Principles for models of neural information processing

Kendrick N. Kay*[1], Kevin S. Weiner[2]

[1]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Twin Cities, Minneapolis, MN, USA

[2]Department of Psychology, Stanford University, Stanford, CA, USA

*Corresponding author (kay@umn.edu)

Abstract

The goal of cognitive neuroscience is to understand how mental operations are performed by the brain. Given the complexity of the brain, this is a challenging endeavor that requires the development of formal models. Here, we provide a perspective on models of neural information processing in cognitive neuroscience. We define what these models are, explain why they are useful, and specify criteria for evaluating models. We also highlight the difference between functional and mechanistic models, and call attention to the value that neuroanatomy has for understanding brain function. Based on the principles we propose, we proceed to evaluate the merit of recently touted deep neural network models. We contend that these models are promising, but substantial work is necessary to (i) clarify what type of explanation these models provide, (ii) determine what specific effects they accurately explain, and (iii) improve our understanding of how they work.

Introduction

There has been a recent surge of excitement in deep neural networks for neuroscience (Kriegeskorte, 2015; Yamins and DiCarlo, 2016). Major advances in training deep neural networks were achieved by the artificial intelligence and computer vision communities, and these networks now achieve unprecedented performance levels on certain computer vision tasks such as visual object recognition (Krizhevsky et al., 2012). Following these developments, neuroscientists studying the visual system have shown that responses of units in deep neural networks correlate strongly with experimentally measured responses in the primate visual system (e.g., (Agrawal et al., 2014; Cadieu et al., 2014; Eickenberg et al., 2016; Güçlü and van Gerven, 2015a; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2016; Yamins et al., 2014)). Due to these correspondences and similarities in architecture between the artificial and biological networks, deep neural networks have been touted as excellent models of biological neural systems.

In this paper, we use the excitement elicited by deep neural networks as an opportunity to think carefully and critically about models of brain function. We step back and consider the broad endeavor of developing models in cognitive neuroscience (Sections 1–2) and provide an assessment of *why* we should develop such models (Sections 3–4). We then highlight the important distinction between functional and mechanistic models (Section 5) and propose specific criteria for evaluating models (Section 6). We end by using the principles we propose to

evaluate the merit of deep neural network models (Section 7). While we write this paper as a *Comments and Controversies* article, some of the points in this paper may be uncontroversial, especially to practitioners of model-based neuroscience. However, we think integrating points into a single document is useful, and we believe this consolidated perspective will be especially useful for those who are interested in understanding modeling, but who do not necessarily engage in it.

1. What is cognitive neuroscience?

Before reasoning about models in cognitive neuroscience, we must first define these terms. Gazzaniga, Ivry, and Magnun define 'cognitive neuroscience' as

> "the question of understanding how the functions of the physical brain can yield the thoughts and ideas of an intangible mind" (Gazzaniga et al., 2014).

It is widely accepted that "thoughts and ideas of an intangible mind," or mental operations more generally, can be viewed as information-processing operations: for example, the brain represents sensory information, stores sensory information, reasons about this information, and uses information to guide motor behavior. Thus, the brain can be characterized as an organ that mediates interactions between an organism and its environment, accepting incoming sensory information and delivering outgoing motor information. With this definition in mind, the broad question in cognitive neuroscience is, *how do the structure (anatomy) and function (physiology) of neurons and their connections enable the brain to carry out information-processing operations?*

At a coarse level, we already know *what* the brain does, that is, what the information-processing operations are. To use an example from visual neuroscience (DiCarlo and Cox, 2007), we know that one information-processing operation performed by the brain is to take complex spatiotemporal patterns of light impinging on the retina and to use this information to decide what the source of these inputs are (e.g. what type of object is present in the environment). Or, to use an example from social neuroscience (Kubota et al., 2012), we know that one information-processing operation performed by the human brain is to form stereotypes about other humans and use these stereotypes to influence future behavior. But without further work, we do not know *how* the brain performs these operations. The beauty and challenge of neuroscience is that we have tools for measuring in detail the inner workings of the physical organ that performs these operations. Thus, we can move from 'what does the brain do?' to 'how does this particular neuron, or population of neurons, contribute to the functions performed by the brain?'.

2. What is a model?

A small but growing number of researchers are using model-based approaches to tackle questions in cognitive neuroscience (e.g., (Brouwer and Heeger, 2013; Forstmann et al., 2011; Huth et al., 2012; Kay and Yeatman, 2017; O'Doherty et al., 2007; Santoro et al., 2014; Sprague

2

1 and Serences, 2013)). We propose a simple, general definition of 'model': a model is a
2 description of a system. In neuroscience, a model would describe how the nervous system is
3 physically structured and/or how its activity changes dynamically over time. In the specific field
4 of cognitive neuroscience, a model would describe how the components of the brain accomplish
5 behaviorally relevant information-processing tasks, such as decision-making or motor control.
6 For example, we might ask, for a given brain region, what stimulus, cognitive, or motor
7 operations are performed by neural activity in that region?
8
9 Given the broadness of our proposed definition, nearly any neuroscience result could be viewed
10 as providing a model. However, models vary substantially in how precise and quantitative they
11 are. For example, models can be qualitative, conceptual, and vague about assumptions (e.g., a
12 description in an introductory textbook, or a 'word' model that involves ill-defined jargon), or
13 models can be quantitative, mathematical, and explicit about assumptions (e.g., a formal
14 implementation of a model in computer code). Models can depend on concepts and labels
15 derived from our own cognitive abilities (e.g., oracle models that involve manually labeling
16 complex audiovisual stimuli (Huth et al., 2012)), or models can provide explicit specification of
17 concepts and labels independent of a human observer (e.g., a computational implementation of
18 stimulus category (Kay and Yeatman, 2017)). Models can describe systems at coarse levels of
19 detail (e.g., overall activity in a brain region) or at fine levels of detail (e.g., ion channels). As
20 cognitive neuroscientists, we all attempt to describe how the brain performs information
21 processing, and so technically we are all 'modelers'. Of course, in practice, when we use the
22 term 'model', we are typically referring to descriptions that have been made precise and
23 quantitative, and we adopt this usage for the rest of this paper.
24

25 ## 3. Models make falsifiable claims
26
27 Models perform real scientific work, and are not simply *ad hoc* appendages to an experimental
28 study (though in some cases they can be). Rather, models make substantive falsifiable claims
29 and can progressively improve in sophistication and detail. Consider the following simple
30 experiment (**Figure 1, left**). We ask a human observer to direct her eyes towards a small dot at
31 the center of a blank display. The small dot changes color periodically and we instruct the
32 observer to press a button when the color changes. Meanwhile, we place a stimulus (e.g. a
33 checkerboard) on the display, and move this stimulus to a variety of different positions. As we
34 manipulate the stimulus, we record neural activity in the observer's occipital cortex using some
35 technology (e.g. fMRI). We discover that there is an increase in activity when the stimulus is
36 present on the display and that there is some variation in activity levels as a function of stimulus
37 position.
38
39 *** *Insert Figure 1 here* ***
40
41 In this example, the system consists of the stimulus, task, observer, behavior, measurement
42 device, and recorded activity. Our goal, as scientists, is to describe this system and, in
43 particular, to describe why the increases in neural activity occur. There are many possible

