Title: Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks

Cameron Buckner
University of Houston
Department of Philosophy
cjbuckner@uh.edu

This is a pre-print of an article published in *Synthese*.
The final authenticated version will be available online at https://doi.org/10.1007/s11229-018-01949-1

Abstract: In artificial intelligence, recent research has demonstrated the remarkable potential of Deep Convolutional Neural Networks (DCNNs), which seem to exceed state-of-the-art performance in new domains weekly, especially on the sorts of very difficult perceptual discrimination tasks that skeptics thought would remain beyond the reach of artificial intelligence. However, it has proven difficult to explain why DCNNs perform so well. In philosophy of mind, empiricists have long suggested that complex cognition is based on information derived from sensory experience, often appealing to a faculty of abstraction. Rationalists have frequently complained, however, that empiricists never adequately explained how this faculty of abstraction actually works. In this paper, I tie these two questions together, to the mutual benefit of both disciplines. I argue that the architectural features that distinguish DCNNs from earlier neural networks allow them to implement a form of hierarchical processing that I call "transformational abstraction". Transformational abstraction iteratively converts sensory-based representations of category exemplars into new formats that are increasingly tolerant to "nuisance variation" in input. Reflecting upon the way that DCNNs leverage a combination of linear and non-linear processing to efficiently perform this feat allows us to understand how the brain is capable of bi-directional travel between exemplars and abstractions, addressing longstanding problems in empiricist philosophy of mind. I end by considering the prospects for future research on DCNNs, arguing that rather than simply implementing 80s connectionism with more brute-force computation, transformational abstraction counts as a qualitatively distinct form of processing ripe with philosophical and psychological significance, because it is significantly better suited to depict the generic mechanism responsible for this important kind of psychological processing in the brain.
Keywords: Abstraction, connectionism, deep learning, convolution, empiricism, nuisance variation, mechanism

"In particular the concept of what is a chair is hard to characterize simply. There is certainly no AI vision program which can find arbitrary chairs in arbitrary images … Such problems are never posed to AI systems by showing them a photo of the scene. A person (even a young child) can make the right interpretation of the photo and suggest a plan of action. But this abstraction is the essence of intelligence and the hard part of the problems being solved. Under the current scheme the abstraction is done by the researchers leaving little for the AI programs to do but search. A truly intelligent program would study the photograph, perform the abstraction and solve the problem." Rodney Brooks (1991, p143), "Intelligence without Representation"

1. Introduction: The deepening of AI

On March 15, 2016, AlphaGo—a system designed by Google's DeepMind research group—defeated Lee Sedol in the final game of a five-game Go competition (Silver et al., 2016). This victory will be remembered as a remarkable milestone in the development of artificial intelligence, even more so than Deep Blue's victory over chess grandmaster Gary Kasparov in 1997. Lee was the second-highest ranking professional Go player at the time, a 9 dan grandmaster, and Go is, in computational terms, a more difficult problem than chess.

The difficulty can be located in its higher “branching-factor”; though the rules of Go are in some sense simpler (with only one type of game piece, rather than six), there are many more possible moves for a player at each turn (there are about 35 possibilities at each turn for chess, compared to around 250 for Go). Considering all possible combinations of moves through only about a third of an average Go game would thus require exploring more board configurations than there are atoms in the universe. Perhaps for this reason, expert Go players are often unable to explain their strategies in terms of the explicit rule-based knowledge presumed by classical approaches to artificial intelligence; they instead express them in terms of abstractions such as “influence”, “connection”, and “stability”—and so it was thought that highest levels of Go play would remain forever beyond the computer’s more algorithmic grasp. Nevertheless, AlphaGo shattered those predicted limitations because it was the first artificial Go-playing system that could model (and in some ways surpass) our own powers of abstraction.

This interpretation might sound incredible, but it follows from a proper understanding of AlphaGo’s success. Its architecture contains two components that its modelers describe in terms of “reflection” and “intuition”: a “reflective” tree-search procedure populates and explores the space of possible moves, and a trio of “intuitive” Deep Convolutional Neural Networks (hereafter DCNNs) guide that search using information abstracted from the model’s extensive previous experience. To elaborate, the tree search component simulates future game trajectories by predicting consecutive moves and countermoves until one of the two players wins out. Some commentators have emphasized this component of AlphaGo’s architecture in explaining its success (e.g. Lake, Ullman, Tenenbaum, & Gershman, 2016); but this approach to search—referred to as “Monte Carlo Tree Search” in machine learning—is by now a traditional, brute-force AI method previously implemented by a variety of other programs (i.e. CrazyStone, Zen, Pachi) that did not approach AlphaGo’s level of skill. In fact, even when this part of its architecture was disabled and AlphaGo selected moves using its policy networks alone, it still defeated Pachi—one of the next-highest ranking Go-playing programs, which deploys more than 100,000 simulated games each turn—85% of the time.[1] Instead,

[1] This is similar to findings in human expertise; for example, Gobet and Simon (1996) found that the skill level of chess grandmasters was little diminished by putting them under time constraints and pitting them against a dozen opponents, compared to the more leisurely pace of tournament play which allows more consideration of alternative moves.

AlphaGo's success is due to its use of DCNNs (LeCun, Bengio, & Hinton, 2015), which are elsewhere exceeding state-of-the-art performance on new problems weekly, from Go play to image classification, speech recognition, autonomous driving, and video game play, to name just a few.

In this paper, I place these engineering advances in the philosophical context of empiricist philosophy of mind, addressing questions in both disciplines. On the engineering side, most agree that DCNNs work astonishingly well on a wide range of tasks, but it has proven difficult to explain why.[2] I argue here that DCNNs are so successful across so many different domains because they model a distinctive kind of abstraction from experience, to be elaborated below. On the philosophical side, this engineering achievement vindicates some themes from classical empiricism about reasoning, while at the same time filling in some of its most vexing lacunae. It supports the empiricist idea that information abstracted from experience enables even high-level reasoning in strategy games like chess and Go. More surprisingly, it bolsters a Lockean approach to abstraction that has long been dismissed by other empiricists. Indeed, the present results go beyond what some of Locke's critics like Hume dreamt possible; for while Hume took reasoning to be a rarified form of association, he notoriously claimed that the means by which the mind selected the most creative associations could never be made intelligible to human understanding. Not only can we now understand the trajectory of these paths by analyzing DCNNs, these models also help us better understand the mechanism that performs this abstraction in the brain.

Two quick caveats about the aims of the paper before beginning: first, this paper does not aim to defend DCNNs as the right, best, or only viable approach to modeling "general intelligence", so many recent criticisms of deep learning (i.e., Marcus, 2018) are beside the point here. My central claim is rather that DCNNs model one crucially important component of intelligence—a form of categorical abstraction that until recently eluded our grasp—but other components may be required for general intelligence. Second, though explaining the success of DCNNs requires us to analyze their success at categorization, I do not here attempt to solve the more ambitious problem of showing how they might implement a Fregean notion of

[2] This question has been raised as the "interpretation problem"; however, this label has been used too broadly and inconsistently to admit of a single solution. Some commentators use it to broach the question addressed here—why do DCNNs succeed where other neural network architectures struggle—while others use it to raise other questions, such as semantic interpretability or decision justification.

concepts, which requires intersubjective agreement and objective standards of correctness (Machery, 2009). I focus more modestly on subjective category representations or "conceptualizations" (Gauker, 2011, p6).

The structure of the paper is as follows: In Section 2, I review the empiricist approach to learning general categories, introducing two classic problems concerning the role of abstraction. In Section 3, I introduce the notion of transformational abstraction as an intuitive solution to these problems by canvassing a series of nontechnical examples. In Section 4, I provide a primer on DCNNs, focusing especially on architectural differences between DCNNs and the three-layer feedforward networks of the 80s and 90s with which most readers are probably more familiar. In Section 5, I explain how these architectural features allow DCNNs to model the way mammalian cortex implements transformational abstraction. With all these pieces in place, in Section 6 we may finally clarify the proper interpretation of DCNNs—as well as the limits of that interpretation—by arguing that they depict specific aspects of the neural mechanisms that implement transformational abstraction in mammalian neocortex.

2. Historical Interlude: Two Classic Problems with Empiricist Accounts of Abstraction

This paper proposes that it is useful to frame recent research in DCNNs with the work of historical empiricists like Locke and Hume—as the neural network modelers themselves have often done (i.e. Silver et al., 2017). This linkage, however, merits some disclaimers—which, in the interests of efficiency and bipartisanship, I echo from Laurence & Margolis, who have written on similar topics from a nativist perspective (2012, 2015). First, the goal of such a historical linkage is not to enter into debates about the best interpretation of these thinkers on their own terms, but rather to extract problems and insights in a way that illuminates current research developments. Second and relatedly, some of the terms of the great debate will thus be due for renovation; for example, we should agree here with Laurence & Margolis (2015) that if empiricists must hold that the mind begins as a truly blank slate lacking any structure whatsoever, then, for quite some time, there shall have been no serious defenders of pure empiricism. A mind lacking any general resources for perception, memory, learning, attention, or association will learn nothing from any amount of experience. Rather, the contemporary nativist/empiricist debate in cognitive science should be construed as a disagreement as to the origins of categorical representations, specifically as to whether those representations are due mostly to domain-specific or domain-general cognitive mechanisms. Thus, the goal of our historical linkage is to see whether problems concerning the acquisition of general category representations so poignantly raised by historical empiricists can in some useful sense be addressed by DCNNs using primarily domain-general learning mechanisms.

Contemporary rationalists are skeptical that this enterprise will succeed, wondering how domain-general mechanisms could ever be bootstrapped or scaffolded to learn the sorts of category representations that tend to feature in more demonstrative reasoning, especially those found in i.e. geometry, arithmetic, or social cognition. The general empiricist strategy for responding to this challenge often appeals to abstraction: we come to learn general principles about, for example, triangles—that they all have three angles (but not necessarily of any particular degree), that they have three sides (but not necessarily of any particular length or orientation), or even such general laws as the Pythagorean Theorem—by considering triangles "in the abstract". One popular idea is that this involves "leaving out" features which are irrelevant to the figures' triangularity—an approach that Gauker (2011) explicates as "abstraction-as-subtraction"—but explaining how humans and animals can even detect the abstract features in specific particulars has proven difficult.

I will begin with *triangle* as a test case, because of its prominence in relevant historical discussions; but such relatively neat geometric categories are too easily recognized by state-of-the-art DCNNs to illustrate their distinctive power,[3] so we shall quickly move on to messier and more challenging categories like *chair*. Prominent researchers like Rodney Brooks have as recently as 30 years ago bemoaned AI agents' inability to recognize chairs in the sorts of natural scenes that humans encounter in daily perception, describing the ability to do so as a the "essence of intelligence" (1987, p143). More recently, *chair* has featured as a routine test category in standardized photograph labeling challenges in machine learning, such as the Pascal Visual Object Classes challenge 2012, on which DCNNs have held the highest benchmark performance. Thus, if we can adequately explain how DCNNs are able to efficiently recognize *chairs* in these photographs and how their ability to do so relates to human and animal cognition, it will have already been a good day's philosophical work.

[3] Even three-layer perceptrons have been trained to categorize triangle exemplars with a high degree of accuracy (Spasojević, Šušić, & Djurović, 2012).

Let us consider one perplexing illustration of the difficulty in providing an empiricist account of abstraction from the British empiricist John Locke. Locke tried explain the origin and structure of the abstract idea of a triangle in a particularly notorious passage of the *Essay*:

> "The ideas first in the mind, it is evident, are those of particular things, from whence, by slow degrees, the understanding proceeds to some few general ones…For when we nicely reflect upon them, we shall find, that general ideas are fictions and contrivances of the mind, that carry difficulty with them, and do not so easily offer themselves, as we are apt to imagine. For example, does it not require some pains and skill to form the general idea of a triangle (which is yet none of the most abstract, comprehensive, and difficult), for it must be neither oblique, nor rectangle, neither equilateral, equicrural, nor scalenon; but all and none of these at once. In effect, it is something imperfect, that cannot exist; an idea wherein some parts of several different and inconsistent ideas are put together." (Locke 1690, IV.7.9).

