Deep Learning: Philosophical Issues

1. Introduction: A wakeup call

Deep learning is currently the most powerful and high-profile approach to Artificial Intelligence (AI). Major technology companies like Google, Microsoft, Facebook, and Amazon have all devoted marquee research & development groups to the study of the technology, and such systems currently control operations as diverse as image search and labeling, speech and face recognition, natural language translation, strategy gameplay, and (semi-)autonomous driving. These feats overcome a variety of benchmarks that skeptics supposed would remain beyond the reach of AI: they can identify complex objects in natural photographs at human-level performance, and they have defeated world champions in games as complex as chess and Go. In addition to their practical achievements, deep learning models are currently regarded as the best models of perceptual similarity judgments in primates. Influential figures in deep learning research—whose achievements are featured monthly in prestige journals like *Science* and *Nature*, and have been covered heavily in the scientific press since at least 2010—have offered ambitious claims about the philosophical and social significance of these achievements, such as that they vindicate classical empiricism about reasoning, or that vast swathes of the human labor force will soon be rendered obsolete. Yet there remains significant uncertainty in the technical literature as to why these networks perform so well, what limits we should expect to discover in their performance, and which ethical constraints we ought to impose upon their use.

It is therefore astounding that, as of 2018, there has been virtually no discussion of deep learning in mainstream philosophy journals. Whereas philosophers have long played a central role in the development of artificial intelligence generally and neural network research specifically, the profession has largely ignored AI's most effective and widespread technology for nearly a decade. When philosophers speak of the method at all, they have tended to recycle discussions of neural networks from the 1980s and 1990s without engaging with any of the specific advances that have allowed deep learning models to vastly outperform their earlier forebears. This must change; this article is offered as a clarion call and opinionated guidebook to this powerful and bewildering new area of AI research.

I begin Section 2 by summarizing the main features that distinguish the most common and reliably successful deep architecture, deep convolutional neural networks (hereafter DCNNs). In Section 3, I consider the scientific and philosophical interpretations of these networks. In Section 4, I provide a list of 10 questions for future philosophical research into the significance and implications of deep learning. Each section aims to briefly introduce the issues in a way to encourage reflection and participation by philosophers in future research.

2. DCNNs: Main features

I now illustrate the main features of the most common and reliably successful deep architecture, DCNNs, by contrasting them with their earlier forebears from the "Golden Age" of network research from 1980-1995 (in the interests of space we will completely forego history, including stock criticisms of earlier neural networks, adequately covered elsewhere—Buckner & Garson, 2018; Schmidhuber, 2015). All neural networks can be thought of as composed of nodes and links, intended to model the behavior of neurons and synapses at some level of abstraction. Processing is performed by passing an input signal to some array of input nodes, which then activate according to their functions and pass an output activation signal up to the next layer in the network along their links, modified by those links' "weights" (which might produce inhibition, should the resulting value be negative). Activation propagates forward in this manner until it reaches a designated output layer of nodes, which is then decoded and taken as the network's "decision" on that input.

An exciting feature of these networks is their ability to discover novel solutions directly from problem data. The most popular learning algorithm since the 1980s has been error backpropagation learning. This method is called "supervised," because it deploys a teaching signal generated by an error function to calculate the distance between the actual and desired output for that exemplar, determined by a training set labeled with the correct answers. For multi-layer networks, that error signal is then further backpropagated through the next previous layer of the network and used to adjust its input link weights, and so on until the error signal reaches the initial layer. Simply by backpropogating error signals and gradually adjusting link weights in this manner, network performance can converge on the solutions to a wide range of classification and decision problems.

These basic features are shared by most Golden Age and state-of-the-art DCNN models. Now for the contrasts. Let us characterize the typical Golden Age network by three properties; they are: 1) shallow, with no more than three or four layers between input and output; 2) uniform, with only one type of node deploying a sigmoidal activation function; and 3) fully-connected, with each node from a lower layer connected to each other node in the next layer up. The typical state-of-the-art DCNN, on the other hand, is: 1) deep, containing anywhere from 5 to 250 (or more) layers; 2) heterogeneous, containing different kinds of processing node deploying different activation functions, especially convolutional nodes, rectified linear units (ReLU), and non-linear downsamplers (usually max-poolers); 3) sparsely-connected, with later layers only taking input from spatially nearby nodes from the previous layer (a fully-connected output layer is the common exception); and 4) deploys explicit regularization techniques to avoid overfitting, such as dropout. Each of these features will be elaborated below, as each likely plays a role in making DCNNs orders of magnitude more efficient on the kind of difficult perceptual discrimination problem at which they reliably succeed.