1   descriptions, or models, that we could propose (**Figure 1, right**). For example, let us consider
2   four potential models:
3

4   *Model 1.* There is visual information on the display (it is not blank). That is why occipital
5   cortex shows increased neural activity.
6   *Model 2.* There is a point-to-point mapping between positions on the display and positions
7   on cortex (Engel et al., 1997). That is why neural activity at a given cortical position
8   increases for some stimulus positions but not others.
9   *Model 3.* Spatial extent is one property of a visual stimulus. For a given cortical position, this
10   property is represented through a mathematical operation that takes the spatial
11   extent of the stimulus and performs a weighted sum using a Gaussian function to
12   generate the activity level (Dumoulin and Wandell, 2008). Thus, neural activity
13   levels are what they are because occipital cortex performs this operation.
14   *Model 4.* Light reflected from the display enters the eye, is refracted by the lens, and is
15   focused onto the retina. Photoreceptors in the retina transduce light energy into
16   electrical voltages. These voltages are communicated by different types of cells to
17   retinal ganglion cells, which send action potentials to the LGN. In turn, neurons in
18   the LGN send action potentials to primary visual cortex. At each stage in this
19   process, sensitivity is local (e.g., photoreceptors are sensitive to light from a
20   restricted region of the visual field, neurons in the LGN receive input from a specific
21   collection of neighboring retinal ganglion cells, etc.). The net result of these
22   processing stages can be summarized by any of the earlier three models.
23

24   Although the above models vary widely in sophistication and detail (and we could go into even
25   further detail with respect to molecular mechanisms), all of the models describe the system
26   under consideration and make substantive falsifiable claims. Each model posits certain
27   variables as being causally related to the observed neural activity and implicitly excludes other
28   variables. The claim is that the visual stimulus matters to the neural activity, but that for
29   example, the auditory background noise that happened to be present during the experiment, the
30   motor behavior, and the internal cognitive state of the observer do not. With additional
31   experimental measurements, we can test whether the models are indeed sufficient or whether
32   modifications to the models are necessary. If we find that variables such as auditory stimulation
33   or cognitive state affect the observed activity, these variables must be included to achieve a
34   complete description of the system.
35

36   The examples provided above, like many studies in cognitive neuroscience, characterize neural
37   activity in specific brain regions. This approach assumes we have already accurately identified
38   the relevant brain regions in a given observer, which can be a challenging endeavor (Benson et
39   al., 2012; Frost and Goebel, 2012; Glasser et al., 2016; Gordon et al., 2016; Sabuncu et al.,
40   2010; D. Wang et al., 2015; L. Wang et al., 2014; Weiner and Grill-Spector, 2012). We stress
41   the importance of careful localization of brain regions before developing models of their function.
42   An increasing number of researchers are developing quantitative models of where distinct
43   regions and networks are located within the brain (Haxby et al., 2011; Huth et al., 2016; Nelson
44   et al., 2010; Yeo et al., 2011). Interestingly, locations of regions and networks in cortex do not

1 appear to be random and are instead very predictable. Recent research indicates this
2 predictability may stem from several types of neurobiological substrates. For example, cortical
3 folding (Benson et al., 2012; Weiner et al., 2014), white matter (Saygin et al., 2011; Yeatman et
4 al., 2014), cytoarchitectonics (Rosenke et al., 2017; Weiner et al., 2016a), and myelination
5 (Glasser et al., 2016) can all contribute to predicting the locations of functional regions.
6
7 **4. Why are models useful?**
8
9 Developing precise and quantitative descriptions of how the brain performs information
10 processing takes effort. In our view, models provide three main benefits: summary, explanation,
11 and prediction. We describe these benefits below, and refer the reader to a concrete example
12 taken from previous work (**Figure 2**).
13
14 *Summary*
15
16 Neural measurements are complex and noisy, and there is no limit to the number of
17 experimental variations that one can investigate. Models can provide compact summaries of the
18 information processing that a neural system is performing. Thus, a major benefit of a model is
19 that one can make inferences on a focused set of parameters that summarize the data, instead
20 of attempting to interpret a large number of noisy individual data points. Parameters derived in
21 this way can then be compared across brain areas (e.g. (Kay et al., 2013b)) or subject
22 populations (e.g. (Schwarzkopf et al., 2014)).
23
24 *Explanation*
25
26 Models posit that specific variables relate to neural activity. As such, models provide
27 *explanations* of measurements of the brain. For example, suppose we find that a neuron is
28 highly active when a clip of rock music is played but is only weakly active when a speech clip is
29 played. Why does this occur? One model could be that the neuron computes overall sound
30 intensity, and the reason we observe weak activity for the speech clip is that it has low sound
31 intensity. Alternatively, there are other candidate models that might explain the phenomenon
32 (e.g., selectivity for guitar tones, selectivity for speech, variations in attentional engagement).
33 With appropriate experimental measurements, we can adjudicate different models and decide
34 which model is most accurate (Naselaris and Kay, 2015).
35
36 *Prediction*
37
38 There are several different senses in which models provide predictive power. One sense comes
39 from cross-validation (Hastie et al., 2001), a procedure that is commonly used in model-based
40 studies. In cross-validation, the researcher sets aside some testing data, fits the parameters of a
41 model on the remaining training data, and then assesses how well the fitted model predicts the
42 testing data. The testing data could reflect distinct trials of the same experimental conditions
43 found in the training data, in which case this demonstrates limited predictive power.