Locke here highlights a difficulty with acquiring even moderately abstract categories: they must be learned from and subsume a number of particulars with mutually-inconsistent manifestations of their characteristic features. However, by leaving the reader with the impression that the abstract idea of a triangle itself contains inconsistent features, this "universal triangle" passage became one of the most snarked about in the history of philosophy (Beth, 1957). Berkeley in particular pulled no punches, opining that "if any man has the faculty of framing in his mind such an idea of a triangle…it is in vain to pretend to dispute him out of it," and "is it not a hard thing to imagine that a couple of children cannot prate together of their sugar-plums and rattles…till they have first tacked together numberless inconsistencies?" (Berkeley, 1710, Introduction 13-14). Let us take this passage from Locke to introduce the first problem that I will address below. Summarized, **Problem #1**: Is there a coherent position that could be attributed to this passage, and what is it to possess the abstract idea of a triangle? More specifically, how can we form representations of abstract categories when our sensory experience consists of only particular exemplars exhibiting mutually-inconsistent properties?

Moving on, Berkeley and later empiricists like Hume took the troublesome triangle passage not so much to present an interesting problem to be solved as evidence that Locke's doctrine of abstraction had gone off the rails. In its place, Berkeley and Hume attempted to eschew the need for abstract ideas entirely by offering a more (what we would now call) exemplar-based approach to abstract reasoning—i.e. what Gauker (2011) explicates as "abstraction-as-representation". On one way of developing this view, when we reason about abstract principles—say, the Pythagorean theorem—it is not that we reflect upon some abstract idea of a triangle which possesses no specific properties or inconsistent combinations of specific properties, and use that idea to deduce the theorem; it is rather that we consider a series of particular triangles (with their idiosyncratic sizes, angles, rotations, and so on) and take them to stand in for the whole category. We then evaluate the truth of the general propositions by considering only the relevant aspects of those specific exemplars (i.e. their *triangularity* or more specifically *three-sidedness*, but not *size* or the particular degrees of their angles), and seeing whether the general proposition holds true of them on those grounds.

While this account gets something right about the course of demonstrative reasoning, it creates further problems regarding its generalizability. For one, how can we be sure that we considered only aspects of the sample figures that were relevant to their triangularity? For another, how can we know that the set of exemplars we considered was broad enough to stand in for the whole class? We must already possess some way of relating abstractions to exemplars to ensure that we have only considered their relevant properties; and we need some way to generate representative exemplars for abstract categories to evaluate a comprehensive sample. Berkeley and Hume both thought that language plays an important role here, but even the availability of explicit definitions for the abstract categories would not be panacea. To take an extreme example, for centuries everyone thought that the principles of Euclidean geometry, such as that two parallel lines never meet, were necessarily true; it fell to Riemann to conceive of consistent arrangement of parallel lines that violated them.

Hume, who was aware of the difficulty of selecting appropriate exemplars, put the matter thusly:

> "One might think that we could see the whole intellectual world of ideas all at once, and that all we did was to pick out the ideas that best suited our purpose. But it may be that really the only ideas that we have at such a moment are the seemingly 'picked out' ones—the very ideas that are thus collected by a kind of magical faculty in the soul. This faculty is always most perfect in the greatest geniuses…but it can't be explained by the utmost efforts of human understanding." (Hume, 1739, I.7)

Unfortunately, where most we need transparency, the lynchpin of this revised empiricist account of reasoning is again obscured in a haze of magic. While we should agree with Berkeley and Hume that we cannot form or reason about mental images with inconsistent properties—which is hopefully not what Locke meant, anyway—we might rightly complain that their exemplar-based alternative has only pushed the mystery to another location. This presents us with **Problem #2**: How does the mind produce appropriate exemplars of abstract categories to represent classes in reasoning?

In the next section, we will explore an intuitive solution to Problems #1 and #2, showing them to be two sides of the same coin. An answer to both, we shall see in Section 4, is suggested by a proper understanding of how DCNNs transform sensory-based representations of particulars, on the basis of accrued experience, into more abstract representational formats—and how they also chart a return journey back to particulars again.

3. An Ecumenical Solution: Transformational Abstraction

We have reached an apparent opposition between two different approaches to abstraction, each confronted by a difficult problem. On the one hand, Locke suggests that our general category representations of properties like *triangle* subsume idiosyncratic exemplars by focusing only on their relevant commonalities (i.e. using abstraction-as-subtraction), but it is difficult to explain how we cope with the inconsistent manifestations of these commonalities in doing so. On the other hand, Berkeley & Hume argue that we reason about abstract categories by selecting or generating specific exemplars to stand in for the class (i.e. using abstraction-as-representation)—but the way we select or generate appropriate exemplars remains mysterious. The first approach concerns the mental travel from specific exemplars to abstract categories, with the mystery concerning how the mind knows what to leave out from consideration. And the second approach concerns the trip from abstract categories to specific exemplars, with the difficulty concerning how the mind produces a representative set of exemplars exhibiting appropriate details.

I will here argue that the mind is capable of travel in both direction—vindicating elements of both approaches—but first it will be useful to consider a third approach that has been popular in accounting for the most abstract mathematical or logical properties, such as *cardinal number* or *valid argument*. A common strategy here deploys the notion of transformational invariance. In this sense, an invariant property is one that is perfectly conserved across some systematic alterations of the target domain. For example, we might discover abstract geometric properties by performing spatial transformations on an object or set of objects—such as scalings, translations, rotations, or reflections—and seeing which properties are conserved across all transformations (in fact, this method provides the basis of trigonometry). In arithmetic and logic, invariance under permutation has featured prominently in the views of neo-logicists and structuralists as a way to distinguish arithmetical or logical concepts; for example, the cardinality of a set is invariant across any order in which a set's elements are counted, and a valid argument form remains valid across any permutation of the argument's domain.[4] Call this third family of approaches "abstraction-as-invariance".

Unfortunately, even if this approach can be used to define the most abstract logical or mathematical properties, a search for perfect invariance cannot solve our problem with everyday categories like *chair*. Psychologists like Rosch (1978) and Barsalou (1999) have argued convincingly that membership in such categories is too graded and idiosyncratic to be defined so cleanly; they are, as Wittgenstein emphasized, "family-resemblance" categories that lack a perfectly invariant essence. However, a key proposal of this paper is that if we treat invariance as a more graded and multi-dimensional notion, a domain-general tolerance for specific sources of variance in perceptual input might provide an empiricist answer to Brooks' challenge concerning *chairs*.

Specifically, computer vision and machine learning researchers have recently noted that *triangle*, *chair*, *cat*, and other everyday categories are so difficult to recognize because they can be encountered in a variety of different poses or orientations that are not mutually similar in terms of their low-level perceptual properties. Without a deeper representation of chairs, a chair seen from the front does not look much like the same chair seen from behind or above; we must somehow unify all these diverse perspectives to build a reliably successful chair-detector. The set of variables on which perspectives can vary tend moreover to be the same for a very wide range of common categories; computer vision researchers have come to call these repeat offenders "nuisance variables", as they routinely pose challenges to bottom-up approaches to cognition. Common examples of nuisance parameters are size, position, and angular rotation in visual recognition tasks, or pitch, tone, and duration in auditory recognition tasks. The challenge facing a computer vision modeler is thus to develop an artificial agent that can reshape its sense of perceptual similarity by controlling for common forms of nuisance variation. An agent that is able to do so should be able to judge a cat seen from

[4] This is but the barest gloss on a rich research area in the foundations of logic and math going back to Hilbert—for a recent overview, see Antonelli (2010).

the front as more similar to a cat seen from behind than to a dog seen from the front, despite initial perceptual dissimilarity.

To spell out the transformational solution to this problem, let us return to the mystery that troubled Locke: How do we form the general idea of a triangle, when we have only been exposed to a series of particular and idiosyncratic exemplars with surface properties—sizes, spatial positions, angular rotations and degrees—that are mutually inconsistent? Berkeley says we accomplish this by directing our attention to aspects of the individual triangles such as their number of angles or lines, and to how these diverse figures either satisfy or fail to satisfy the definition of a triangle. But if we consider the scope of the empiricist problem in the way it was more recently posed by Quine, or as confronted by a computer vision model—as beginning with "stimulation of [the] sensory receptors" by "certain patterns of irradiation in assorted frequencies" and terminating in high-level category knowledge (1971, pp. 82-83)—then even an explicit definition of triangle wouldn't help; "line" and "angle" would already too abstract and diverse in their presentations to take for granted as representational primitives.

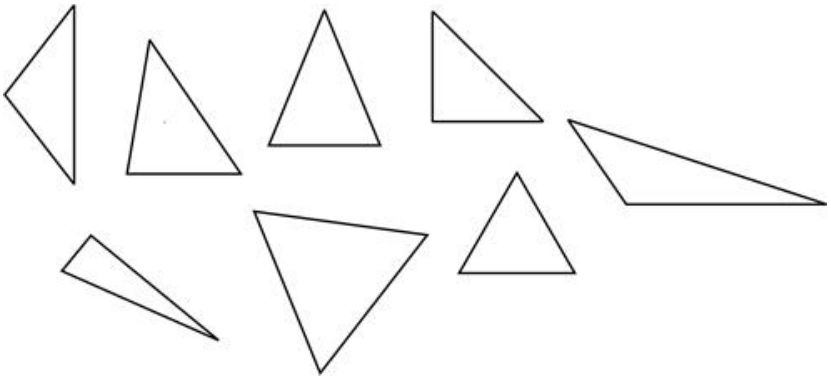

Figure 1. Examples of triangle exemplars with mutually-inconsistent features.

One way to make the problem more tractable would be to view triangles as hierarchically composed all the way down, from pixels and color channels to lines, angles, and shape, and develop a series of filters that could detect and compose each layer of features into those at the next layer of abstraction.[5] Contrasts and shadings can be built from pixels, lines from suitably composed contrasts, and so on. Barsalou notes that

[5] Such hierarchical forms of composition have also been considered a fourth method of abstraction (also attributed to Locke by Gauker (2011) as "abstraction-as-composition")—but as three-layer perceptrons can perform this form of abstraction just as well, I do not emphasize it here.

such a method might work in defense of Locke, as "three qualitative detectors for lines, coupled spatially with three qualitative detectors for vertices that join them, could represent a generic triangle" (1999, p. 585). This method does not solve our Problem #1 without elaboration, however, because those lines could be of diverse lengths, and those angles could be located in diverse locations; a "qualitative" detector of lines and angles would need to be able to cope with this diversity. We need a method that can look past this variation to group together diverse triangle exemplars despite the idiosyncratic manifestations of their characteristic features.

When dealing with geometric figures like triangles, the problem can be solved by learning an appropriate set of geometric transformations. The right series of affine transformations—contractions, expansions, dilations, rotations, or shears—could transform any arbitrary triangle into any other arbitrary triangle. Cognitively, however, this would be an overkill, for we do not require that the exemplars of a common category be rendered perceptually indistinguishable from one another. We only more modestly require that they be rendered more mutually similar to one another than to members of opposing categories with which they might initially appear more similar—such as a square or rhombus with lines of the same length and angular rotation—from which they need be reliably distinguished.