A. Depth

The most obvious distinguishing mark of DCNNs is their depth. In the 80s, the transition from two-layer perceptrons to three-layer networks with a "hidden layer" was computationally profound; it allowed neural networks to compute linearly inseparable functions such as XOR that had been shown to be beyond the reach of two-layer perceptrons (Minsky & Papert, 1969). The advantage of further depth is not quite as obvious as allowing them to compute an entirely new kind of function, but it does afford exponential reductions in the computational complexity of learning and inference, because it allows networks to iteratively transform and simplify input signals as they move through the network's layers. Another way to put the point has become more popular: any computation performed by a node at an earlier layer can be recursively re-used by later layers exponentially many times in terms of the network's depth, as each later layer can compose increasingly complex functions from the simpler building blocks of the previous layer (Montufar, Pascanu, Cho, & Bengio, 2014). As a result, on problems which benefit from hierarchical processing, deep networks can be exponentially more efficient than shallower networks with the same number of nodes (e.g. up to 60,000 times or more in Big-O notation on visual discrimination—see Bengio, Courville, & Goodfellow, 2016 for a worked example). Though some may scoff at these "mere efficiency gains", they would be decisive in evolutionary or behavioral competitions involving limited time and resources.

B. Heterogeneity

State-of-the-art DCNNs use at least three different types of nodes: convolutional nodes, ReLUs, and max-poolers. These different kinds of nodes are often layered in a series, with a convolutional node passing input to a ReLU, and then with a max-pooler taking input from several convolutional/ReLU combinations with overlapping spatial or temporal sensitivity. Understanding the cooperation between these different activation functions in series is crucial to understanding the computational power of DCNNs, so I elaborate each in turn (for a longer discussion, see Buckner, 2018).

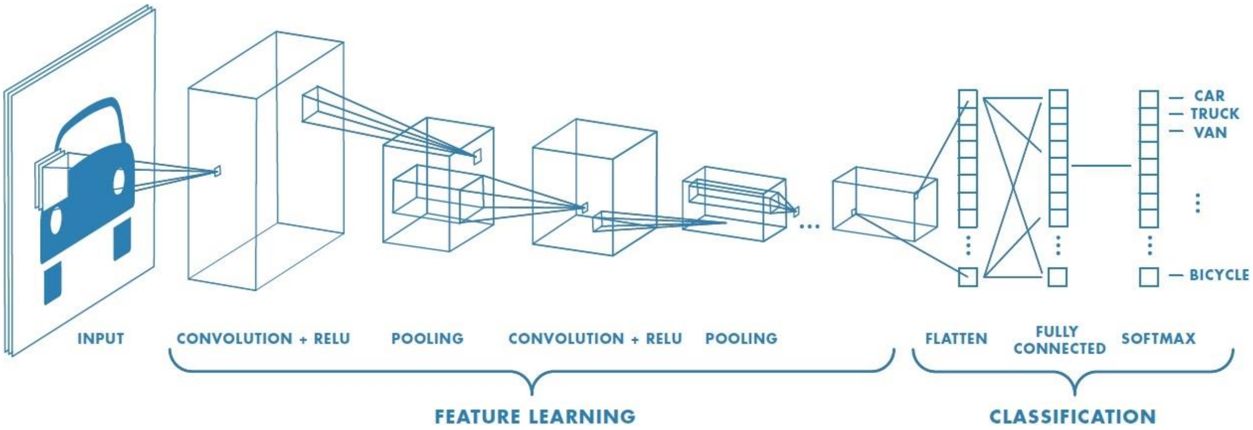

Figure 1. The generic processing flow of a Deep Convolutional Neural Network, whose architecture involves transforming input signals through many sequences of locally-connected convolutional, ReLU, and pooling nodes, before finally passing to a fully connected classification layer that assigns final category labels. Image credit: https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html .

Convolution is a linear algebra operation that transforms some chunk of the network's input to amplify certain values and minimize others. In DCNNs, it is typically applied to a "window" of perceptual input data, such as a rectangle of pixels in an image or a snippet of audio information in a sound file (for ease of exposition, I will limit discussion to images in what follows). Pixels are themselves usually vectors of RGB color channel information at that point; the convolution operation returns a transformed matrix of RGB vectors that amplifies the presence of a certain feature, such as (at early layers) contrasts or shadings. These convolutional units are called "filters" or "kernels". The output of each convolution operation is then typically passed to a ReLU unit, which activates according to a simple function called rectification if the result of the convolution exceeds a preset threshold. This is sometimes called the "detector" stage of processing, as activation is only passed up the hierarchy if the kernel's activation indicates that it found the feature at that location. To illustrate by example, suppose we had a kernel that amplified vertical lines; if the "window" of this convolution operation were tiled over the whole image and each passed to a different ReLU node, one chunk at a time, it would "filter" the raw input and return a transformed image that showed all and only the image's vertical lines.