1   Alternatively, the testing data could reflect completely novel experimental conditions, which
2   demonstrates stronger predictive power.
3
4   A different sense in which models provide predictive power is if a model developed in one study
5   is able to predict the results of a new study that does not involve exactly the same subjects,
6   stimuli, and task design used in the first study. For example, we have shown that a model
7   developed using simple artificial stimuli and fMRI measurements successfully generalizes to
8   complex naturalistic stimuli (Kay et al., 2013a) and data obtained from a different measurement
9   technique (Winawer et al., 2013). As another example, we have shown that a model that
10 describes structural-functional relationships in one group of subjects can successfully generalize
11 to a new group of subjects (Rosenke et al., 2017; Weiner et al., 2016a).
12
13 A third and deep sense in which models provide predictive power is if a model is able to predict
14 the consequences of physical perturbations to the brain. If we had accurate and detailed
15 descriptions of how neural systems in the brain coordinate to perform information-processing
16 operations, we should be able—in principle—to predict, for example, the effects of lesions made
17 in specific brain areas (Gallant et al., 2000), surgical removal of entire brain areas (Weiner et al.,
18 2016b), the effects of enhancement (Salzman et al., 1990) or disruption (Pascual-Leone and
19 Walsh, 2001) of neural activity, or the effects of psychoactive drugs (Rokem and Silver, 2010).
20 These are not easy predictions to make, assuming we are careful to avoid the illusion of
21 predictive power that comes from making "predictions" after looking at the data. A model
22 conjured to explain effects that have already been observed generates "postdictions" and
23 should be treated with skepticism (since the model has been de facto fit and tailored to the
24 data).
25
26 *** *Insert Figure 2 here* ***
27

5. Functional vs. mechanistic models

29
30 It is important to distinguish between *functional models* and *mechanistic models* of neural
31 information processing (Albrecht et al., 2002; Carandini, 2012; Carandini and Heeger, 2011).
32 Functional (or computational) models characterize the transformation between input and output
33 performed by a neuron or population of neurons (Wu et al., 2006), reminiscent of the concept of
34 functions in mathematics or programming. Mechanistic (or biophysical or circuit) models
35 characterize the details of the method or mechanism by which a neuron or population of
36 neurons carry out such a transformation (Priebe, 2016). To illustrate, recall Models 1–3 from the
37 previous example. These models are all stimulus-referred (Heeger et al., 1996; Wandell et al.,
38 2015) in the sense that they specify how the stimulus relates to activity in occipital cortex. Thus,
39 the models can be viewed as functional models that characterize the transformation between
40 input (stimulus) and output (neural activity). In contrast, Model 4 concerns not only the stimulus,
41 but also the series of physical events that intervene between the stimulus and neural activity.
42 This model can therefore be viewed as a mechanistic model that characterizes how the brain
43 carries out the transformation described by Models 1–3. There may be multiple possible

1   mechanistic models that are all consistent with a given functional model. Functional models can
2   be rigorously established for a system, even if the underlying mechanisms are not yet known.
3
4   Functional and mechanistic models are complementary to one another and should be judged on
5   their own merits. The value of functional models is that they emphasize the outcomes and
6   meaning of neural information processing. The significance of a signal carried by a neuron or
7   population of neurons ultimately lies in what that signal conveys about sensory or motor
8   information for the observer. For example, if an organism encounters a predator, what matters is
9   successful detection of the predator so that motor behavior can be appropriately guided; how
10 that detection is accomplished is of secondary importance. Focusing on mechanisms without
11 addressing sensory or motor significance would generate an incomplete picture of neural
12 information processing. On the other hand, the value of mechanistic models is that they are
13 necessary for a complete understanding of a system. To fully describe a functioning brain, we
14 must specify not only what it and its parts compute, but also how they actually compute it. The
15 presumption—at least under conventional neuroscientific thinking—is that the overall
16 information-processing tasks performed by the brain are ultimately the result of neurons and
17 their interactions, which are governed by basic principles of chemistry and physics. However,
18 therein still lies immense complexity, given the sheer number of neurons, cell types, and
19 connections in the brain.
20
21 These points are directly related to David Marr's well-known levels of analysis where distinctions
22 are made among computational, algorithmic, and implementation levels (Marr, 1982). Slightly
23 generalizing our definition of 'mechanistic' to refer to the details of how something is
24 accomplished, we see the algorithmic level serves as a mechanistic model for the computational
25 level and the implementational level serves as a mechanistic model for the algorithmic level. For
26 example, imagine a situation where an organism is attempting to determine the location of a
27 predator from auditory inputs. We can describe the system at a computational level by
28 characterizing the problem that the organism is trying to solve: given auditory inputs, detect the
29 predator and determine the direction of the predator. We can describe the system at an
30 algorithmic level by identifying the specific set of auditory and decision-making algorithms that
31 the brain uses to solve the problem. Or we can describe the system at an implementational level
32 by identifying the specific configurations of neurons and connections that implement those
33 algorithms. Each level provides details as to how the level above is accomplished. Presumably,
34 we should strive to build accurate models at all levels of analysis.
35
36 Many studies in cognitive neuroscience develop functional models and ignore anatomical
37 implementation. For example, a researcher might use fMRI to investigate how patterns of
38 population activity represent a stimulus, irrespective of details of how this activity is spatially
39 organized across cortex. Or, a researcher might use electrophysiology to study how individual
40 neurons respond to experimental conditions, irrespective of details of cell types or the circuit
41 that a neuron participates in. Whether this is a fruitful or fruitless approach is open for debate.
42 Although it is not entirely known how anatomical details—from cortical folding patterns down to
43 the level of neuron density, cell types, and ion channels—constrain and shape functional
44 processing (Van Essen, 1997), we believe that anatomy may hold valuable clues to function. If

7

1 the brain spends so much neurobiological energy organizing structure and function across
2 spatial scales, presumably this orderliness is useful for something. For example, perhaps the
3 specific way that functional properties are clustered in the brain enables faster and more
4 efficient readout of information for a particular task (Grill-Spector and Weiner, 2014). However,
5 we cannot rule out that in some cases, there may be anatomical structures without any
6 particular functional meaning (Horton and Adams, 2005). The path to understanding anatomical
7 implementation will be difficult, as even the heavily studied functional property of orientation
8 selectivity in primary visual cortex has not been resolved in terms of circuit-level mechanisms
9 (Priebe and Ferster, 2008). Nonetheless, a deeper understanding of anatomical implementation
10 is critical, especially if we want to be able to predict the effects of physical perturbations to the
11 brain or make predictions that successfully span vast differences in spatial scale of
12 measurement in neuroscience (Sejnowski et al., 2014).
13
14 **6. What makes a good model?**
15
16 Thus far, we have addressed what models of neural information processing are, why they are
17 useful, and the distinction between functional and mechanistic models. Now suppose in our
18 daily work, we come across a model put forth by a researcher in the field. How should we
19 evaluate the merit of the model? We propose two criteria, accuracy and understanding.
20
21 *Accuracy*
22
23 The first criterion is accuracy, which refers to how well a given model performs in matching the
24 system under investigation (for example, see **Figure 2b**). To assess accuracy, we collect
25 experimental data at some spatial and temporal scale, perform proper preparation and binning
26 of those data, and then quantify whether the predictions of a model match the data. Typically,
27 models have free parameters whose values are not known *a priori* and must be set to obtain
28 quantitative predictions. These parameters are usually set based on experimental data, and in
29 such cases, it is crucial to control for overfitting. This can be done by evaluating predictive
30 performance on left-out data (i.e. cross-validation) or by using techniques that penalize
31 goodness-of-fit based on number of free parameters (e.g. Akaike Information Criterion). It is
32 sometimes disparagingly remarked that a model is 'just fitting the data'—on the contrary,
33 quantitatively matching experimental measurements is exactly what a model ought to do.
34
35 Beyond quantifying predictive power for a set of data, we should consider the range and
36 diversity of the experimental manipulations represented by those data. A model should describe
37 how the brain carries out information processing in a broad range of situations, not just the
38 specific situations used in one or a few particular studies (Felsen and Dan, 2005; Kay et al.,
39 2013b). For example, suppose visual sinusoidal gratings are presented to an observer and a
40 model is proposed in which the activity of a neuron is calculated as a weighted sum of the
41 luminance values of the stimulus (Carandini et al., 2005). This model posits that neural activity
42 reflects a specific visual attribute and, by implication, does not reflect other visual, cognitive, or
43 motor attributes. Evidence for the accuracy of the model would be greatly strengthened if we
44 performed a diverse range of experimental tests—for example, using naturalistic visual scenes