This more subtle view has been attributed to ventral stream visual processing in the mammalian brain by neuroscientists; to articulate this solution, we need to develop the notions of a perceptual similarity space and category manifolds as regions therein (for development and critical examination of this conceptual apparatus, see Churchland, 1989; Gärdenfors, 2004; Gauker, 2011). Perceptual similarity space is a multi-dimensional vector space—with each dimension standing for a perceptually discriminable feature—that plots an agent's perceptual experience of each exemplar to a unique vector. Vector distance in this space marks the degree of perceived similarity between the different exemplars. A "manifold" in perceptual similarity space is a region of this vector space, and category representations can be construed as manifolds. Conceived in this way, the problem facing perceptual categorization is that, as DiCarlo, Zocolan, & Rust (2012) put it, nuisance variation causes "the manifolds corresponding to different [abstract categories to] be 'tangled' together, like pieces of paper crumpled into a ball" (p. 417). The task of both the brain and artificial agents is to find a series of operations—systematic transformations of this space—that reliably "unfold" the papers corresponding to different categories so that they can be more easily discriminated. More specifically, agents must learn a series of transformations of perceptual similarity space that map disparate triangles into a nearby points in a transformed *triangle* manifold, and ensure that this manifold marks a region that is linearly separable from the manifolds corresponding to *square* or *rhombus* (DiCarlo & Cox, 2007).

Like most of the processing of the visual cortices, the assumption is that these transformations and the manifolds they create are subpersonal and not available to introspection. We should thus not assume that their structure will be familiar or intuitively easy to understand; but we might obtain some rough metaphorical purchase by considering examples from the visual arts. Visual artists have long meditated on the difficulty in representing general categories from multiple perspectives and higher levels of abstraction. Matisse's *The Back Series*, for example (Fig. 2), consists of four bas-relief sculptures of increasing levels of abstraction. From left to right, the sculptures provide an increasingly indefinite representation of a woman viewed from behind, gradually adjusting for idiosyncratic positioning of the limbs and hips. By the fourth sculpture, we have a representation which perhaps looks unlike any particular woman viewed from behind, but bears more aggregate similarity to the full range of different positions such a figure might occupy. This provides a visual example of the sort of transformed, pared-down, late-stage representation that has often been cited as the essence of abstraction in artistic and scientific creativity (Camp, 2015; Chatterjee, 2010).

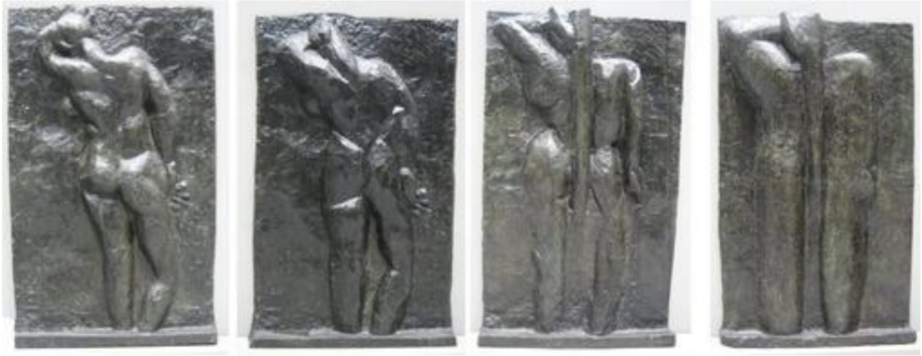

Figure 2. Matisse's *The Back Series*, discussed also in Patel et al. (2016).

In the history of art, one might find similar trends in the transition from Impressionism to proto-Cubism to analytical Cubism, which progressed through discoveries that increasingly abstract subjects can be represented with a series of swirling brushstrokes, geometric shapes viewed from inconsistent angles, and finally as a jumble of heterogeneous features in inconsistent poses. The most abstract representations are difficult to recognize as figures at all, and resemble no specific exemplar; but they can somehow capture the gist of a category, such as Picasso's portrait of a bullfighting fan (*l'Aficionado*) in 1912, which consists of a motley assemblage of visual themes from bullfighting arrayed in roughly appropriate spatial locations (Fig. 3). These sorts of images might be the closest we can come to visualizing the structure of abstract subpersonal category manifolds; and if we tried to describe them with words, we would end up with just the sort of jumble that Berkeley would mock. Yet these comparisons provide suggestive evidence that such transformational abstraction is performed by humans—that perhaps expert artists have, by experimenting with their aesthetic responses to diverse transformations, reverse-engineered the intermediate forms of abstract representation implemented by their subpersonal perceptual systems.

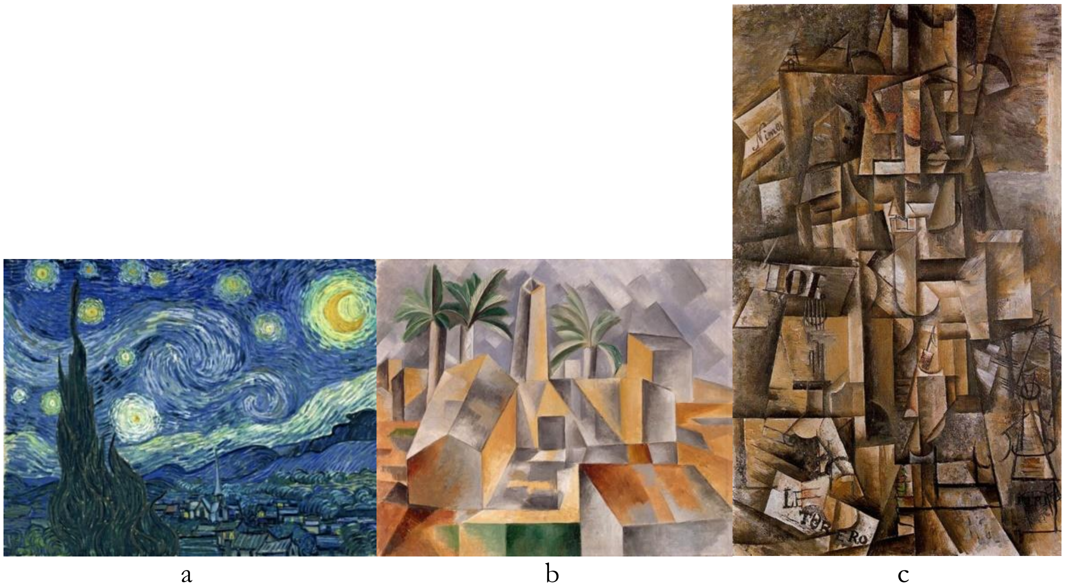

Figure 3. Series of artworks arranged from less to more abstract: (a) Van Gogh, *The Starry Night* 1889, (b) Picasso, *Brick Factory at Tortosa*, 1909, and Picasso, *L'Aficionado*, 1912.

Returning to Locke, we now have a sturdier conceptual foundation to support the troublesome triangle passage. Against Berkeley and Hume, Locke need not be interpreted as here suggesting that the general category representation of a triangle is an introspectible mental image with inconsistent properties. Rather, the general idea of a triangle might be something more subpersonal, like a transformed category manifold that, if it could be coherently imaged at all, would look more like the abstract art pieces just discussed. This is the sense in which the Lockean representation of *triangle* might involve both all and none of those variations; it controls for them by transforming idiosyncratic exemplars into an abstract representational format that adjusts for nuisance variations, locating exemplars of a common category as nearby points in a transformed manifold. The general manifold itself, however, consists in a whole region of similarity space that—like Picasso's *L'Aficionado*—should not be interpreted as depicting a single coherent exemplar with some particular configuration of nuisance parameters. This analysis provides us with an alluring solution to Problem #1; Locke's comments might be seen as struggling to express a theory of abstraction that was beyond the reach of his day's philosophical and mathematical lexicon.

Fascinatingly, the resources required to solve Locke's Problem #1 suggest a corresponding solution to Hume's Problem #2—for similar transformations might be used to generate particular exemplars for demonstrative reasoning, as required by the theories of Berkeley and Hume. If the system retains the unique location of the exemplar vector within the transformed category manifold, then it might be able to return that exemplar to its original configuration of nuisance parameters by deploying those transformations "in reverse". Returning to DiCarlo et al.'s paper-folding metaphor, if a system that learned to "unfold" the manifolds for different categories could "re-fold" them in a similar manner, then each vector in an abstract category manifold could be remapped to its original perceptual representation with the appropriate values of its original nuisance parameters like pose, position, scale, and so on. This begins to look more like the theory of abstraction provided by Kant (and the contemporary Kantian, Barsalou—see Gauker 2011, p. 67), who emphasized the need for "rules of synthesis" to generate a range of specific possible exemplars corresponding to an abstract category. Yet with bidirectional transformations in hand, the task might be performed without any explicit rules; and if the transformations are learned from experience using domain-general mechanisms (and we refrain from transcendentalist indulgences), the view still counts as empiricist in the relevant sense. This would in turn provide us with a straightforward way to generate exemplars from abstractions, providing a solution to Problem #2. Call the method which provides these two solutions to Problem #1 and Problem #2 "transformational abstraction".

The discussion of these solutions has thus far been mostly intuitive and metaphorical. If not by explicit rules, it remains to be explained how these transformations are actually achieved. In the next two sections, I will show that DCNNs perform just these sorts of transformations to categorize exemplars according to abstract categories and generate specific or novel exemplars of abstractions. This approach vindicates elements of the Lockean, Berkeleyan/Humean, and Kantian views; but again we are here less interested in historical scorekeeping than in explaining the distinctive success of DCNNs. The next two sections will explore the characteristic features of DCNNs to illustrate how they distinctively suit them, compared to other neural network architectures, to model transformational abstraction.

4. DCNNs: A Primer

DCNNs are a type of neural network that can broadly be placed under the rubric of connectionism; however, their characteristic structure crucially differs from the 3-layer perceptron networks that were ubiquitous during the 80s and 90s, and so may be more familiar to most readers. In the next section, I will argue that these differences allow DCNNs to model core aspects of transformational abstraction in the mammalian brain. This argument will hinge on three features which jointly differentiate DCNNs from other kinds of networks: depth, convolution, and pooling. Each of these features provides significant gains in computational efficiency and representational resource consumption when deployed on tasks with high nuisance variation, compared to networks that lack these features. In this section, we will elaborate these three features and distinguish deep convolutional networks from their intellectual forebears.

To contrast them with DCNNs, let us characterize traditional networks in terms of two features: shallowness (only 1-3 layers) and uniformity (containing only one type of processing node, usually involving a sigmoidal activation function). The engineering success of these networks led to a corresponding Golden Age of innovation in philosophy of mind and cognitive science. This work introduced to contemporary discussion many ideas which were revolutionary by the standards of the more classical, symbol-based methods they aimed to supplant: massive parallelism, soft constraints, distributed representation, gradual learning, graceful degradation, and the importance of reproducing subject's errors as well as their successes (for review and discussion, see Clark, 1989). Today, the appeal of these properties is largely taken for granted by much cognitive science—and they are all simply inherited by DCNNs. However, as with the more symbolic, logic-based AI before it, the early promise of "shallow" connectionism gradually began to fizzle as the limitations of such networks become apparent.[6]

It was long speculated that just as the limitations of one-layer perceptrons could be overcome by going to two- or three-layer networks, the addition of even more layers could allow these networks to perform better still. Such speculations were bolstered by the findings that mammalian neocortex possesses a 6-layer structure and that visual processing in the ventral stream passes hierarchically through a series of cortical regions, roughly corresponding to the detection of less abstract features like contrast differences and boundaries, such as orientation and contrast in V1, lines and borders in V2, angles and colors in V4, shapes in TEO/PIT (posterior inferotemporal), and finally to figures and objects in TE/AIT (anterior inferotemporal) (comparisons which hold up well today—see Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016--and see Fig. 4). Perhaps one problem, in other words, was that the shallower Golden Age networks were not yet deep enough to replicate the kind of hierarchical sensory processing characteristic of mammalian cortical processing.