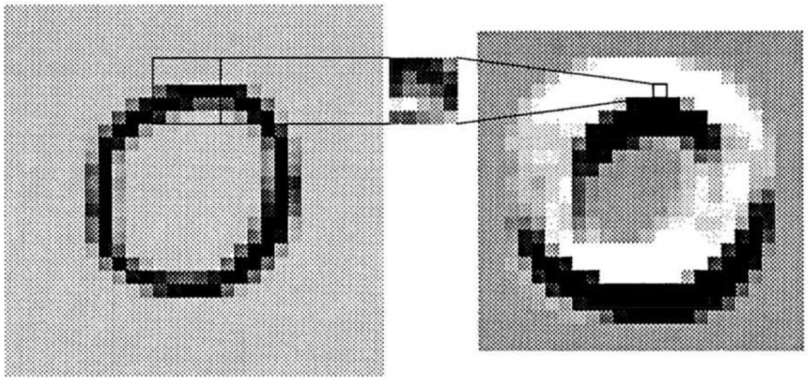

Figure 2. An example of the visual output of a single convolutional kernel on real handwritten digit data in a trained convolutional network (reproduced from LeCun et al., 1990, 399). This (learned) kernel detects something like curves at the top or bottom of a digit.

Much of the distinctive power of DCNNs, however, is to be found in their ability to detect features in a variety of different poses. In other words, we typically do not just want to detect vertical lines, but rather lines in any orientation. This can be achieved by passing each filter+ReLU output to a third kind of node whose function is to aggregate and downsample the activity of several different filters with overlapping spatial receptivity. The most popular downsampling function in current DCNNs is max-pooling (also called "max-out"), which sends up activation only from its most highly-activated input. In other words, if a max-pooling unit receives input from a vertical line kernel and a horizontal line kernel at a particular location, then it will only pass activation to the next layer for whichever of the two was most strongly activated. Combining all three operations, we could produce a simplified, transformed representation of the source image which consisted not only of vertical lines in a particular spot, but of all lines in the original image in any location or orientation. These abstracted features then become available for processing at the next layer of the network, which performs a similar series of operations to detect yet more complex features. For example, the next layer of convolutional units could then build filters to detect angles from the transformed and simplified information about lines and their locations.

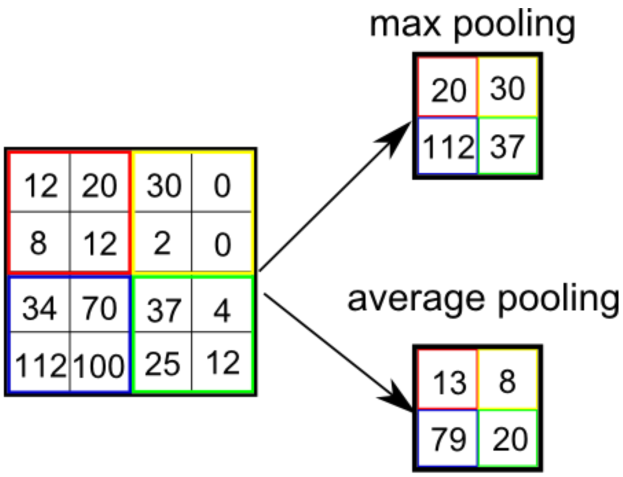

Figure 3. A comparison of max-pooling with average-pooling for aggregation across activation received from the same receptive fields (reproduced from Singhal, 2017).

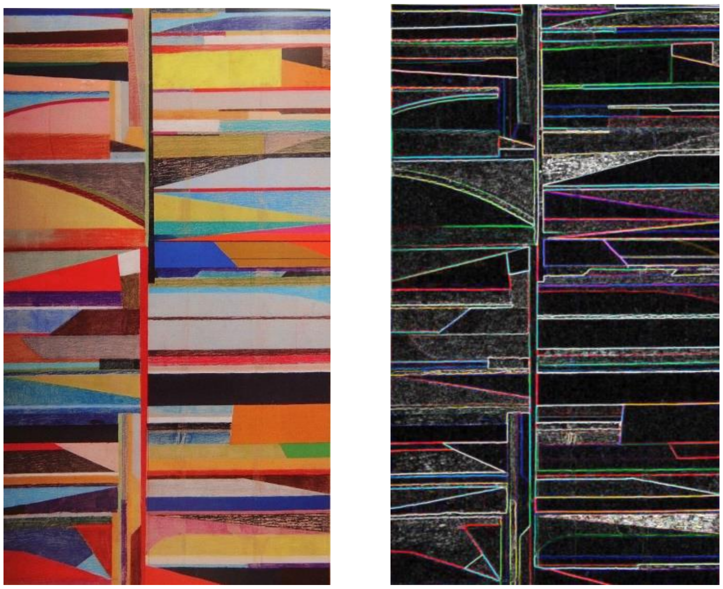

Figure 4. *21st Century* by Viktor Romanov, with and without a Sobel edge-detection filter, one popular convolutional edge-detection algorithm. (Operation performed by author in Gimp 2.0 Freeware image editing program. Image credit: Yugra News Publishing, Wikimedia Commons CC License.)