1   to deliver luminance stimulation (David et al., 2004), manipulating the internal cognitive state of
2   the observer (McAdams and Reid, 2005), or allowing visual stimulation to occur simultaneously
3   with motor responses—and still found that the same model (with exactly the same parameters)
4   accurately predicts neural activity. By performing stringent tests of a model, we gain confidence
5   in its accuracy.
6
7   We comment briefly on the topic of phenomenological models. It is possible to have a model
8   that accurately matches a set of data, but performs no actual explanatory work. Such models
9   (which can also be termed 'purely descriptive models') may be useful for comparison purposes,
10   but do not provide neuroscientific insight (Albrecht et al., 2002). For example, suppose we are
11   investigating how neural responses to stimuli change as a function of the cognitive task that a
12   subject is performing (Kay and Yeatman, 2017). We could propose a model that allows each
13   task to induce an additive offset to neural responses, and this model could be fit and evaluated
14   like any other model. However, the model does not make a substantive claim about the specific
15   property of the tasks that is responsible for the additive offsets, and therefore has limited
16   neuroscientific value. (Imagine trying to predict responses for a novel cognitive task—the model
17   would be incapable of doing so because it does not provide any insight into the nature of
18   cognitive tasks.)
19
20   *Understanding*
21
22   The second criterion for the merit of a model is understanding, which refers to how well we, as
23   scientists, grasp the relationship between the components of a given model and the outcomes
24   that the model predicts. Or, in simpler terms, *do we know how the model works?* To illustrate,
25   suppose we observe neural activity is higher in one experimental condition compared to
26   another. A model that describes this system should indicate what property of the first condition
27   leads to increased neural activity. If the model successfully conveys what this property is, we
28   will have *understood* why the effect occurs. In practice, models can be mathematically or
29   algorithmically complex, and it may take effort to determine which specific model component is
30   responsible for a given effect (for an example of how this can be done, see **Figure 2**).
31
32   It is helpful to consider examples where model understanding is poor. Suppose we wish to
33   characterize the relationship between two continuous variables, x and y. One approach is to
34   characterize y as a weighted sum of the outputs of nonlinear basis functions defined on x (e.g.,
35   the weighted sum of a large number of Gaussian functions). Another approach is to simply
36   characterize y as a linear function of x. Now suppose the relationship between x and y is, in fact,
37   linear. Both the complex nonlinear model and the simple linear model are identical in their
38   behavior and equally accurate in matching the data. However, the complex model has less
39   value because it provides less understanding: to understand the model, we would have to
40   expend additional effort analyzing the tuning properties of the basis functions and the weights
41   associated with the basis functions.
42
43   As another example, suppose we have two code implementations of a functional model of
44   neural information processing, one set of code being short, concise, and well-documented, the

1   other set of code being long, convoluted, and undocumented. Both sets of code behave
2   identically in their input-output behavior and achieve the same accuracy in matching
3   experimental data. However, the longer code has less value because it provides less
4   understanding: to figure out what model the code implements, we have to pore through and
5   digest computer code. More generally, this example highlights the importance of clarity in
6   descriptions of models. Clarity should exist at the verbal or conceptual level (scientific prose),
7   the mathematical level (equations), and for models that are algorithmically complex, the
8   computational level (code). An accurate model is useless if not clearly described.
9
10   What are some practical methods for improving understanding of a model? One is to simply
11   observe the model's behavior. Observing how a model behaves across different experimental
12   manipulations is useful, even if empirical measurements of those manipulations are not
13   available. For example, a functional model of visual processing could be probed using a variety
14   of different stimulus manipulations, such as changing the orientation of a bar, changing the
15   semantic category of a object, etc. Carefully controlled experimental manipulations help isolate
16   and identify what effects are explained by a given model (Rust and Movshon, 2005). A second
17   method for improving understanding is to manipulate the model and examine the effect on the
18   model's behavior (Kay et al., 2013b; Nishimoto and Gallant, 2011). If we remove a certain
19   model component or change a certain model parameter, does the model fail to account for the
20   effect of interest? If the model fails, we have learned that the identified component or parameter
21   is critical (for an example, see **Figure 2**). If the model still works, we have learned that the
22   identified component or parameter is not critical, and we could remove it to obtain a simpler and
23   easier-to-understand model. A third method is to model the model, that is, perform simulations
24   of the model's behavior and attempt to develop a simpler model that accounts for the observed
25   behavior. For instance, in the previously described example involving variables x and y, we
26   could take the complex nonlinear model, perform simulations, and eventually realize that a
27   simple linear model can reproduce the model's behavior.
28

29   7. The case of deep neural networks
30
31   Now that we have covered principles for assessing models of neural information processing, we
32   turn to the specific case of deep neural networks (DNNs). These networks, inspired by
33   properties of biological visual systems (Fukushima, 1980; Serre et al., 2007), consist of multiple
34   layers of processing, where each layer is composed of units that perform relatively simple linear
35   and nonlinear operations on the outputs of previous layers. Connections between units are
36   typically designed such that a convolution is performed in the linear weighting step (same
37   weights are applied at different positions), which parallels the visual system. Parameters of the
38   networks are typically set using supervised learning techniques, optimizing performance on
39   specific tasks such as predicting the object category associated with the visual input (Yamins et
40   al., 2014). Researchers have demonstrated high levels of correlation between activity exhibited
41   by DNN units and measurements of activity in visual cortex in response to naturalistic objects
42   and scenes (Eickenberg et al., 2016; Güçlü and van Gerven, 2015b; Khaligh-Razavi and
43   Kriegeskorte, 2014; Yamins et al., 2014).
44