In fact, a distinct tradition in connectionist research in the 1970s—an alternative branch of the family tree that eventually produced DCNNs—had already demonstrated the computational payoff that could be achieved by deepening networks. This tradition was inspired by an anatomical discovery in cat visual cortex by Hubel & Wiesel (1962). Using single-cell recordings in cat areas V1 and V2, they identified two different cell types, which they dubbed "simple" and "complex" cells, based on their differential firing patterns. Whereas simple cells seemed to detect a low-level feature like an edge or grating in a particular orientation and position, complex cells took input from many simple cells and fired in response to the same features but with a greater degree of spatial invariance. Neuroscientists at the time speculated that many layers of these simple and complex cells might be found interspersed in the visual processing stream, and their interplay might explain our own ability to recognize increasingly abstract features in diverse locations and poses.

[6] In the interests of space, we move quickly over the history here; for more background and discussion, see (Buckner & Garson, 2018; Schmidhuber, 2015)

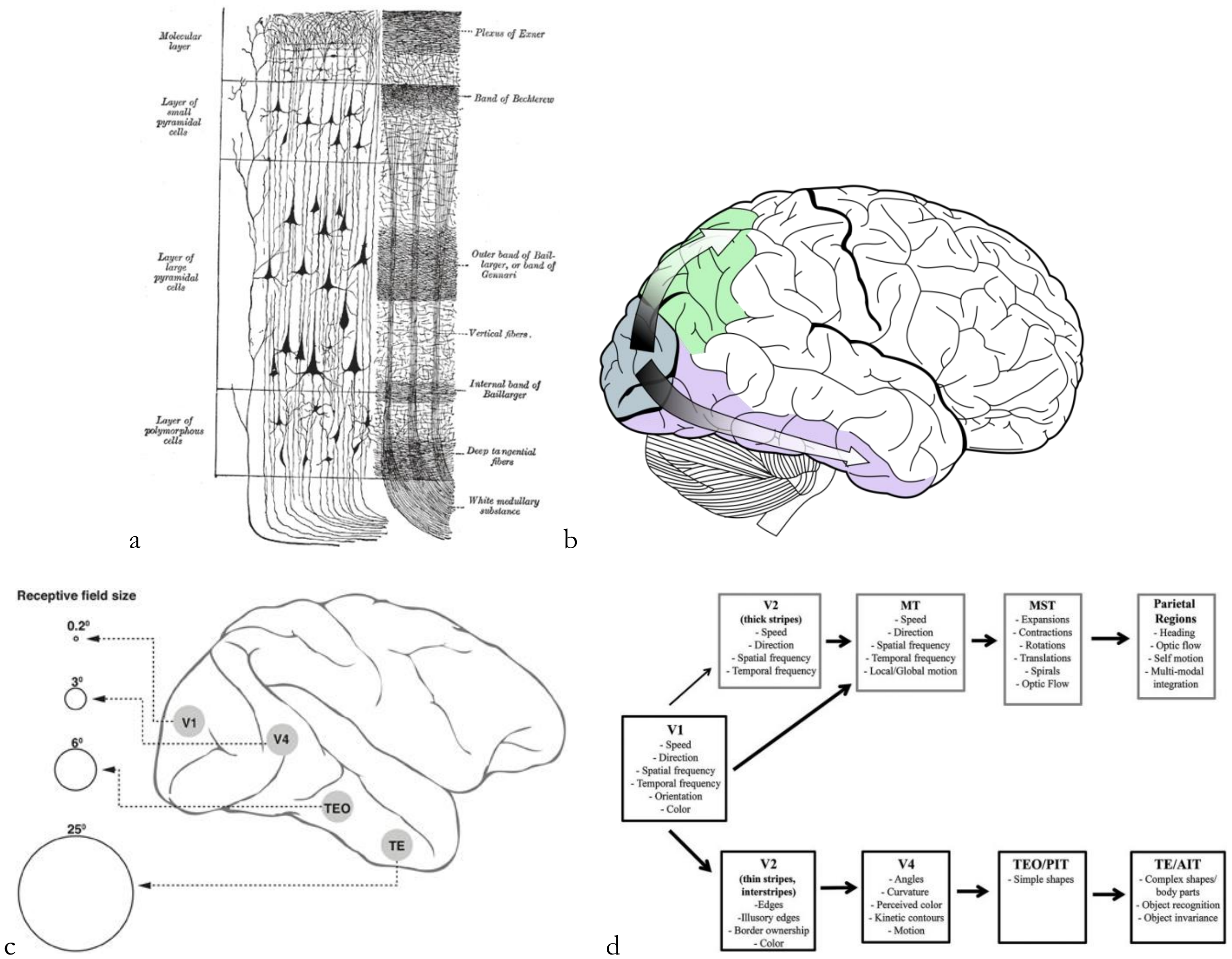

Figure 4. Images detailing laminar cortical structure and hierarchical processing flow in dorsal and ventral streams. Mammalian neocortex has a well-known six-layer laminar structure (a) and sensory information streaming in from visual and auditory sources proceeds through a processing cascade in early anterior sensory areas in V1 to late processing areas in TE/AIT (b). As it proceeds through the hierarchy, the receptive field size of the areas grows larger (c), processing larger and more configural features and focusing on increasingly abstract information (d). (Figures from various sources pending permissions requests.)

To demonstrate the computational power of this neuroanatomical division of labor, Fukushima (1979) designed a new kind neural network called Neocognitron. Neocognitron was perhaps the first network that was truly "deep" (with 4-10 layers, depending on how they are counted), but its most powerful innovation was the way it leveraged two different types of operation—linear convolutional filters and nonlinear downsampling—to combine two different types of processing in a single network. Though at some level of description these operations are all just mathematical operations, it can be useful to conceptualize these two different kinds of processing as taking place in two different kinds of nodes corresponding to the simple and complex cells in the mammalian visual cortex. Fukushima's "simple" nodes performed the convolution operation (a type of linear algebra operation elaborated below) to detect features at particular locations and in particular poses; and his "complex" units took input from many spatially nearby simple units, aggregating their activity to detect those features across small shifts in its location or pose. Several layers of such paired convolution and subsampling layers were iterated hierarchically, such that processing gradually detected more and more abstract features across a wider range of visuospatial variance. With these innovations, Neocognitron was able to outperform the perceptrons of the day on difficult tasks characterized by high variance—such as handwritten digit recognition—by modeling the hierarchical processing cascade of mammalian neocortical processing streams.

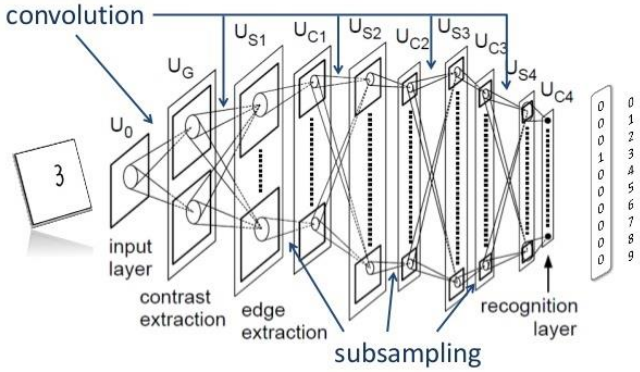

Figure 5. A depiction of Neocognitron's architecture as applied to digit recognition, from Fukushima (2003).

Each of these operations bears elaboration, for the germ of DCNN's distinctive computational potential is already present in any network which combines them in this manner. Let us begin with convolution. Perceptual input is typically passed to such a network in a gridlike structure—a 2-D grid for image processing tasks, or time-slices of audio information for auditory processing tasks. For ease of exposition, let us focus on visual examples in what follows. The smallest unit of information in a visual grid is often a pixel, which itself is typically a multi-dimensional vector of Red, Green, and Blue color channel intensity detected at that location. Convolution is a linear algebra matrix operation that can be performed to transform the vector values for a spatial chunk of pixels (usually a rectangle) in a way that maximizes some values and minimizes others. In practice, the convolution operations that are useful for image recognition are those that tend to amplify the presence of a certain feature and minimize other information for a given chunk. These convolutional nodes are called filters or kernels; a useful convolution for a vertical-edge kernel might be one that maximizes values corresponding to a horizontal edge, and minimizes other values. Each convolution operation is then typically passed to a rectified linear unit (ReLU)—this is sometimes called the "detector" stage of filtering—which activates using the rectification function (Figure 6) if the output of convolution exceeds a certain threshold. In other words, the output from convolution only passed up the processing hierarchy if the feature is detected at that location (for simplicity, I hereafter write as though the convolution and detection operations are performed by a single node). The net effect of passing this vertical-edge kernel across the whole image would be a representation that shows all and only the vertical edges.

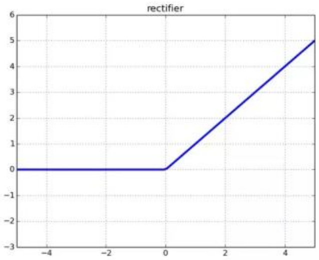

Figure 6. The rectification activation function, argued to be more biologically plausible than sigmoidal activation because the synapse fires and then gradually ramps up activation.

Typically, however, the recognition of a general category requires more than merely recognizing vertical edges; we need to detect a wider diversity of presentations, such as edges in different orientations, and to assemble that wider range of presentations into useful composites like shapes or digits. The addition of Fukushima's "complex" units help us achieve this grander goal by taking input from many spatially nearby convolutional nodes below, and using a downsampling technique (Neocognitron uses spatial averaging) to activate if any one of their inputs were sufficiently active. Using downsampling, we can now efficiently express the fact that an edge occurred approximately here in some spatial orientation, irrespective of exactly where it appeared or how it was oriented. The net effect of multiplying the input by a variety of edge-detecting kernels and combining their outputs using downsampling is like applying an edge-detector filter in a digital photograph editing program; the result is a simplified image representation that reveals all the edges wherever located and however oriented, and "subtracts out" all other information (Fig. 9).

$$\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\right)[2,2] = (i \cdot 1) + (h \cdot 2) + (g \cdot 3) + (f \cdot 4) + (e \cdot 5) + (d \cdot 6) + (c \cdot 7) + (b \cdot 8) + (a \cdot 9).$$

Fig 7. Example of a component of a convolution operation (reproduced from Goodfellow, Bengio, & Courville, 2016), where the filter (kernel) and receptive field are both represented by 3x3 matrices. Convolution is the process of flipping both rows and columns of the kernel and multiplying locally similar entries and summing. The result of the sums would then be stored in the convolved matrix output. This shows the summation operation for the [2,2] location of the resulting image matrix weighted by the convolutional filter.

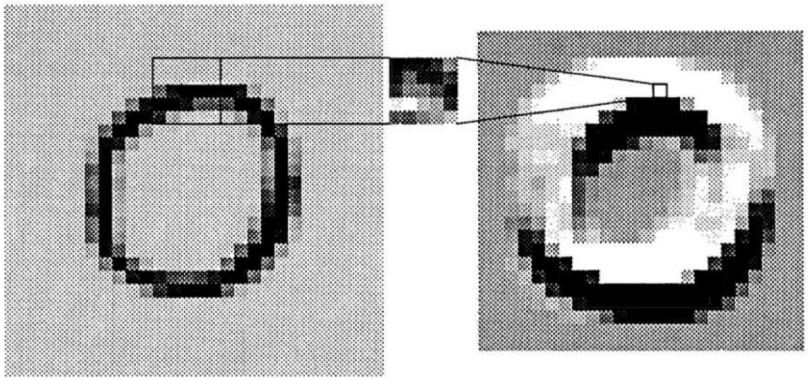

Figure 8. An example of the visual output of a single convolutional kernel on real handwritten digit data in a trained convolutional network (reproduced from LeCun et al., 1990, 399). This (learned) kernel detects something like curves at the top or bottom of a digit image. When combined with other kernels using downsampling, it might detect curves at digit periphery irrespective of angular orientation.