C. Sparse connectivity

A third characteristic feature of DCNNs is their sparse connectivity between layers. Whereas nodes in Golden Age networks were often fully connected to each node in the layers above and below, nodes in DCNNs are usually only locally connected to nodes with spatially or temporally nearby input receptivities, as if retinotopically (the only exception is a fully-connected layer that typically precedes the final categorization). This sparse connectivity greatly reduces the number of parameters that need to be learned, relative to a fully-connected network with the same number of nodes; it also makes computation more efficient when a trained network is asked to classify a test image, because the activation functions which need to be computed have far fewer inputs.

D. Regularization

All machine learning seeks a tradeoff between underfitting and overfitting training data; training data is underfit if the model does not learn enough structure from the training exemplars to predict its category labels, and overfit if it learns so much idiosyncratic structure from the training data that it fails to generalize to novel exemplars outside the training set. Overfitting is a concern with DCNNs, as analyses have shown that deep networks possess enough storage capacity to simply memorize mappings between exemplars and labels for even very large training corpora, even when exemplars are assigned random labels or consist entirely of random noise (Zhang, Bengio, Hardt, Recht, & Vinyals, 2016).

To avoid overfitting, modelers deploy a variety of explicit regularization techniques. One simple method involves slightly modifying the training exemplars, such as by adding noise or shifting or rotating images. This forces the network to learn that categorizations must be robust to small changes in input details. Another popular method is dropout, which causes some nodes in the network to periodically become inactive, forcing the network's categorizations to not depend too much on any particular regularity (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). A third method modifies the error function to favor simpler or sparser solutions; "ℓ1" regularization, for example, adds a penalty term that (roughly) causes link weights to fall back to zero if not maintained by a large gradient during learning (Krizhevsky, Sutskever, & Hinton, 2012). As a result, the performance of the network is biased against learning many precise details, and feature representations tend to become more localized (that is, involving fewer nodes and links) and focused on more generalizable properties. Finally, there are also "implicit" regularization techniques, such as early stopping, which halts learning according to some measure of when generalization improvement starts to plateau and begins to overfit (Zhang et al., 2016). Recent empirical analyses have suggested that explicit and implicit regularization are requirements for effective DCNNs (Achille & Soatto, 2018).

3. Interpretation and explanation

With these basic features of DCNNs in place, we can now ask: why do these networks tend to work so well? And what, if anything, should they be thought to model in the mind or brain? Let us consider three popular accounts of DCNN function. Though they are sometimes offered as competitors, it is not clear that they are in conflict, and may illuminate different or complementary aspects of the same underlying problem.

A. Hierarchical feature composition

The most traditional explanation of the effectiveness of DCNNs is that they work like visual processing in the mammalian ventral stream, by hierarchically composing more complex features from simpler and less abstract ones. Early DCNNs were inspired by Hubel & Wiesel's discovery that early ventral stream neurons seemed to be sensitive to very specific and local features like shadings and contrasts in V1, which was later enhanced with a variety of imaging methods to suggest a whole processing cascade, from lines and borders in V5, to angles and colors in TEO/PIT (posterior inferotemporal), to figures and objects in TE/AIT (anterior inferotemporal). In this respect DCNNs and ventral stream processing compare favorably, with both appearing to recover and compose increasingly abstract features as one moves up the hierarchy of processing (Yamins & DiCarlo, 2016).

B. Systematic transformations of input to adjust for nuisance variation

Another interpretation of DCNNs holds their characteristic features to implement a set of infinitely-strong, domain-general prior probability estimations that help networks control for systematic variation in input. Consonant with the previous subsection, depth enforces the assumption that complex features are built from simpler ones. Passing the windows of kernels over the whole image at each layer applies the assumption that features can be re-used many times in the composition of more complex ones, and that features can occur anywhere in the image. Max-poolers render the precise pose or location of a feature irrelevant to an exemplar's categorization. Sparse connectivity and regularization impose the assumption that generalizable features should not depend on too many long-distance relations or subtle contextual details. During training, DCNNs simply ignore hypotheses that violate these assumptions; but when the solution to a problem satisfies them, DCNNs can find that solution exponentially more efficiently than more thorough explorations of the hypothesis space.

