of atoms or of quantum physics but the more agent-salient world of tables, chairs, football games, dogs, cats, and (as J. L. Austin famously quipped) "medium-sized dry goods." This is a world (as we saw in section IV) parsed for human action and choice, and revealed by the top-down deployment of a hierarchical generative model. Seen in this light, the account on offer shares as much (or so it seems to me) with direct as with indirect views of perception, for it delivers a genuine form—perhaps the only genuine form that is naturally possible—of "openness to the world."[68] Against this, however,

[66] Moore, "The Status of Sense-Data," *Proceedings of the Aristotelian Society*, XIV (1913–1914): 355–80, reprinted in Moore, *Philosophical Studies*.

[67] Friston, "Beyond Phrenology: What Can Neuroimaging Tell Us about Distributed Circuitry?," *Annual Review of Neuroscience*, XXV, 1 (2002): 221–50. The quoted passage is from pp. 237–38.

[68] Considered as delineating a sub-personal mechanism that thus delivers "openness to the world," the present view might even (strange as this may sound) be cast as a representationalist version of "direct perception." Such a representationalist version of direct perception would be fundamentally different from those championed by Gibson and the ecological psychologists (see James J. Gibson, *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin, 1979)) precisely insofar as those accounts seem to reject the explanatory need to appeal to complex internal representational cascades.

it must be conceded that extensive reliance on the top-down cascade makes veridical perception sub-personally inferential and highly dependent upon prior knowledge. I shall not attempt further to adjudicate this delicate issue here.[69] If a label is required, it has been suggested (Michael Rescorla, in personal communication) that the implied metaphysical perspective may most safely be dubbed "not-indirect perception."

It is revealing, finally, to notice that content fixation in these accounts is arguably externalist in nature. Perceptual states function to estimate properties and features of the distal environment (including, for these purposes, states of our own bodies and the mental states of other agents). Such states are individuated by reference to the world actually sampled. Thus Michael Rescorla notes that Bayesian approaches to perceptual psychology do not "type-identify twins that differ in their representational capacities."[70] To see this, consider the case, described by Geoffrey Hinton,[71] of a trained-up neural network whose high-level internal states are "clamped," that is, forced by the experimenter into some specific configuration. Activity then flows downwards in a generative cascade, resulting in a state of experimenter-induced hallucination. But what is the content of that state? What is represented, Hinton argues, is best captured by asking *how the world would have to be were such a cascade to constitute veridical perception.* A perceptual state, as here depicted, is thus nothing but "the state of a hypothetical world in which a high-level internal representation would constitute veridical perception."[72] Importantly, the only way for the theorist to populate such a world is by invoking the very world from which the training samples were originally drawn. It is this world, and not the world of any neurologically identical twin, that thus provides the resources that enable us to raise and then answer the question Hinton poses.

These considerations suggest a twist upon the notion of perception as "controlled hallucination." For it would be much better, I suggest, to describe hallucination as a kind of uncontrolled (hence mock) perception. In hallucination, all the machinery of perception is brought to bear but either without the guidance of sensory

[69] For some resources, see Tim Crane, "What Is the Problem of Perception?," *Synthesis Philosophica*, xx, 2 (December 2005): 237–64.

[70] Rescorla, "Bayesian Perceptual Psychology."

[71] Hinton, "What Kind of a Graphical Model Is the Brain?," in *IJCAI-05: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (San Francisco: Morgan Kaufmann, 2005), pp. 1765–75.

[72] *Ibid.*, p. 1774.

prediction error at all or with malfunctioning prediction-error circuitry.[73] In such cases the agent really does enter a state of what David Smith calls "mock sensory awareness."[74] Smith claims, however, that this is an elusive and inadequate notion, since an agent may attend to various aspects of a hallucinated scene, a feat that (he claims) strongly suggests some kind of virtual (he calls it an "intentional") object of awareness. But this, we can now see, is not strictly required.[75] Instead of positing a mock (virtual) object for the mock sensing, we can cash out the content (and any attention-modulated shifts of content) using Hinton's strategy, namely, by asking how the world (this very world) would have to be for that flow of systemic states to constitute veridical perception. The mere possibility of attention-based shifts does nothing to undermine this. That such mock sensory states might evolve in ways consistent with (but not here caused by) shifts of attention to some real-world object is unsurprising, since the malfunctioning internal states are indeed the states of generative models. Such models embody rich sets of expectations concerning how visual input should vary with (for example) movements of eyes and body, and even with covert shifts of attention. Given such resources, mock sensory awareness (nonveridical sensory awareness without a mock object) is no more surprising than mock tango-ing. An experienced dancer knows and can reproduce the moves, even those apparently responsive to a partner, without requiring some surrogate source of push, pull, and resistance. An experienced perceiver, likewise, can enter into sequences of mental states that would be veridical perceptions were the world to contain such-and-such objects, and were those objects subject to such-and-such actions (including the act of shifting attention).