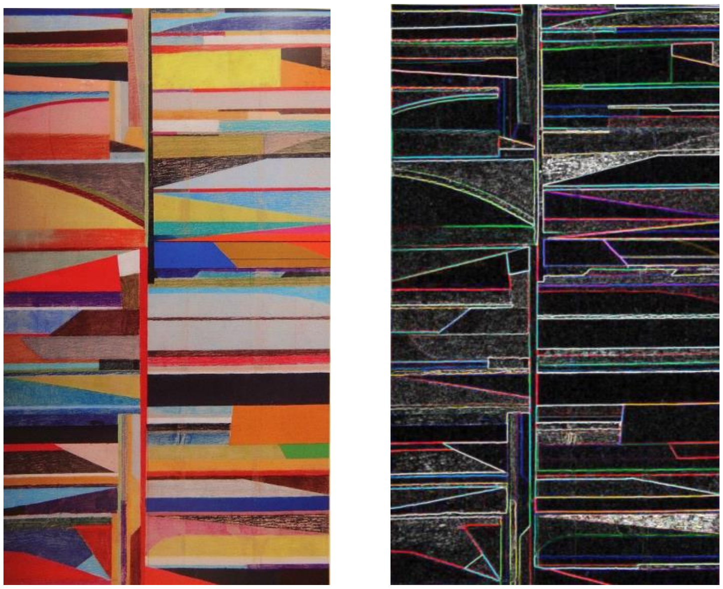

Figure 9. *21st Century* by Viktor Romanov, with and without a Sobel edge-detection filter, one popular convolutional edge-detection algorithm. (Operation performed by author in Gimp 2.0 Freeware image editing program. Image credit: Yugra News Publishing, Wikimedia Commons CC License.)

Despite Neocognitron's intriguing architecture and impressive performance, DCNNs all but vanished from the spotlight for more than two decades, for it proved difficult to train them. Fukushima calibrated Neocognitron using a complex combination of manual pre-wiring and unsupervised learning. Despite some early successes (LeCun et al., 1989), when backpropagation was applied to such deep networks, learning tended to settle into poor local minima or fail to stabilize at all (DeMers & Cottrell, 1993). Modelers came to describe the problem as that of "vanishing or exploding gradients"; as error signals would either shrink rapidly or grow out of bounds as they were backpropagated through too many successive layers (Hochreiter, 1991), resulting in chaotic link modification during training that failed to converge on optimal performance. DCNNs thus did not re-appear in the spotlight until two developments overcame the problem of training. First, exchanging sigmoidal activation functions for the more biologically-plausible rectification function helped minimize the possibility that gradients would vanish through multi-layer backpropagation (because the gradient of a rectification function is always 0 or 1—Hahnloser, Sarpeshkar, Mahowald, Douglas, & Seung, 2000). Second, Hinton and colleagues (Hinton & Salakhutdinov, 2006) discovered a way to use unsupervised pre-training to initialize weights in the intermediate layers of deep networks to the general statistical properties of their environment, which further improved performance (though this step is no longer necessary in state-of-the-art networks). This coupled with a general increase in computational processing power provided by powerful, specialized, and affordable graphics processing units sent deep convolutional networks roaring back onto the scene after 2010.

Though DCNNs are perhaps the most widely successful approach in machine learning today, until recently it has been difficult to explain why they work so well or what kind of qualitative operation they were performing. An answer to this question is complicated by the fact that over the last ten years, innovations in DCNNs have been often driven by practical, performance-oriented considerations other than biological plausibility. As a result, many state of the art networks include dozens of heterogeneous tweaks, including pre- and post-training weight adjustment, different kinds of modules and sub-networks, rules for adding or deleting nodes as performance improves or deteriorates, and the use of many more layers (up to 250) than could plausibly be attributed to cortical processing. Fortunately for the purposes of cognitive modeling, however, some prominent DCNN modelers have begun to dial back to more biologically-plausible parameters and principles (e.g. Hassabis, Kumaran, Summerfield, & Botvinick, 2017). At any rate, to analyze the shared computational power of DCNNs, we should focus on core features shared by networks that reliably succeed on the kinds of perceptual and intuitive judgment tasks on which humans and animals excel.

Following several other analyses (DiCarlo et al., 2012; Montufar, Pascanu, Cho, & Bengio, 2014; Patel, Nguyen, & Baraniuk, 2016; Schmidhuber, 2015), I characterize the computational core of these networks in terms of three features: (1) many layers of hierarchical processing which interpolates two different kinds of computational nodes, (2) linear convolutional filters and (3) non-linear "poolers".[7] In state-of-the-art networks, the downsampling role in Neocognitron is performed by an operation called max-pooling. Max-pooling is even more biologically plausible and effective at filter aggregation than Neocognitron's spatial averaging, for its operation is computationally very simple—simply pass along the greatest activation (above a critical threshold) amongst the inputs taken from filter nodes at spatially nearby locations (Figure 10). Other pooling functions—such as averaging or min-pooling—have been explored, but max-pooling tends to produce better results on the kinds of perceptual classification or intuitive decision-making tasks at which humans and animals distinctively excel.[8] These are the shared core principles of DCNNs we shall focus on below.

[7] Note that when DCNNs are deployed for categorization or other forms of decision-making, the final layer of the network will typically be a fully-connected classifier that takes input from all late-stage features (i.e. a fully connected layer of nodes or set of category-specific support-vector machines). These are used to draw the boundaries between the different category manifolds in the transformed similarity space. Since these components are deployed in many other machine learning methods that do not model transformational abstraction, I do not discuss them further here.

[8] An important current point of controversy is whether specifically max-pooling is required to reduce the search space and avoid overfitting, or whether other methods might be as effective. For two poles in this debate, see (Patel, Nguyen, & Baraniuk, 2016; Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). The present paper holds that even if alternative solutions are also practically effective, biologically-relevant networks must somehow implement the aggregative role of complex cells—though max-pooling is perhaps only one possible technique in a family of implementations that can accomplish this (DiCarlo & Cox, 2007).

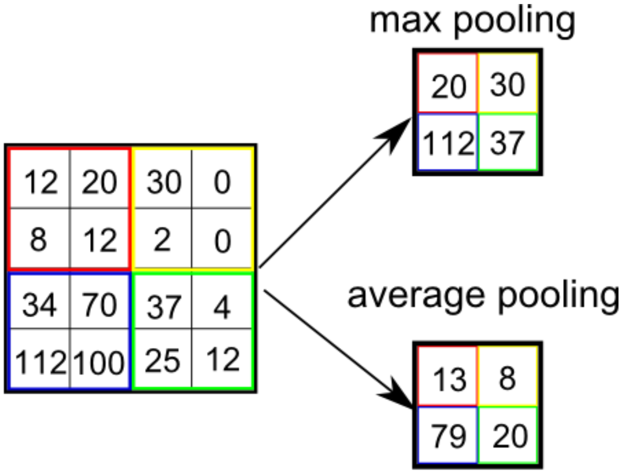

Figure 10. A comparison of max-pooling with average-pooling for aggregation across activation received from the same receptive fields.

5. Transformational Abstraction in Deep-Learning Convolutional Networks

So, why do DCNNs perform so well, relative to neural networks which lack these features? The effect of their three computational components is best understood cooperatively and holistically; the key, I argue, is that they jointly implement a form of hierarchical abstraction that reduces the complexity of a problem's feature space (and avoids overfitting the network's training samples) by iteratively transforming it into a simplified representational format that preserves and accentuates task-relevant features while controlling for nuisance variation. In short, they perform transformational abstraction, as characterized in Section 3.

The operations needed for transformational abstraction's solutions to Problems #1 and #2 above are performed by the cooperation between the two different kinds of processing that we identified with DCNNs above. The classical empiricists never specified a plausible mechanism that could perform the crucial "leaving out" of irrelevant information highlighted in abstraction-as-subtraction. This, I argue, is the role played by max-pooling units, which provide a computationally simple and neurally-plausible method. At each convolutional layer, nearby linear filter nodes detect some feature-type in different idiosyncratic presentations, such as lines of different positions, sizes, or angular orientations. The non-linear pooler nodes then take input from the various configurations in which a feature can present itself, and pass along the maximum activation to the next layer of linear filters. Later layers can now tolerate the idiosyncrasies amongst the ways a feature might present itself and focus computational resources on configuring the next level of features from these more complex blocks, such as building angles from adjacent lines, figures from lines, and so on up to the most abstract features.

As concerns the travel from exemplars to abstractions, DCNNs structured around these basic principles have achieved remarkable levels of accuracy in identifying and categorizing objects in natural photographs, even the specific category of *chair* highlighted by Brooks. Performance of a DCNN is often evaluated using curated image collections, such as PASCAL VOC. The 2012 version of this dataset contains 11,530 training and validation images (though many nets pre-train on much larger datasets such as ImageNet) and 20 object classes—and the most difficult of these to discriminate, judging by average accuracy results, is indeed *chair*.

Figure 11 shows some examples of the images coded as containing at least one chair in the dataset. As the reader can see, there is an incredible amount of diversity in the presentations of chairs in terms of their nuisance parameters. Nevertheless, a variety of DCNNs have achieved accuracy rates from 50-70% of labeling chairs in these complex images (Fig. 12). Though there is intense debate as to whether further gains in DCNN performance have plateaued or are about to plateau, this is already a remarkable achievement, one that until recently seemed out of reach. And though there is a considerable variation in the implementation details of the top-ten DCNN models in the VOC 2012 leaderboard, they all involve the core features discussed in the previous section.

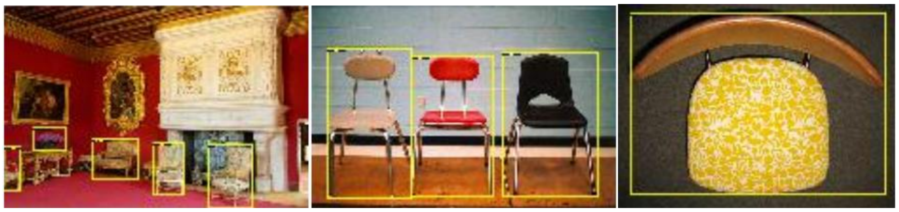

Figure 11. Three images from Pascal VOC 2012 that contain at least one chair.

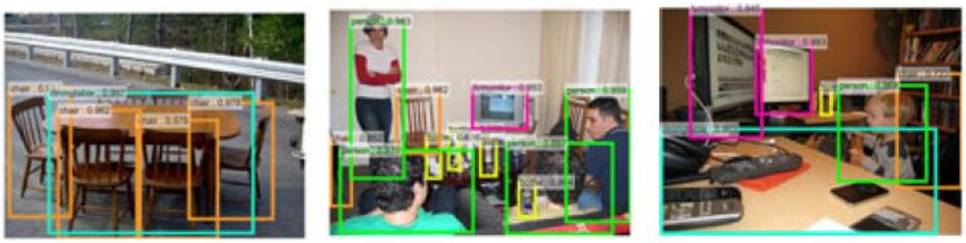

Figure 12. Chairs identified by one particularly successful DCNN, Faster R-CNN, on three Pascal VOC test images (Ren, He, Girshick, & Sun, 2017). Faster R-CNN speeds categorization by deploying an additional DCNN to direct processing to region proposals for objects (intended to simulate the role of attention); in this case, the orange region proposal boxes indicate chairs identified by the network.

During training of these DCNNs, their link weights converge on the sequence of transformations that do the best job of solving the widest range of categorization problems, by increasing the net distance in feature space between the manifolds for categories which must be discriminated. This transformational ability explains how DCNNs can recognize the abstract similarities shared amongst all chairs—and remarkably, no explicit definitions are involved or required for it to do so. Moreover, the network itself discovers the right series of transformations to perform from labeled training data; as Goodfellow et al. 2016 put it, "[pooling] over the outputs of separately parametrized convolutions [allows the network] to learn which transformations to become invariant to" (2016, 337) for some range of categorization tasks. Moreover, insofar as those transformations and the intermediary features they reveal are useful for the recognition of other categories as well—the transformations useful for recognizing chairs may closely resemble those useful for recognizing other objects like tables or beds—the network will enjoy accelerated performance on those other categories as well.