Thus, perhaps DCNNs work so well because a wide class of classification and decision problems satisfy these assumptions. In particular, machine learning researchers have noted that many visual and auditory tasks are characterized by "nuisance factors", repeatable sources of variation that are not diagnostic of decision success. For visual classification tasks, common nuisances include size, pose, location, and rotation[1], or for auditory tasks, pitch, tone, pronunciation, and duration. Decision procedures that succeed on these tasks must learn to systematically adjust for these sources of variance. More cognitive or amodal tasks can fit these assumptions as well; Go strategy, for example, should also be tolerant to small changes in the position or rotation of stone placement patterns, whether board patterns are perceived visually, auditorially, or inputted through amodal symbols. How far the utility of these priors extends into territory traditionally ascribed to higher or amodal cognition remains unknown.

C. Number of linear regions

[1] More precisely, DCNNs are efficient at overcoming nuisance parameters with a group-like structure. Non-group-like sources of variation include occlusion or changes in illumination, on which DCNNs are no more efficient than alternative neural network architectures.

A third, related account of deep learning functionality that is becoming increasingly popular emphasizes the number of linear regions they can map in a problem's input space, relative to networks which do not possess their characteristic features. This idea requires some elaboration; a linear region is a piece of a decision function implemented by a neural network that is linear, i.e. where its slope is constant. The goal of training a neural network for classification is to discover a global function composed of individual nodes' activation functions and associated link weights that can draw boundaries between the exemplars of categories that need to be discriminated. The ability to learn more linear regions in this function is advantageous because it allows neural networks to draw more complex boundaries between categories with subtle hierarchical structure.

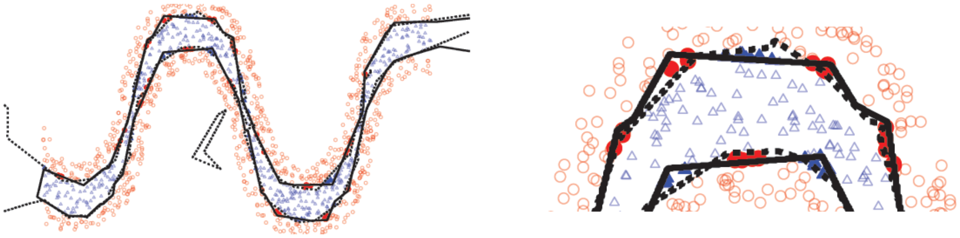

Figure 5. From (Montufar et al., 2014) – function borders (zoomed in on right) from a shallow (with 20 hidden units, solid line) or a deep network (two hidden layers of 10 units each, dashed line). Filled markers are errors made by the shallow but not the deep model.

A well-known analysis by Montufar et al. (2014) uses a paper-folding metaphor to explain how the discrimination functions of deep neural networks can have exponentially more linear regions than shallower networks with the same number of nodes. The key is that by finding symmetries in the input and "folding" the input space to align them, these networks can re-use linear regions exponentially many times in terms of the network's depth. Deep neural networks thus effectively fold input space recursively starting with the first layer of the network, with each successive fold multiplying the number of times that a previously-computed linear region can be re-used in drawing fine discriminations between categories. The learning process thus identifies which folds of the input space exploit symmetries to produce the lowest loss functions, while at the same time exponentially reducing the computational complexity of the functions that must be computed by later layers. This explanation for DCNNs' success is not necessarily incompatible with the first two—hierarchical feature composition and nuisance adjustment may just be two members of a class of tasks that benefit from recursive reuse of linear regions—but there may be others, giving this third explanation potentially broader applicability.

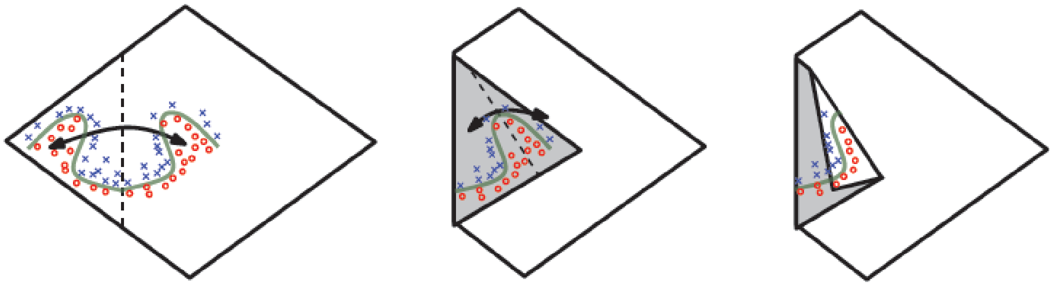

Figure 6. A visual representation of paper folding metaphor demonstrating how depth allows linear regions to be re-used (from Montufar et al. 2014).