VI. CONCLUSIONS

I have presented an account of perception as generative-model-based prediction and shown that such an account satisfies reasonable constraints upon perception (distinguishing it from mere sensor-based response). Systems that operate in this manner realize powerful forms of hierarchical Bayesian inference and are able to learn their own priors from the data. Moreover, they simultaneously learn

[73] For a detailed account of how this might occur, see Paul C. Fletcher and Chris D. Frith, "Perceiving Is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia," *Nature Reviews: Neuroscience*, x, 1 (January 2009): 48–58.

[74] A. David Smith, *The Problem of Perception* (Cambridge: Harvard, 2002), p. 224.

[75] For some critical discussion of this aspect of Smith's view, see Susanna Siegel, "Direct Realism and Perceptual Consciousness," *Philosophy and Phenomenological Research*, LXXIII, 2 (September 2006): 378–410.

at multiple levels of abstraction, enabling them to induce abstract domain-specific (and perhaps even domain-general) knowledge in advance of "filling in" the details. Detailed learning then proceeds just as if it had been constrained by apt bodies of innate knowledge.

Nothing in the account I have presented rules out the presence of rich bodies of innate knowledge. But it demonstrates that the potent, accelerated, domain-specific learning profiles often associated with such knowledge may also be displayed by systems that begin from much more minimal bases. The precise balance between innate and learnt expectations remains a matter for empirical research. But the HBM accounts on offer share the singular virtue of accommodating many empiricist intuitions (for example, those concerning flexibility in the face of new environmental inputs) while leaving room for as much innate knowledge as well-controlled experimental studies may (or may not) eventually mandate. Such knowledge need not, and in all probability will not, take the form of encoded propositions or rules. Instead, it is likely to consist in a set of probabilistically couched expectations governing the general shape of some of the basic hypothesis spaces that we explore during early learning. Expectations, whether learnt or innate, concerning the shape of these emerging hypothesis spaces plausibly (as we saw in section III above) explain our abilities to learn rapidly from statistically limited samples.

Prediction-based hierarchical Bayesian regimes learn to construct the sensory signal by combining probabilistic representations of hidden causes operating at many different spatial and temporal scales. Like SLICE*, they must match the incoming sensory signal by constructing that signal from combinations of hidden causes (latent variables). The so-called "transparency" of perception emerges as a natural consequence of such a process when it is conditioned by an embodied agent's lifestyle-specific capacities to act and to choose. We seem to see dogs, cats, chasings, pursuits, captures, and (for that matter) handwritten digits, because these feature among the interacting, nested structures of distal causes that matter for human choice and action. Prediction-driven learning is what allows us to lock on to such distal structures (using knowledge that may, as we saw, remain hidden from the agent). Raw sensor perturbations, on the other hand, *cannot* be perceived. Instead, they form the baseline that needs to be matched from the top down using these inherently world-revealing resources.

How, assuming these accounts are on track, should we characterize the relation between perceiver and world? Perception was here revealed as an active process involving the (sub-personal) prediction

of our own evolving neural states. Such a thoroughly inward-looking process of self-prediction may seem unpromising as a model of how perception reaches out to the world. On the contrary, however, the pressure actively to account, across multiple time scales, for our changing inner states itself brings into focus the structured and intelligible world that we then encounter as an arena for action and choice. Somewhat paradoxically, it may thus be processes of inward-looking prediction that enable us (both in learning and in online response) to be perceptually open to an external world. The perceiver-world relationship that results is perhaps most reasonably glossed as one of "not-indirect perception," for these regimes provide a mechanistic account of how brains like ours allow agents like us to encounter our world. More importantly, they show how this can be possible despite the continual presence of ambiguity, uncertainty, and noise. This is perceptual openness for real agents, confronting real worlds.

ANDY CLARK

University of Edinburgh