These hierarchical transformations explain why DCNNs have achieved a leap in computational power that can overcome even the extreme branching factor of a game as complex as Go. The distinctive power cannot be expressed quite so simply as that they, for example, implement a different class of Turing machine, or learn a type of function that remains in principle beyond the reach of shallower and more homogeneous networks. At some level of mathematical abstraction, all neural networks perform pattern recognition or dimensionality reduction, and the universal approximation theory shows that in principle a three-layer perceptron with an unlimited number of nodes can approximate any continuous function

(Cybenko, 1989; Hornik, 1991). However, this does not entail that these networks could be trained to approximate the operations of a DCNN in practical conditions or perform their categorizations at anywhere near the same speed. State-of-the-art DCNNs, on the other hand, learn these transformation themselves from messy natural images using domain-general learning mechanisms, and can compute category classifications for an arbitrary photograph in real time.

The cooperative gains enabled by convolutional filtering, pooling, and depth can be understood by how they enable DCNNs, when compared to earlier NN architectures, to have much sparser connections, share parameters more widely, and iteratively simplify the perceptual similarity space in which category membership is computed (for a longer discussion, see Goodfellow et al., 2016, Ch9). First, because the receptive fields of filter kernels are smaller than the whole image, both filter and pooling units need only be connected to several spatially nearby units in the input below, further reducing the number of link weights that need to be learned and stored. Second, the parameters for many different filter nodes can be shared across the whole network, greatly reducing learning time and memory usage at every convolutional layer. If there are m inputs and n outputs, then the matrix operations of a fully-connected network would normally require m x n parameters, and the computational complexity of computing output would be O(m x n). If the kernel requires only k connections, the runtime is O(k x n); and since k is often orders of magnitude less than m, these gains can be quite large.[9] Moreover, because the same kernels are applied to every chunk of the input image, this precludes the network from having to learn and store new parameters to detect every feature in every spatial location, providing further reductions in memory usage and gains in statistical efficiency. Finally, these benefits are cumulative as we proceed through layers of the network, for once the features are recognized by the filter nodes and aggregated by the pooler nodes, later layers of the network need only consider these simplified image representations.

To return to our frame of empiricism vs. nativism, DCNN modelers often express the joint effect of these gains in terms of infinitely-strong, domain-general priors (i.e. Goodfellow et al. 2016, p337). Parameter sharing, max-pooling, and sparse connections all capture the prior probability estimation that the precise pose

[9] For a worked example, see Goodfellow et al. (2016, p. 334), who show that edge detection alone can be roughly 60,000 times more computationally efficient when performed by a DCNN, compared to a traditional 3-layer perceptron.

or location of a feature is not relevant to an object's general category membership(s). Whenever the recognition of many different categories is complicated by the presence of similar nuisance parameters, DCNNs will tend to perform much better than alternative methods that do not somehow implement these priors. Environmental regularity thus provides the other side of the equation in explaining DCNNs' success, for the forms of variation that must be factored out for discrimination success are often common for many different categories. This analysis provides precise, graded, and multi-dimensional definition of abstraction: one representational format is more abstract than another, relative to some range of classification tasks, if it is more tolerant to the forms of nuisance variation that must be overcome to succeed at those tasks.

At this point, we are in a position to ask whether the form of transformational abstraction performed by DCNNs is merely perceptual, or whether it might be extended to cover more theoretical or amodal category representations. Notably, the definition of abstraction just provided applies to the most abstract properties like *cardinality* or *logical validity*, for the forms of variation that might need to be overcome to recognize those properties might include complete permutation of the model's domain. However, it is an open question whether an unaided DCNN could discover or implement the full range of transformations required to detect these properties—a daunting task, because it is unclear how convolution and pooling could be bootstrapped to evaluate complete permutations of a set or domain. These properties require transformation at a kind of logical limit that may be difficult for any method to achieve without explicit formulation of hypotheses regarding the set of permutations that must be evaluated for a complete assessment, or quantificational resources to describe that set. Additional components corresponding to these resources might need to be added to DCNNs for them discover mathematical or geometric properties themselves.

On the other hand, it would also seem that the capacities of this form of abstraction already exceed the strictly perceptual, covering any domain with sources of nuisance variation that could be mapped to geometric dimensions and overcome by convolution and pooling. Board configurations in Go, for example, are not limited to visual or tactile modalities, and they were fed to AlphaGo in symbolic form. Some critics worry that this renders AlphaGo's achievement less impressive since its input was partially pre-digest (i.e.

Marcus 2018); but on the other hand, it shows that the kind of transformations enabled by DCNN are not narrowly limited to information vehicled in visual or auditory sensory modalities. Moreover, AlphaGo's success at recognizing the kinds of board-wide abstractions that helped it defeat Lee depended upon its DCNNs' transformational ability to recognize patterns across rotations, reflections, and dislocations. Just as with recognizing features in natural images, Go strategies need to be nuisance-tolerant, for abstractions like "influence", "connection", and "stability" are largely preserved across rotations and small shifts in spatial location. I thus consider that DCNNs have solved Locke's Problem #1 for at least mid-range abstract categories and in a way not strictly limited to individual perceptual modalities, and they do so using domain-general learning mechanisms.[10]

Let us now turn to the solution transformational abstraction provides to Hume's Problem #2; can a DCNN's transformations be reversed to generate exemplars from abstract categories? The prospects here might not initially sound promising; a simplistic understanding of abstraction-as-subtraction might suppose that the gains in memory usage and computational efficiency just reviewed require DCNNs to completely discard nuisance information as it transforms similarity space up the network hierarchy. In other words, DiCarlo's paper-folding metaphor might be thought to break down here; while no information is permanently lost by folding a paper, the gains in computational efficiency as we go up a DCNN's hierarchy must be bought at the expense of some lost information. Remarkably, however, the point to which an exemplar is mapped in an abstract manifold appears to contain more information about its nuisance parameters than can be recovered from the raw perceptual input fed to the network itself. Empirical analysis of DCNNs trained on visual discrimination has found that deepest layers actually contain the most accessible information about nuisance parameters (Hong et al. 2016). An exemplars' specific point in an unfolded category manifold appears to contain more information about nuisance variations specific to that abstract category; for intuitively, it is easier to represent any property of an object—even nuisance properties—after representing the object itself. The ease with which this preserved nuisance information can be repurposed to generate a

[10] One could also worry here that AlphaGo did not learn the rules of Go from experience, but this does not impugn the point. What is claimed is rather that once these rules were provided, a DCNN can learn play strategies without any domain-specific strategy heuristics (which knowledge of the rules do not provide). This is especially driven home by AlphaGo Zero, which acquired strategies entirely through self-play (Silver et al., 2017).

range of different but coherent exemplars has produced its own research area, and deep generative networks have already succeeded at a variety of tasks like handwriting simulation and artistic rendering. To tie back to our central example from Brooks, deep convolutional networks can now not only recognize chairs in diverse presentations, but also generate a range of novel but visually plausible exemplars along a range of diverse nuisance presentations (Fig. 13).

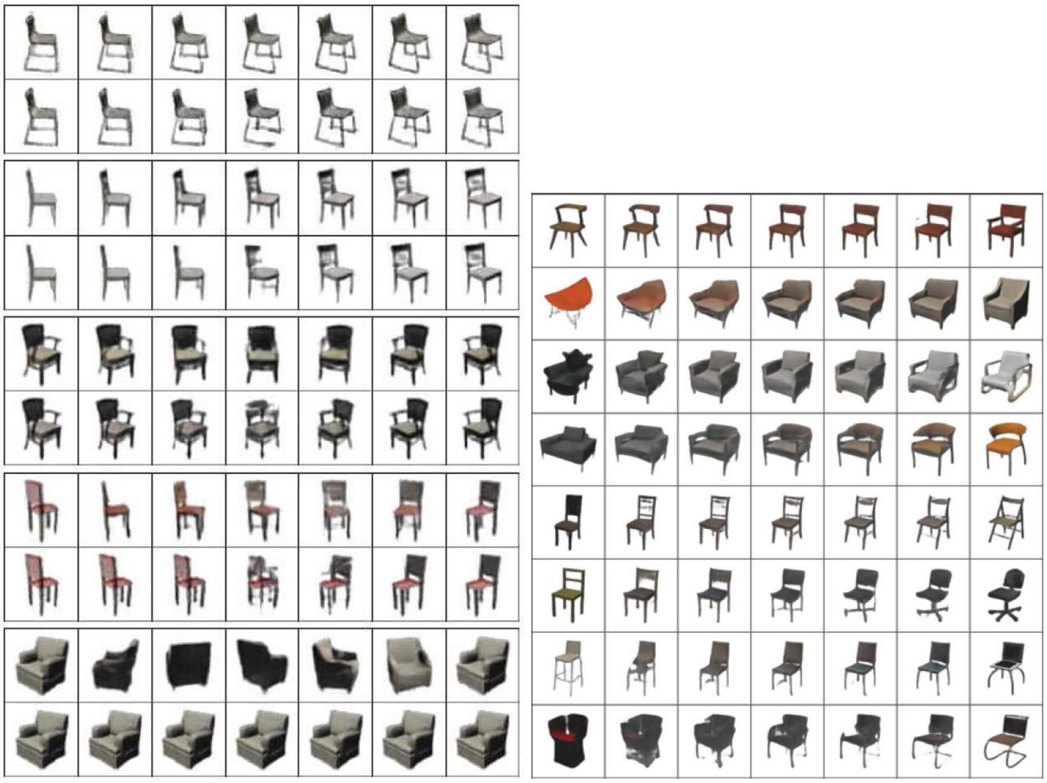

Figure 13. A range of chair exemplars generated by a network deploying "deconvolution" and "unpooling" operations. On the left are arrayed similar chair exemplars generated by the network in a variety of distinct nuisance presentations, and on the right a range of novel intermediate exemplars between two distinct chair exemplars. (Images reproduced from Dosovitskiy, Springenberg, & Brox, 2015).

To be frank, there are some obvious challenges in reversing the transformations performed by discriminative DCNNs; their transformations do discard some information in order to simplify the search for features as we go up the network, and merely "refolding" an abstracted exemplar vector produces some well-known errors. For example, a well-known side effect of naïve deconvolution is the "checkerboard" error, which reveals the costs of downsampling images in discretized windows (Fig. 14, Odena, Dumoulin, & Olah, 2016). "Undoing" a convolution to transform feature values closer to initial perceptual presentation can be achieved by applying its transpose; but reversing max-pooling is a bit trickier, since it does discard some information about the exact way in which the feature was detected. However, a variety of methods can now overcome these difficulties; solutions range from supplementing unpooling with prior probability estimations to produce the likeliest manifestation of properties at each level of abstraction, given some assumptions—even domain-general priors such as "contrast levels tend to be similar in nearby pixels" help considerably—to adding new layers of generative network separately trained to reconstruct a range of novel but coherent ensembles of feature presentations. These discriminative/generative hybrids demonstrate that the opposition between the Lockean and Humean approaches to abstraction is unnecessary; mathematically, they are two sides of the same coin, and both are required for a full empiricist account of demonstrative reasoning that allows bi-directional travel between exemplars and abstractions.