1   Do DNNs have merit as models of biological visual systems? The answer depends on the
2   specific claim that is being made. Suppose the claim is simply that activity in visual cortex
3   reflects a series of processing operations that are performed on visual inputs provided to an
4   observer. This minimal interpretation, that 'visual cortex is a multi-layer neural network that
5   processes visual inputs', is a simplistic qualitative model that is neither exciting nor
6   objectionable, but nevertheless counts as a valid model (see Section 2). Presumably, there is a
7   deeper, more substantive claim that one wants to make regarding DNNs, and the merit of this
8   claim will depend heavily on details of the architecture and parameters used in a DNN. Do we
9   wish to adopt the extreme claim that every parameter value in a DNN is critical and every DNN
10 unit corresponds to a specific neuron or neural population in the brain? If not, what is the
11 proposed interpretation?
12
13 An important distinction that affects the interpretation of DNNs is whether they are intended as
14 functional or mechanistic models (see Section 5). Suppose DNNs are intended only as
15 functional models of how stimuli (inputs) relate to neural responses (outputs). In this case, there
16 are a number of open questions to address regarding the accuracy of DNNs. Thus far,
17 researchers have examined large-scale datasets involving a diversity of complex naturalistic
18 stimuli and demonstrated general correspondence between artificial and biological responses.
19 However, much work in visual neuroscience has characterized in detail how specific visual
20 areas represent specific stimulus dimensions, such as contrast (Albrecht et al., 2002), spatial
21 extent (Kay et al., 2015), curvature (Brincat and Connor, 2004), color (Horwitz and Hass, 2012),
22 and spatial frequency (Lennie and Movshon, 2005), just to name a few. Do DNNs accurately
23 account for these effects? Furthermore, we should scrutinize the range and diversity of
24 experimental manipulations that have been examined (see Section 6). DNNs provide potential
25 explanations of stimulus-driven activity, but these are ultimately incomplete descriptions given
26 that visual activity is affected by non-stimulus factors, such as attention (Luck et al., 1997),
27 imagery (O'Craven and Kanwisher, 2000), and working memory (Harrison and Tong, 2009).
28
29 Suppose instead that DNNs are intended as mechanistic models that not only characterize
30 stimulus-response transformations, but also the way in which the brain accomplishes those
31 transformations (see Section 5). If this is the intention, we again are faced with a number of
32 open questions. What is the proposed mapping between individual units in a layer of a DNN and
33 the neurons in a given brain area? Are DNNs attempting to account for variations in the physical
34 sizes of different visual areas (Dougherty et al., 2003)? Do layer-to-layer connections in a DNN
35 accurately reflect physical connections in biological visual systems, e.g., the spatial extent of V1
36 neurons that project to a V2 neuron (Sincich et al., 2003)? How can we reconcile DNNs with the
37 existence of bypass routes in corticocortical connections (Felleman and Van Essen, 1991)
38 which violate a strictly hierarchical organization? Can DNNs account for different cell types, the
39 laminar organization of cortex, and the existence of extensive feedback projections?
40
41 In addition to accuracy, we should also evaluate our understanding of DNNs. The computational
42 capabilities of DNNs depend critically on the specific parameters used in the models (Coates et
43 al., 2011; Pinto et al., 2009). However, DNNs have many thousands (or even millions) of free
44 parameters, and so understanding DNNs is not an easy task. If we fail to understand DNNs

1 (either because they are too complex or because we do not want to) and treat these models as
2 'black boxes', they perform the function of prediction but fail to perform the functions of
3 summary and explanation (see Section 4). They do not summarize well because there are too
4 many potential parameters that contribute to the description of the system; they do not explain
5 well because it is not clear which specific parts of the models are necessary to explain a given
6 effect.
7
8 Overall, we are not claiming that DNNs have no utility, but we are highlighting open questions
9 and limitations that apply to DNNs (as well as other potential models of neural information
10 processing). There is substantial work to be done towards clarifying what type of explanation
11 these models are supposed to provide and determining what specific experimental effects they
12 accurately explain. We need also to improve our understanding of how these models work.
13 There are concrete steps we can take towards improving understanding (as discussed in
14 Section 6). We can observe the models (e.g., inspect responses to controlled stimuli
15 (Eickenberg et al., 2016)), we can manipulate the models (e.g., perturb parameters and
16 examine the consequences (Cichy et al., 2016)), and we can model the models (e.g., perform
17 simulated experiments 'in silico' and derive simpler models that achieve the same behavior).
18 The concerns we voice here have guided our own research approach which produces models
19 that explain specific effects and have components that are well understood (e.g., (Kay et al.,
20 2013b; Kay and Yeatman, 2017)).
21

Conclusion

23
24 We wrote this perspective at a broad level to remove us from the messy, often confusing, details
25 of different measurement methods (e.g., fMRI, EEG/MEG, electrophysiology), different data
26 analysis approaches (e.g., multivariate pattern analysis, representational similarity analysis,
27 voxelwise modeling, functional connectivity), and jargon (e.g., encoding, decoding). Although
28 technical details matter (Naselaris et al., 2011), the goal of this paper is to emphasize the larger
29 point that we should be using measurements of the brain to build models of how neurons and
30 neural populations perform complex information-processing operations. These models should
31 accurately predict what happens under a broad range of experimental manipulations, and we
32 should understand these models through clear description, observation, and manipulation.
33 When we encounter a model in the literature, we should carefully consider questions such as:
34 How well do they account for the data? How extensive are the experimental manipulations?
35 How clear is the link between the components of the proposed model and the observed effects?
36 Is the model attempting to provide a functional or mechanistic explanation?
37
38 It is useful to draw inspiration from other domains of science. In chemistry, we know if we mix a
39 certain amount of one chemical with another, we will observe certain outcomes, such as
40 emission of heat. This is because we have working models of the relevant variables (e.g.,
41 molecular composition of the chemicals) and how these variables interact. In astronomy, we
42 know if we observe two celestial bodies headed towards each other, we will observe certain
43 outcomes, such as collision or trajectory deviation. This is because we have working models of
44 the relevant variables (e.g., mass, velocity, presence of other nearby bodies) and how these

1 variables interact. In cognitive neuroscience, suppose we developed models that could predict
2 the behavioral and neural outcomes of an arbitrary experiment involving stimuli and task
3 instructions. Such models would predict how fast and accurate an observer will be at the task,
4 what levels of neural activity will be found in different brain areas, and how these neural activity
5 levels relate to the sensory, cognitive, and motor processes involved. Once we achieve such
6 models, we might be able to claim to know how the brain works.
7