4. Open questions, future research directions, and live debates

Much ink has been spilt prognosticating extreme future trajectories for DCNN research: either that they have been dramatically oversold and we are about to enter another "AI winter", or that they will soon usher us into a Singularity of exponentially-increasing levels of intelligence. Both of these extreme predictions are probably exaggerated, however, wrapped-up as they are with a nebulous and necessarily speculative question about future discoveries: Could some unknown modification of deep learning methods produce general intelligence? Even if the answer to this question is negative, however, I have recently argued that current DCNNs already model a crucially important component of general intelligence—a particular kind of abstraction—that until now eluded our grasp (Buckner, 2018). Many pressing and more answerable questions still concern the proper interpretation of established DCNN methods, and the epistemic, social, and ethical questions we will face as they become an even more pervasive presence in our lives. I end by canvassing ten that could be profitably explored by different sub-areas of philosophy in the near future.

A. Philosophy of Mind: DCNNs and empiricism

DCNN modelers have often made ambitious claims about the relevance of their research to old debates between empiricism and nativism in philosophy of mind, sometimes citing the work of classical empiricists like John Locke or David Hume (Silver et al., 2017). If this ancient debate is construed as a debate about the origins of categorical knowledge, then DCNNs are obviously relevant; they provide at least a proof of concept that high-level category information and strategy can be derived from low-level sensory experience; and if DCNNs model perceptual cortex, then they show how this abstract knowledge might be discovered in the mammalian brain (Buckner 2018). However, these claims merit more scrutiny from those familiar with the philosophical subtleties of this debate.

B. Philosophy of Cognitive Science: Reinforcement learning and embodiment

Critics have worried that backpropagation learning relies on very large, labeled training sets (e.g. ImageNet), which limits their applicability to problems where such data sets are forthcoming (Lake, Ullman, Tenenbaum, & Gershman, 2016; Marcus, 2018). Most recently, however, systems such as DeepMind's AlphaGo, AlphaZero, and the Atari-game-playing DQN system learn instead using the much more biologically-plausible method of unsupervised reinforcement learning (Mnih et al., 2015; Silver et al., 2018). However, reinforcement learning algorithms require a valuation function that tells the system which outcomes are desirable and undesirable, and so far success has been limited to domains that conveniently provide an objective, denumerable utility currency such as game score. Further gains will require modelers to reflect on the ways that evolution has subtly tweaked the body and brain to provide an array of rich, multi-dimensional valuation signals, connecting to recent work philosophical work on the importance of embodiment and embeddedness to cognitive performance (Shapiro, 2010).

C. Philosophy of Cognitive Science: Episodic Memory Buffers and Episodic Replay

Another of most promising recent extensions of DCNN performance involved the use of episodic buffers to overcome the need for very large training sets (Blundell et al., 2016; Vinyals, Blundell, Lillicrap, Kavukcuoglu, & Wierstra, 2016). Mammalian brains are thought to store short term memories in the medial temporal lobes and repeatedly replay them over interleaved training sessions for a longer consolidation period of months or years. Network modelers have long supposed that this allows mammals to learn from a small number of focused learning episodes, as if they were re-experienced thousands of times. Indeed, DCNNs with episodic buffers have achieved expert-level human performance on Atari video games with much smaller, human-like levels of gameplay experience. This connects with recent work in philosophy of memory about the importance of episodic memory to future performance (Michaelian, 2016).

D. Philosophy of Perception: Attention

Another of the most effective recent improvements to DCNN performance was enabled by emulating attentional selection. Specifically, many recent computer vision DCNN models have achieved dramatic improvements in performance by adding secondary networks that direct the processing of a primary classification network to "region proposal boxes" where objects are deemed especially likely to appear, as occurs with eye saccades and attentional focus in animals (i.e. Ren, He, Girshick, & Sun, 2015). Philosophers of perception and consciousness may have much to add to this debate in providing specific ideas about how to properly model attention, and interpreting the results of the models that are implemented.