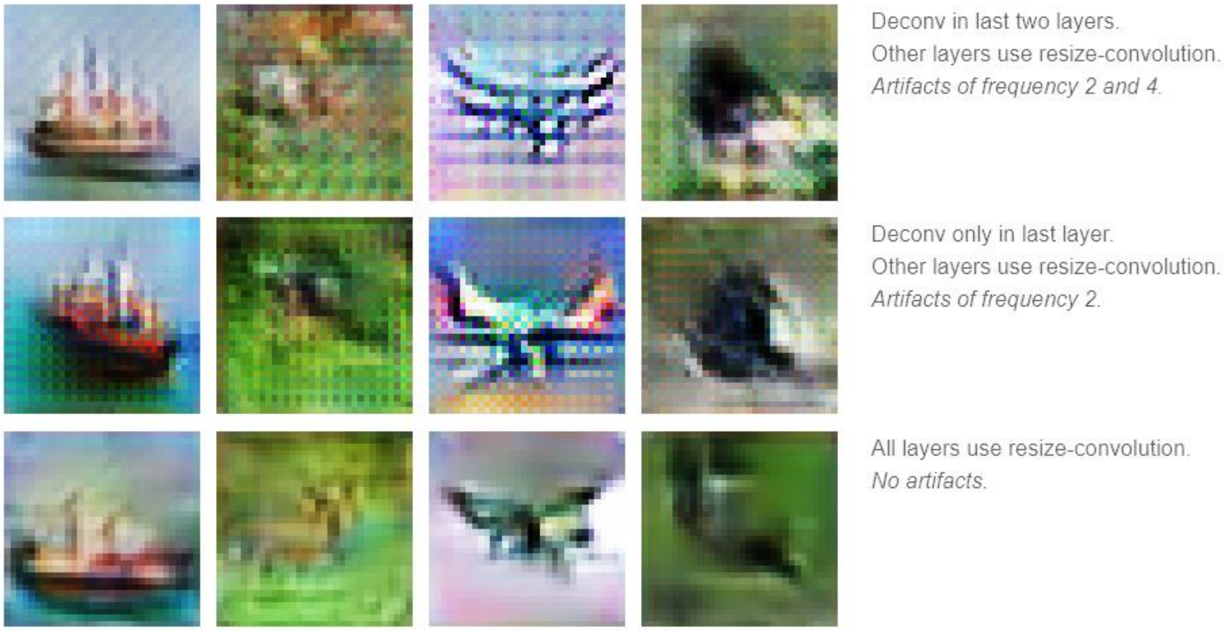
Deconv in last two layers.
Other layers use resize-convolution.
Artifacts of frequency 2 and 4.
Deconv only in last layer.
Other layers use resize-convolution.
Artifacts of frequency 2.
All layers use resize-convolution.
No artifacts.

Figure 14. Examples of checkerboard artifacts produced by image deconvolution with and without corrective measures, reproduced from Odena et al. (2016).

To illustrate the generative potential of the transformational knowledge contained in these networks, consider a particular application that can produce sophisticated artworks using networks trained only for classification. In the DeepArt project, modelers found that, once trained on an object recognition task, a deep convolutional network contained abstract information not only about the objects contained in the input image, but also about what we would normally refer to as an input image's "style" (Gatys, Ecker, & Bethge, 2016). Modelers were able to recover a separate, hierarchical style representation for images by adding a parallel feature map for each layer that recorded only correlations between filter nodes at that layer. In other words, each layer of the primary convolutional network contained increasingly abstract information about what was depicted in the image, whereas the correlational feature maps contained increasingly abstract information about the way in which those objects were depicted. By taking the high-level abstract representation of the objects depicted in a photograph and reversing the transformations performed by the network so as to maximize the correlations found in a different, style-source image (that in essence served as the source of the hierarchical rendering priors), the modelers could flexibly depict the objects portrayed in the photograph as having been painted in a variety of different source styles. For example, a photograph of a row of buildings was flexibly be depicted as painted in the style of Van Gogh and Munch (Fig. 15).[11] Again, the authors speculate that this method works because the convolutional network has to adjust for stylistic variance—as yet another form of nuisance—to understand the objects depicted in diverse images.

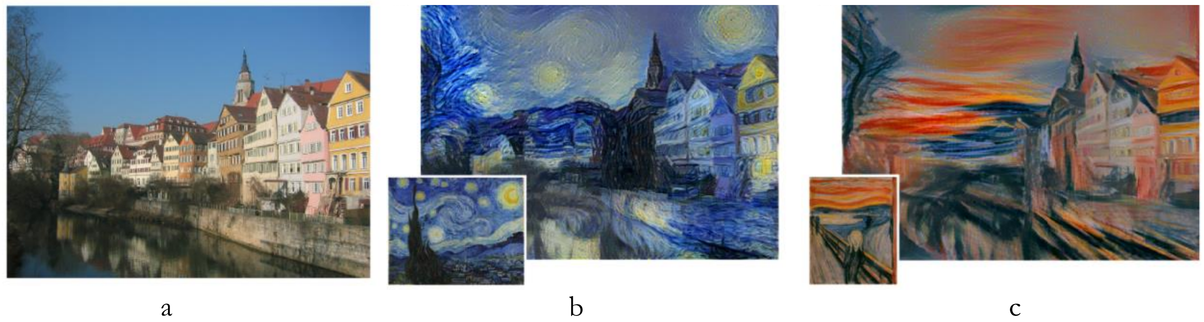

Figure 15. Examples of output from the DeepArt project, with an image source (a) depicted in two different source-image styles: from Van Gogh's *Starry Night* (b) and Munch's *The Scream* (c).

6. DCNNs as Cognitive Models: What Exactly Is Sketched?

I thus take it to be established that DCNNs implement a powerful form of bidirectional translational abstraction that provides practical, non-magical solutions to Problems #1 and #2, and that they can deploy this ability to achieve remarkable levels of success on a range of tasks taken to require abstraction in humans and animals. The final question addressed by this paper is now one from philosophy of science: do these networks model transformational abstraction as performed by mammalian cortex? I approach this question

[11] Interestingly, the DeepArt team found that average-pooling was a more effective activation function than max-pooling when the network was in generation mode.

with tools from the literature on mechanistic explanations in psychology. Though the characteristic features of DCNNs arose from biological inspirations and computer vision modelers have used DCNNs to model human perceptual similarity judgments, two complications—revisions to Hubel & Wiesel's story about simple and complex neurons, and the discovery of adversarial examples—have recently challenged the ability of DCNNs to model mammalian neocortex. Below, we consider each in turn.

The naïve story about DCNNs as a model of perceptual neocortex goes something like this: DCNNs depict the mechanism that performs transformational abstraction in mammalian neocortex. To provide a mechanistic explanation, a model must depict the components and their organization in the target system that produce the phenomenon of interest. In this case, the components of the DCNN models must localize to components of the mammalian neocortex whose orchestrated functioning produces transformational abstraction, and those parts must produce those behaviors by really possessing the causal/structural features that characterize those parts in the model and (causally) interacting with one another as the model depicts.[12] Such a model should allow us to answer a range of what-if-things-had-been-different questions, including especially the predicted behavioral effects of a range of interventions on those components. DCNNs in fact do this, so the simple story goes (perhaps grounding a new method of "virtual brain analytics"—Hassabis et al 2017), because convolutional and pooling nodes correspond to simple and complex cells in the mammalian neocortex, which are organized in hierarchical layers as depicted in the layers of a DCNN. Thus, DCNNs provide a mechanistic model of abstract categorization and perceptual similarity judgments in mammals, or at least the aspects of those processes implemented by perceptual neocortex.

The first major roadblock for this interpretation of DCNNs is that the story about simple and complex cells has undergone significant revision since Hubel & Wiesel's pioneering work in 1962. Hubel & Wiesel surmised that the bimodal distribution of firing preferences in simple and complex cells was produced by a correspondingly bimodal distribution of structural connectivity patterns, with simple cells synapsing with a small number of spatially nearby input cells from the previous layer, and complex cells in turn synapsing with a number of simple cells with overlapping receptive fields. While more recent neuroanatomical work has

[12] This general characterization washes across differences in various canonical accounts of mechanism; see (Bechtel & Abrahamsen, 2005; Glennan, 2002; Machamer, Darden, & Craver, 2000).

upheld the bimodal distribution of firing patterns for simple and complex cells, it has instead found a unimodal (evenly distributed) set of connectivity patterns. Specifically, the structural explanation for the bimodality of the firing patterns appears to be explained instead by an interaction between unimodal connectivity patterns and those cells' nonlinearly thresholded membrane potentials (Priebe, Mechler, Carandini, & Ferster, 2004). The exact details are not important here, only the fact that this attenuates the ability of the filter and pooling nodes in DCNNs to depict the structure and organization of simple and complex cells in the mammalian brain, since the wiring of the typical DCNN was inspired by the aspect of the Hubel & Wiesel story that has been shown to be incorrect.

Second, the ability of DCNNs to model perceptual neocortex must be further tempered by the existence of adversarial examples that have demonstrated resilience in their ability to cause incorrect (and highly confident) categorization decisions in DCNNs, but that do not fool humans. An adversarial example in this sense is created by inserting a relatively small amount of noise into an image that was originally classified (with high confidence) as an instance of one category, and later becomes classified (with high confidence) as an instance of another category, despite the images appearing almost indistinguishable to humans (Fig. 16). Excepting adversarials, DCNNs have been touted for their ability to predict many aspects of human category learning and perceptual similarity judgments: they produce similar rank-orders for how difficult a category is to learn and when sorting typical exemplars by mutual similarity, and they even display the same systemic learning biases as human children, such as preferring to categorize objects by shape rather than color (Ritter, Barrett, Santoro, & Botvinick, 2017). There is also the high degree of correlation between features recoverable at different layers of ventral stream processing and of deep layers of a network (Yamins & DiCarlo, 2016). Adversarials, however, show at least one striking form of dissimilarity in the behavior of DCNNs and human perception. Moreover, subsequent research has shown that rather than being an idiosyncratic and easily-eliminated quirk, adversarials are counterintuitively robust: they can be built without god's-eye access to model parameters, they transfer (with labels) to DCNNs with very different architectures and training sets (and in fact to entirely different machine learning techniques), and they cannot be eliminated with simple regularization techniques (Goodfellow, Shlens, & Szegedy, 2014). Though there is still intense debate over their significance—whether they are irrelevant from a modeling perspective because they would not occur naturally, or whether humans might also be vulnerable to certain adversarials (Elsayed et al., 2018)—they show at the very least that at the most specific grain of detail, DCNNs and human perceptual cortex do not produce exactly the same phenomena.

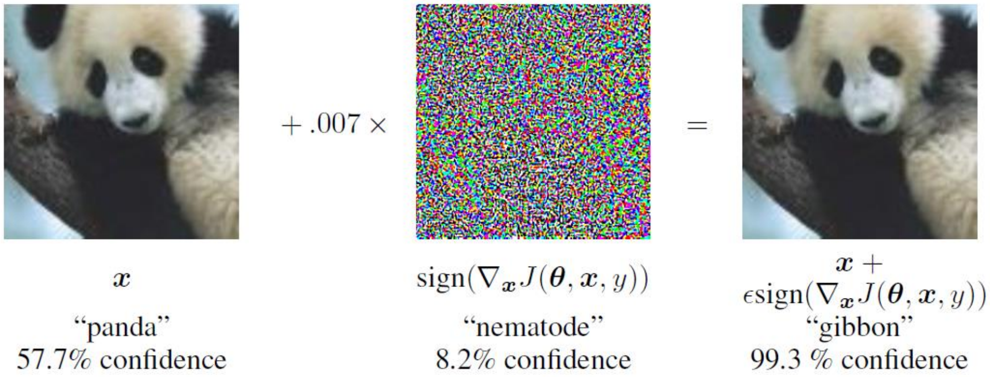
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 16. Adversarial example, reproduced from Goodfellow et al. (2015).