Acknowledgments

9

12
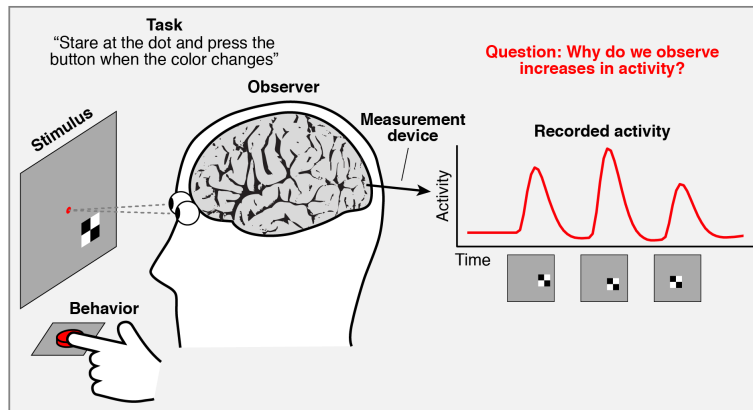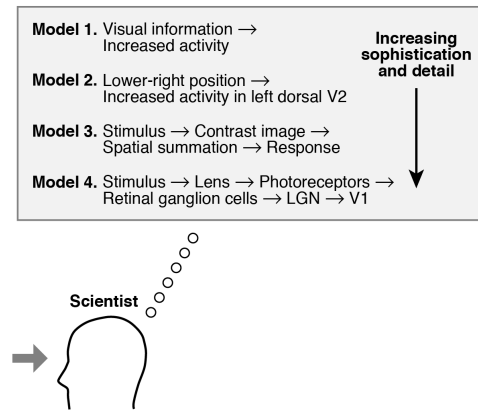
Figure 1. Models describe systems at various levels of sophistication and detail. A typical cognitive neuroscience experiment consists of a stimulus, task, observer, behavior, measurement device, and recorded activity (*left*). A scientist attempts to develop a model of the system, that is, a description of the events that are occurring in the system (*right*). Of particular interest is to characterize why specific levels of neural activity are observed. A variety of different models can be proposed, ranging in sophistication and detail.
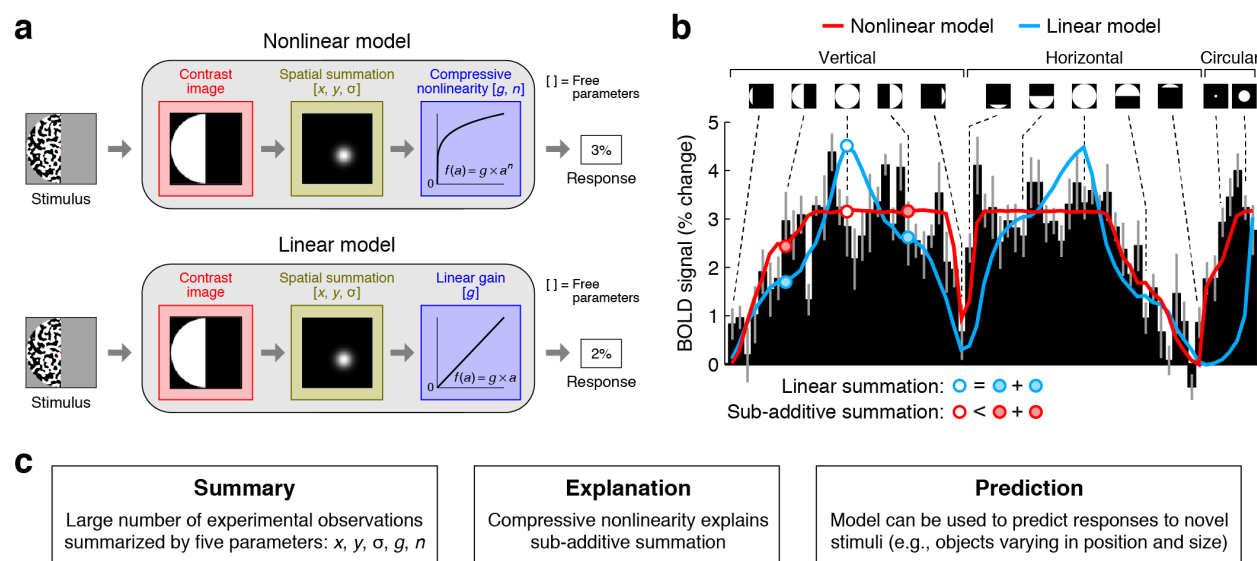
Figure 2. A concrete example of how models provide summary, explanation, and prediction. Figure adapted from (Kay et al., 2013a). *a*, Two potential models of how spatial extent of visual stimuli relates to neural responses. The nonlinear model starts with a contrast image representing stimulus location, computes a weighted sum of this contrast image using a 2D Gaussian, and applies a compressive nonlinearity. The linear model is identical except that the compressive nonlinearity is removed, leaving a linear gain. *b*, Data and model predictions for a voxel in visual area TO-1. Black bars indicate measured BOLD responses to different stimulus locations (depicted by small icons). Leave-one-stimulus-out cross-validation was used to fit the models, and thick lines indicate model predictions. An effect of interest is whether the response to a full stimulus (open dots) is equal to the sum of the responses to two partial stimuli (filled dots). The data support sub-additive summation, which is captured by the nonlinear model. *c*, Three functions performed in this modeling example. (1) The nonlinear model *summarizes* the large set of noisy measurements using just five parameters. (2) The removal of the compressive nonlinearity leads to linear summation, which does not match the data; thus, the compressive nonlinearity is necessary for, and *explains*, sub-additive summation. (3) The nonlinear model *predicts* responses to novel stimuli. For example, the nonlinear model predicts specific levels of tolerance in responses to objects varying in position and size, and experimental measurements have confirmed this prediction (Kay et al., 2013a).

15

References

Agrawal, P., Stansbury, D., Malik, J., Gallant, J.L., 2014. Pixels to voxels: modeling visual representation in the human brain. arXiv preprint arXiv:1407.5104.

Albrecht, D.G., Geisler, W.S., Frazor, R.A., Crane, A.M., 2002. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. Journal of neurophysiology 88, 888–913.

Benson, N.C., Butt, O.H., Datta, R., Radoeva, P.D., Brainard, D.H., Aguirre, G.K., 2012. The retinotopic organization of striate cortex is well predicted by surface topology. Curr Biol 22, 2081–2085. doi:10.1016/j.cub.2012.09.014

Brincat, S.L., Connor, C.E., 2004. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nature neuroscience 7, 880–886.

Brouwer, G.J., Heeger, D.J., 2013. Categorical clustering of the neural representation of color. J. Neurosci. 33, 15454–15465. doi:10.1523/JNEUROSCI.2472-13.2013

Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS computational biology 10, e1003963. doi:10.1371/journal.pcbi.1003963

Carandini, M., 2012. From circuits to behavior: a bridge too far? Nature neuroscience 15, 507–509. doi:10.1038/nn.3043

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., Rust, N.C., 2005. Do we know what the early visual system does? J. Neurosci. 25, 10577–10597.