E. Philosophy of Perception: Adversarial Examples

One of the most perplexing discoveries arising from research on DCNNs is the phenomenon of "adversarial examples" (Szegedy et al., 2013). These are artificially-crafted exemplars that cause dramatic apparent misclassifications in DCNNs, but are supposed to not fool humans. A pessimistic interpretation holds that adversarial examples show that rather than learning a category the way humans do, DCNNs merely build "a Potemkin village that works well on naturally occurring data, but is exposed as fake when one visits points in space that do not have a high probability" (Goodfellow, Shlens, & Szegedy, 2015). Though they would occur only rarely by chance in nature, adversarial examples are of great practical interest because hackers and other malicious agents could use them to fool automated vision systems. Adversarial examples have many counterintuitive properties: they cannot be defeated with simple regularization techniques, they can transfer with labels to other DCNNs with different architectures and training sets, and other forms of machine learning are also vulnerable to them. More recent papers have suggested that a DCNN's take on adversarial examples is not so alien to human perception, either by producing adversarial examples that fool humans or by showing that humans can easily "adopt the machine perspective" and predict a DCNN's preferred labels for a range of adversarial examples (Zhou & Firestone, 2018). Despite intense research, adversarial examples and their implications remain mysterious, and would benefit from further philosophical reflection.

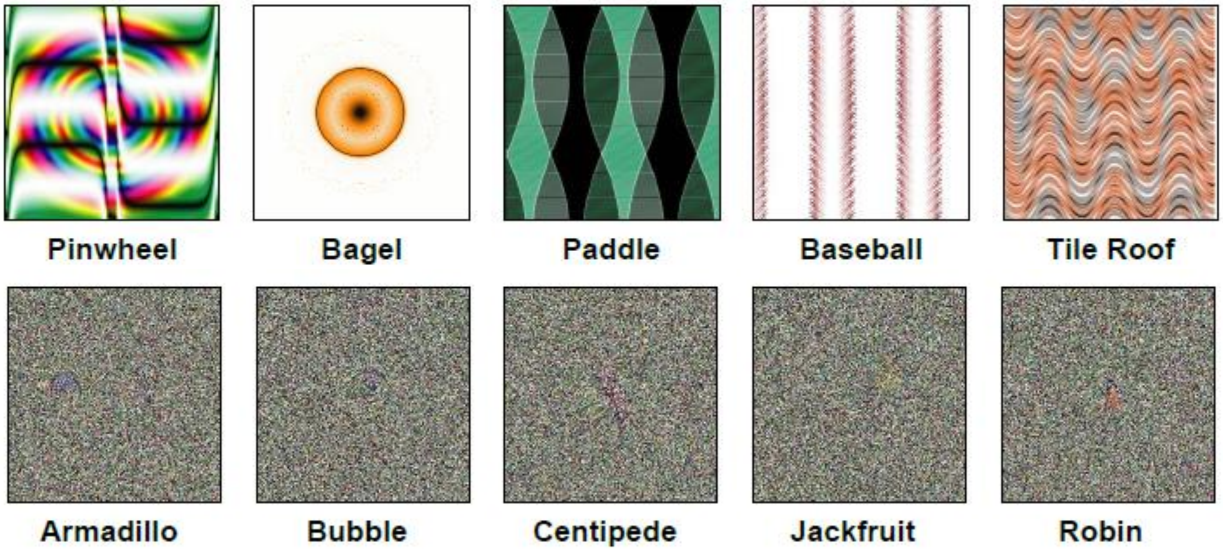

Figure 7. Examples of adversarial examples tested by Zhou & Firestone (2018) generated by two different methods, with preferred DCNN labels.

F. Philosophy of Science: Do DCNNs explain cortical processing; and if so, how?

In philosophy of science, a wave of "new mechanism" about explanation is cresting (Glennan, 2017). This movement holds that explanations in the life sciences work by locating structures and their organization in a target system whose coordinated operations regularly produce the phenomenon of interest. DCNNs, however, are pitched at a high level of abstraction from perceptual cortex, which could lead one to doubt that they succeed at providing mechanistic explanations for even perceptual processing (though see Boone & Piccinini, 2016; Stinson, 2016). Others have more recently advocated a role for functional, fictional, non-causal, or mathematical explanations in the life sciences (Weiskopf, 2011), and indeed in computational neuroscience more specifically (Chirimuuta, 2017). While computational neuroscientists have made extensive comparisons between DCNNs and perceptual cortex, there remain many unresolved questions as to whether and how they might explain the processing that occurs in perceptual cortex, and whether those explanations are mechanistic, functional, causal, mathematical, fictional, or dynamical in nature.

G. Philosophy of Science: Data analysis

Notably, scientists are one of the primary populations that are already relying upon deep neural networks and other machine learning methods for their daily tasks. FMRI interpretation, gene sequencing, protein folding simulations, cancer diagnosis, and many other sciences operating on highly complex systems depend upon

DCNNs for their findings. The opacity of these models raises a host of questions for the epistemology of science: Can such DCNNs explain scientific phenomena? Should scientists trust their results? When are such methods misused? Are there visualization or network analysis methods that we can apply to such models to render their processing more intelligible?