So, are DCNNs hopeless as a source of mechanistic explanation for transformational abstraction in mammalian neocortex? It may be helpful here to remember some truisms about modeling in philosophy of science. Models never seek to reproduce their target phenomena in full detail (this would be better termed "simulation"), and mechanistic models specifically never seek to replicate every structural feature of the systems they depict. Like other forms of explanation in science, modelers aim for simplicity, ceteris paribus; they aim to depict only structural aspects of systems that are relevant to the system's ability to produce the phenomenon being explained. Phenomena themselves can be characterized with different grains of detail, and the ability of any explanation to generalize across a wider range of phenomena will depend upon a tradeoff between this grain of detail and the degree of underlying structural similarity in the systems being explained. Some mechanists with more normative ambitions will automatically score models that explain a wider range of phenomena as possessing more explanatory power (Boyd, 1999; Buckner, 2011; Stinson, 2016; Ylikoski & Kuorikoski, 2010), whereas others of more naturalistic stripe leave such matters entirely to the preference of the modelers (Craver & Kaplan, 2018). All agree, however, that reproducing irrelevant details of a system is more simulation than explanation, and that all explanations should search for a perspicuous grain of abstraction at which to depict the relevant aspects of the system's structure that make a difference to producing the target phenomenon of interest.

With these concerns in mind, there remain two plausible options to conserve the mechanistic explanatory relevance of DCNNs to transformational abstraction in human neocortex. Both require us to claim that the phenomenon we are interested in explaining is not human perceptual similarity and categorization judgments at the highest degree of resolution—which, after all, would show poor ability to generalize across other mammalian species, from individual to individual, or even the same individual at different times. The explanation would broadly address how mammals separate the manifolds of different categories using transformational abstraction, and would include rank-order judgments, systematic biases, and feature recoverability using linear discrimination methods at different levels of a hierarchy, but could not explain these judgments in the case of unnatural images or for those judgments at the highest degree of resolution. The failure to predict these aspects of perceptual categorization, the DCNN modeler might say, are as irrelevant to the phenomenon of interest as the failure of classical models of grammatical competence to predict increased grammatical errors in a low-oxygen environment. In some sense, the components of the mechanism which produces grammatical competence depend for their functioning upon the presence of oxygen, but this particular what-if-things-had-been-different question exceeds the boundary conditions of our interest. Even if adversarial examples are pragmatically significant because we must protect against hackers, they may be similarly insignificant when our goal is only to explain transformational abstraction at a grain of detail broad enough to cover a wide swath of mammals, and limited to natural perceptual situations.

Two approaches from mechanistic explanation in psychology are useful in completing this picture. First, we could say that the DCNNs are "mechanism sketches" of human perceptual neocortex (Kaplan & Craver, 2011; Piccinini & Craver, 2011), specifically where the synaptic connections of simple and complex cells are black-boxed rather than depicted with structural fidelity. This would be an admission that we are not with DCNNs explaining the functional selectivity of simple and complex cells by highlighting a structural isomorphism with the connectivity patterns of filter and max-pooling nodes in a DCNN; but it is a misunderstanding of mechanistic explanations to suppose that they must always aim for this kind of maximal completeness irrespective of other factors (Craver & Kaplan, 2018). This idealization about connectivity could be done away with if needed, but if our explanandum covers only the perceptual similarity and categorization judgments of whole agents, doing so may complicate the model's implementation without explanatory payoff.

An even more promising route, however, is provided by Catherine Stinson's notion of a "generic mechanism", developed in part to make sense of connectionist explanations in psychology and neuroscience. Stinson notes that connectionists typically aim both to reproduce important patterns in the target phenomenon and to do so using biologically plausible constraints—"as though they are deploying an inferential pincer movement" to narrow the space of possible explanations from two directions at once (Stinson, 2018). The pincer movement aims to locate an optimal balance in the balance between generality and inductive unity. The goal of such modeling is never to reproduce subject's behavior at the highest grain of resolution or to achieve a correspondence with actual brains down to the level of individual neurons and their connections, but rather to reveal the brain's "basic principles of processing" (McClelland, 1988, p. 107) by exploring "computational constraints that govern the design of the nervous system" (Sejnowski, Koch, & Churchland, 1988, p. 1300). In the present case, the sweet spot for a generic mechanism usefully describable by DCNNs is independently provided by what DiCarlo et al. dub the family of "linear-non-linear encoding models" (LN), which untangle category manifolds by alternating linear operations (which achieve feature sensitivity) with nonlinear operations (which achieve tolerance for variance). As DiCarlo et al put it, "while LN-Style models are far from a synaptic-level model of a cortical circuit, they are a potentially powerful level of abstraction in that they can account for a substantial amount of single-neuron response patterns in early visual, somatosensory, and auditory cortical areas" (DiCarlo et al. 2012, p. 415). While it is still possible that mammalian neocortex does not instantiate even the generic LN-model structure, this seems unlikely given current empirical evidence, and the hypothesis is not seriously challenged by the discovery of unimodal connectivity distribution or the existence of adversarials.[13]

7. Conclusion

This paper argued that the key to the varied successes of DCNNs lies in their capacity for transformational abstraction—that is, their ability to learn from experience to hierarchically transform perceptual input into increasingly abstract representational formats that are more tolerant to nuisance variation. This account both makes contact with previous ideas in artificial intelligence while also grounding the interpretation of DCNNs in traditional empiricist philosophy of mind. The linkage in turn suggests solutions to classical problems that caused long-standing disputes between different strands of empiricism, especially between Locke's account of abstract ideas and Berkeley & Hume's emphasis on exemplars in abstract reasoning. DCNNs point the way toward an ecumenical resolution to this dispute that allows bi-directional travel between exemplars and abstractions. Specifically, DCNNs show that the generic kind of neural mechanism found in mammalian perceptual cortex can learn and deploy abstract category representations using only domain-general learning mechanisms—vindicating a key theme of empiricism. Thus, contrary to recent critiques, DCNNs provide more than a mere brute-force augmentation of three-layer perceptrons, and offer a promising, multidisciplinary future ripe with philosophical and empirical significance.

References

Antonelli, G. A. (2010). Notions of invariance for abstraction principles. *Philosophia Mathematica*, *18*(3), 276–292.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

[13] A likelier critical outcome is that both DCNNs and mammalian neocortex are members of the LN generic mechanism family, but there are other members in this family besides DCNNs that provide a tighter fit in performance and structure to humans. For example, while a more recent study by DiCarlo and co-authors confirmed that DCNNs predict many low-resolution patterns in human perceptual similarity judgments and do so using the same sorts of features that are found in late-stage ventral stream processing in V4/5 and IT, they found that these models were not as predictive of high-resolution, image-by-image comparisons in humans as were rhesus monkeys (Rajalingham et al., 2018). They speculate that an alternative but nearby subfamily of models that tweaks one or more typical features of DCNNs—i.e. their diet of training on static images, or lack of recurrent connections between layers—might provide an even better mechanistic model of human perceptual similarity and categorization judgments without unduly complicating the model. However, whether this prospect will pay off—and do so without inhibiting the ability of DCNNs to generalize to non-primate species—remains an open empirical question, and DCNNs remain the most successful mechanistic model of primate visual perception that we have to date.

Berkeley, G. (1710). *A treatise concerning the principles of human knowledge*. RS Bear. Retrieved from https://scholarsbank.uoregon.edu/xmlui/handle/1794/653

Beth, E. W. (1957). Uber Lockes "Allgemeines Dreieck." *Kant-Studien*, *1*(48).

Boyd, R. (1999). Kinds, complexity and multiple realization. *Philosophical Studies*, *95*(1-2), 67–98.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1-3), 139–159.

Buckner, C. (2011). Two approaches to the distinction between cognition and "mere association." *International Journal of Comparative Psychology*, *24*(4).

Buckner, C., & Garson, J. (2018). Connectionism: Roots, Revolution, and Radiation. In M. Sprevak & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind*.

Camp, E. (2015). Logical concepts and associative characterizations. *The Conceptual Mind: New Directions in the Study of Concepts*, 591–621.

Chatterjee, A. (2010). Disembodying cognition. *Language and Cognition*, *2*(1), 79–116.

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. MIT press.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* (Vol. 6). MIT Press. Retrieved from http://uclibs.org/PID/8692

Craver, C., & Kaplan, D. M. (2018). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *The British Journal for the Philosophy of Science*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

DeMers, D., & Cottrell, G. W. (1993). Non-linear dimensionality reduction. In *Advances in neural information processing systems* (pp. 580–587). Retrieved from http://papers.nips.cc/paper/619-non-linear-dimensionality-reduction.pdf

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dosovitskiy, A., Springenberg, J. T., & Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1538–1546). https://doi.org/10.1109/CVPR.2015.7298761

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial Examples that Fool both Human and Computer Vision. *arXiv Preprint arXiv:1802.08195*.

Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron. *IEICE Technical Report, A*, *62*(10), 658–665.

Fukushima, K. (2003). Neocognitron for handwritten digit recognition. *Neurocomputing*, *51*, 161–180.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423). Retrieved from http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html

Gauker, C. (2011). *Words and images: An essay on the origin of ideas*. OUP Oxford.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(S3), S342–S353.

Gobet, F., & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychological Science*, *7*(1), 52–55.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Book in preparation for MIT Press. *URL!` Http://www. Deeplearningbook. Org*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv Preprint arXiv:1412.6572*.

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*(6789), 947.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen*. Technische Universität München, München.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Hume, D. (1739). *A Treatise on Human Nature*. Oxford University Press.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). https://doi.org/10.1371/journal.pcbi.1003915

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, 1–101. https://doi.org/10.1017/S0140525X16001837

Laurence, S., & Margolis, E. (2012). Abstraction and the Origin of General Ideas. *Philosopher's Imprint*, *12*(19). Retrieved from http://hdl.handle.net/2027/spo.3521354.0012.019

Laurence, S., & Margolis, E. (2015). Concept Nativism and Neural Plasticity. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (pp. 117–147). MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396–404).

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Machery, E. (2009). Doing without concepts. Retrieved from http://books.google.com/books?hl=en&amp;lr=lang_en&amp;id=smC3a_I8ZPAC&amp;oi=fnd&amp;pg=PR11&amp;dq=DOING+WITHOUT+CONCEPTS&amp;ots=9Mjsoogc9q&amp;sig=40MWFY-KiMx8BQoT9AmIl36Gtms

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv:1801.00631 [cs, Stat]*. Retrieved from http://arxiv.org/abs/1801.00631

McClelland, J. L. (1988). Connectionist models and psychological evidence. *Journal of Memory and Language*, *27*(2), 107–123.

Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems* (pp. 2924–2932).

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

Patel, A. B., Nguyen, M. T., & Baraniuk, R. (2016). A probabilistic framework for deep learning. In *Advances in Neural Information Processing Systems* (pp. 2558–2566). Retrieved from http://papers.nips.cc/paper/6231-a-probabilistic-framework-for-deep-learning

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Priebe, N. J., Mechler, F., Carandini, M., & Ferster, D. (2004). The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, *7*(10), 1113.

Quine, W. V. (1971). Epistemology naturalized. *Akten Des XIV. Internationalen Kongresses Für Philosophie*, *6*, 87–103.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 240614.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv Preprint arXiv:1706.08606*.

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

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

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science*, *241*(4871), 1299–1306.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Bolton, A. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Spasojević, S. S., Šušić, M. Z., & Djurović, Ž. M. (2012). Recognition and classification of geometric shapes using neural networks. In *Neural Network Applications in Electrical Engineering (NEUREL), 2012 11th Symposium on* (pp. 71–76). IEEE.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv Preprint arXiv:1412.6806*. Retrieved from https://arxiv.org/abs/1412.6806

Stinson, C. (2016). Mechanisms in psychology: ripping nature at its seams. *Synthese*, *193*(5), 1585–1614.

Stinson, C. (2018). Explanation and Connectionist Models. In M. Colombo & M. Sprevak (Eds.), *The Routledge Handbook of the Computational Mind*. New York, NY: Routledge.

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

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201–219.