Carandini, M., Heeger, D.J., 2011. Normalization as a canonical neural computation. Nature Rev. Neurosci. 13, 51–62. doi:10.1038/nrn3136

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6, 1. doi:10.1038/srep27755

Coates, A., Lee, H., Ng, A.Y., 2011. An analysis of single-layer networks in unsupervised feature learning, in:. Presented at the Journal of Machine Learning Research, pp. 215–223.

David, S.V., Vinje, W.E., Gallant, J.L., 2004. Natural stimulus statistics alter the receptive field structure of V1 neurons. J. Neurosci. 24, 6991–7006.

DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends in cognitive sciences 11, 333–341.

Dougherty, R.F., Koch, V.M., Brewer, A.A., Fischer, B., Modersitzki, J., Wandell, B., 2003. Visual field representations and locations of visual areas V1/2/3 in human visual cortex. Journal of vision 3, 586–598.

Dumoulin, S.O., Wandell, B., 2008. Population receptive field estimates in human visual cortex. NeuroImage 39, 647–660. doi:10.1016/j.neuroimage.2007.09.034

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2016. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage. doi:10.1016/j.neuroimage.2016.10.001

Engel, S.A., Glover, G.H., Wandell, B., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex 7, 181–192.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Felsen, G., Dan, Y., 2005. A natural approach to studying vision. Nature neuroscience 8, 1643–1646.

Forstmann, B.U., Wagenmakers, E.-J., Eichele, T., Brown, S., Serences, J.T., 2011. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract?

Trends in cognitive sciences 15, 272–279. doi:10.1016/j.tics.2011.04.002

Frost, M.A., Goebel, R., 2012. Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59, 1369–1381. doi:10.1016/j.neuroimage.2011.08.035

Fukushima, K., 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics 36, 193–202.

Gallant, J.L., Shoup, R.E., Mazer, J.A., 2000. A human extrastriate area functionally homologous to macaque V4. Neuron 27, 227–235.

Gazzaniga, M.S., Ivry, R.B., Mangun, G.R., 2014. Cognitive Neuroscience: The Biology of the Mind, 4 ed. W. W. Norton & Company, New York.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. doi:10.1038/nature18933

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. Cereb. Cortex 26, 288–303. doi:10.1093/cercor/bhu239

Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nature Rev. Neurosci. 15, 536–548. doi:10.1038/nrn3747

Güçlü, U., van Gerven, M.A.J., 2015a. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci. 35, 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015

Güçlü, U., van Gerven, M.A.J., 2015b. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage 145, 329–336. doi:10.1016/j.neuroimage.2015.12.036

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635. doi:10.1038/nature07832

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The elements of statistical learning: data mining, inference, and prediction, Springer series in statistics. Springer, New York.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. doi:10.1016/j.neuron.2011.08.026

Heeger, D.J., Simoncelli, E.P., Movshon, J.A., 1996. Computational models of cortical visual processing. Proceedings of the National Academy of Sciences of the United States of America 93, 623–627.

Horton, J.C., Adams, D.L., 2005. The cortical column: a structure without a function. Philosophical transactions of the Royal Society of London 360, 837–862. doi:10.1098/rstb.2005.1623

Horwitz, G.D., Hass, C.A., 2012. Nonlinear analysis of macaque V1 color tuning reveals cardinal directions for cortical color processing. Nature neuroscience 15, 913–919. doi:10.1038/nn.3105

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. doi:10.1038/nature17637

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. doi:10.1016/j.neuron.2012.10.014

Kay, K.N., Weiner, K.S., Grill-Spector, K., 2015. Attention reduces spatial uncertainty in human ventral temporal cortex. Curr Biol 25, 595–600. doi:10.1016/j.cub.2014.12.050

Kay, K.N., Winawer, J., Mezer, A., Wandell, B., 2013a. Compressive spatial summation in

1   human visual cortex. Journal of neurophysiology 110, 481–494. doi:10.1152/jn.00105.2013

2   Kay, K.N., Winawer, J., Rokem, A., Mezer, A., Wandell, B., 2013b. A two-stage cascade model
3   of BOLD responses in human visual cortex. PLoS computational biology 9, e1003079.
4   doi:10.1371/journal.pcbi.1003079

5   Kay, K.N., Yeatman, J.D., 2017. Bottom-up and top-down computations in word- and face-
6   selective cortex. Elife 6, e22341. doi:10.7554/eLife.22341

7   Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models
8   may explain IT cortical representation. PLoS computational biology 10, e1003915.
9   doi:10.1371/journal.pcbi.1003915

10   Kriegeskorte, N., 2015. Deep neural networks: A new framework for modeling biological vision
11   and brain information processing. Annual Review of Vision Science 1, 417–446.

12   Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep
13   Convolutional Neural Networks 1097–1105.

14   Kubilius, J., Bracci, S., Op de Beeck, H.P., 2016. Deep Neural Networks as a Computational
15   Model for Human Shape Sensitivity. PLoS computational biology 12, e1004896.
16   doi:10.1371/journal.pcbi.1004896

17   Kubota, J.T., Banaji, M.R., Phelps, E.A., 2012. The neuroscience of race. Nature neuroscience
18   15, 940–948. doi:10.1038/nn.3136

19   Lennie, P., Movshon, J.A., 2005. Coding of color and form in the geniculostriate visual pathway
20   (invited review). Journal of the Optical Society of America 22, 2013–2033.

21   Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R., 1997. Neural mechanisms of spatial
22   selective attention in areas V1, V2, and V4 of macaque visual cortex. Journal of
23   neurophysiology 77, 24–42.

24   Marr, D., 1982. Vision: A computational investigation into the human representation and
25   processing of visual information, henry holt and co. Inc., New York, NY 2, 4.2.

26   McAdams, C.J., Reid, R.C., 2005. Attention modulates the responses of simple cells in monkey
27   primary visual cortex. J. Neurosci. 25, 11023–11033. doi:10.1523/JNEUROSCI.2904-
28   05.2005

29   Naselaris, T., Kay, K.N., 2015. Resolving Ambiguities of MVPA Using Explicit Models of
30   Representation. Trends in cognitive sciences 19, 551–554. doi:10.1016/j.tics.2015.07.005

31   Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI.
32   NeuroImage 56, 400–410. doi:10.1016/j.neuroimage.2010.07.073

33   Nelson, S.M., Cohen, A.L., Power, J.D., Wig, G.S., Miezin, F.M., Wheeler, M.E., Velanova, K.,
34   Donaldson, D.I., Phillips, J.S., Schlaggar, B.L., Petersen, S.E., 2010. A parcellation scheme
35   for human left lateral parietal cortex. Neuron 67, 156–170.
36   doi:10.1016/j.neuron.2010.05.025

37   Nishimoto, S., Gallant, J.L., 2011. A three-dimensional spatiotemporal receptive field model
38   explains responses of area MT neurons to naturalistic movies. J. Neurosci. 31, 14551–
39   14564. doi:10.1523/JNEUROSCI.6801-10.2011

40   O'Craven, K.M., Kanwisher, N., 2000. Mental imagery of faces and places activates
41   corresponding stimulus-specific brain regions. Journal of cognitive neuroscience 12, 1013–
42   1023.

43   O'Doherty, J.P., Hampton, A., Kim, H., 2007. Model-based fMRI and its application to reward
44   learning and decision making. Ann. N. Y. Acad. Sci. 1104, 35–53.
45   doi:10.1196/annals.1390.022

46   Pascual-Leone, A., Walsh, V., 2001. Fast backprojections from the motion to the primary visual
47   area necessary for visual awareness. Science 292, 510–512. doi:10.1126/science.1057099

48   Pinto, N., Doukhan, D., DiCarlo, J.J., Cox, D.D., 2009. A high-throughput screening approach to
49   discovering good forms of biologically inspired visual representation. PLoS computational
50   biology 5, e1000579.

51   Priebe, N.J., 2016. Mechanisms of Orientation Selectivity in the Primary Visual Cortex. Annual

Review of Vision Science 2, 85–107.

Priebe, N.J., Ferster, D., 2008. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. Neuron 57, 482–497.

Rokem, A., Silver, M.A., 2010. Cholinergic enhancement augments magnitude and specificity of visual perceptual learning in healthy humans. Curr Biol 20, 1723–1728. doi:10.1016/j.cub.2010.08.027

Rosenke, M., Weiner, K.S., Barnett, M.A., Zilles, K., Amunts, K., Goebel, R., Grill-Spector, K., 2017. A cross-validated cytoarchitectonic atlas of the human ventral visual stream. NeuroImage. doi:10.1016/j.neuroimage.2017.02.040

Rust, N.C., Movshon, J.A., 2005. In praise of artifice. Nature neuroscience 8, 1647–1650.

Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., Haxby, J.V., 2010. Function-based intersubject alignment of human cortical anatomy. Cereb. Cortex 20, 130–140. doi:10.1093/cercor/bhp085

Salzman, C.D., Britten, K.H., Newsome, W.T., 1990. Cortical microstimulation influences perceptual judgements of motion direction. Nature 346, 174–177. doi:10.1038/346174a0

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS computational biology 10, e1003412. doi:10.1371/journal.pcbi.1003412

Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D.E., Saxe, R.R., 2011. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. Nature neuroscience 15, 321–327. doi:10.1038/nn.3001

Schwarzkopf, D.S., Anderson, E.J., de Haas, B., White, S.J., Rees, G., 2014. Larger extrastriate population receptive fields in autism spectrum disorders. J. Neurosci. 34, 2713–2724. doi:10.1523/JNEUROSCI.4416-13.2014

Sejnowski, T.J., Churchland, P.S., Movshon, J.A., 2014. Putting big data to good use in neuroscience. Nature neuroscience 17, 1440–1441. doi:10.1038/nn.3839

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., Poggio, T., 2007. A quantitative theory of immediate visual recognition. Progress in brain research 165, 33–56.

Sincich, L.C., Adams, D.L., Horton, J.C., 2003. Complete flatmounting of the macaque cerebral cortex. Visual neuroscience 20, 663–686.

Sprague, T.C., Serences, J.T., 2013. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nature neuroscience 16, 1879–1887. doi:10.1038/nn.3574

Van Essen, D.C., 1997. A tension-based theory of morphogenesis and compact wiring in the central nervous system. Nature 385, 313–318. doi:10.1038/385313a0

Wandell, B., Winawer, J., Kay, K.N., 2015. Computational modeling of responses in human visual cortex. Brain Mapping 651–659.

Wang, D., Buckner, R.L., Fox, M.D., Holt, D.J., Holmes, A.J., Stoecklein, S., Langs, G., Pan, R., Qian, T., Li, K., Baker, J.T., Stufflebeam, S.M., Wang, K., Wang, X., Hong, B., Liu, H., 2015. Parcellating cortical functional networks in individuals. Nature neuroscience 18, 1853–1860. doi:10.1038/nn.4164

Wang, L., Mruczek, R.E.B., Arcaro, M.J., Kastner, S., 2014. Probabilistic Maps of Visual Topography in Human Cortex. Cereb. Cortex bhu277. doi:10.1093/cercor/bhu277

Weiner, K.S., Barnett, M.A., Lorenz, S., Caspers, J., Stigliani, A., Amunts, K., Zilles, K., Fischl, B., Grill-Spector, K., 2016a. The Cytoarchitecture of Domain-specific Regions in Human High-level Visual Cortex. Cereb. Cortex 27, 146–161. doi:10.1093/cercor/bhw361

Weiner, K.S., Golarai, G., Caspers, J., Chuapoco, M.R., Mohlberg, H., Zilles, K., Amunts, K., Grill-Spector, K., 2014. The mid-fusiform sulcus: a landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. NeuroImage 84, 453–465. doi:10.1016/j.neuroimage.2013.08.068

1   Weiner, K.S., Grill-Spector, K., 2012. The improbable simplicity of the fusiform face area.
2   Trends in cognitive sciences 16, 251–254. doi:10.1016/j.tics.2012.03.003
3   Weiner, K.S., Jonas, J., Gomez, J., Maillard, L., Brissart, H., Hossu, G., Jacques, C., Loftus, D.,
4   Colnat-Coulbois, S., Stigliani, A., Barnett, M.A., Grill-Spector, K., Rossion, B., 2016b. The
5   Face-Processing Network Is Resilient to Focal Resection of Human Visual Cortex. J.
6   Neurosci. 36, 8425–8440. doi:10.1523/JNEUROSCI.4509-15.2016
7   Winawer, J., Kay, K.N., Foster, B.L., Rauschecker, A.M., Parvizi, J., Wandell, B., 2013.
8   Asynchronous broadband signals are the principal source of the BOLD response in human
9   visual cortex. Curr Biol 23, 1145–1153. doi:10.1016/j.cub.2013.05.001
10 Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory
11   neurons by system identification. Annual review of neuroscience 29, 477–505.
12 Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand
13   sensory cortex. Nature neuroscience 19, 356–365. doi:10.1038/nn.4244
14 Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014.
15   Performance-optimized hierarchical models predict neural responses in higher visual cortex.
16   Proceedings of the National Academy of Sciences of the United States of America 111,
17   8619–8624. doi:10.1073/pnas.1403112111
18 Yeatman, J.D., Wandell, B., Mezer, A.A., 2014. Lifespan maturation and degeneration of human
19   brain white matter. Nat Commun 5, 4932. doi:10.1038/ncomms5932
20 Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M.,
21   Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L.,
22   2011. The organization of the human cerebral cortex estimated by intrinsic functional
23   connectivity. Journal of neurophysiology 106, 1125–1165. doi:10.1152/jn.00338.2011
24