H. Epistemology and Action Theory: Explainable AI

Even more ambitiously, DARPA has recently issued its "Explainable AI" (XAI) challenge, which has focused especially on decisions produced by deep neural networks (Gunning, 2017). The goal of this initiative is to "enable human users to understand, appropriately trust, and effectively manage" systems driven by machine learning. These goals involve an admixture of what philosophers might distinguish as explanatory and justificatory concerns that pull in different directions (on the explanatory side, systems should answer queries "Why did you do that?", "Why not something else?", and on the justificatory side "When do you succeed?", "When can I trust you?"). We might also explore principles to fairly compare these systems' explanations to humans' rationalizations for their behavior, which psychology has revealed to be inferential in nature and not as reliable as we might have initially supposed (Carruthers, 2011).

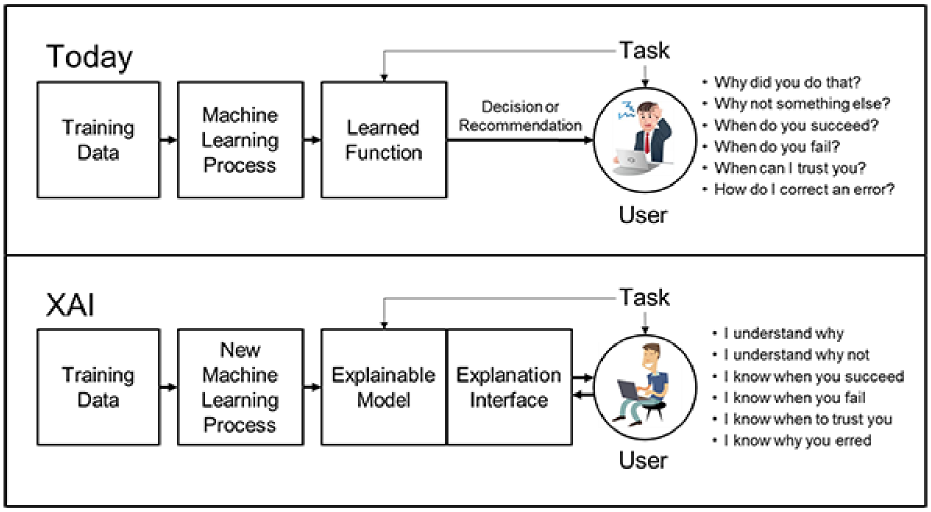

Figure 8. The DARPA XAI concept (from Gunning 2017).

I. Social and Political Philosophy: Automatization of human labor

As DCNNs (and other machine learning methods) succeed on a wider range of problems, they will inevitably cause greater disruption to human labor and education. Many jobs that have not yet been automated but rely on perceptual skill, motor expertise, or hierarchical pattern detection may now be replaced by DCNNs: diagnosticians, paralegals, drivers, actuaries, and so on. This will in turn have a ripple effect for business ethics and legal theory as we explore the managerial uses and legal consequences of automating all this work. We very badly need a new generation of social and political philosophers who are savvy in the technical details of neural networks, and can help determine which jobs could be replaced (Damian, Spengler, & Roberts, 2017)—and perhaps even more contentiously, which jobs should not be replaced for ethical or legal reasons.

J. Applied Ethics: Ethical and legal issues

Finally, one of the few sub-disciplines of philosophy that has already begun thinking seriously about the implications of deep learning is applied ethics (Lin, 2018). When a vehicle autonomously-driven by a DCNN-based-system hits a pedestrian, who is liable? When a system is confronted by a real-world "trolley-problem", and must choose between a course correction that will strike a single individual on the sidewalk vs. hitting several jaywalking children, what should it do? Can we trust such systems to make decisions involving lethal force, in situations of law enforcement or warfare? Given that training sets can reproduce the demographic biases of the cultures that produce them (Zou & Schiebinger, 2018), what countermeasures should be adopted to ensure that the decisions of systems that deploy deep learning are more equitable and just?

So ends this (abridged) list of proposed questions; let us waste no more time in attempting to answer them.

References

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, *19*(50), 1–34.

Bengio, Y., Courville, A., & Goodfellow, I. (2016). Deep learning. Book in preparation for MIT Press. *Current Version Available at Http://Www. Deeplearningbook. Org*.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., … Hassabis, D. (2016). Model-free episodic control. *ArXiv Preprint ArXiv:1606.04460*.

Boone, W., & Piccinini, G. (2016). Mechanistic abstraction. *Philosophy of Science*, *83*(5), 686–697.

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese*, *195*(12), 5339–5372.

Buckner, C., & Garson, J. (2018). Connectionism and post-connectionist models. In M. Sprevak & M. Columbo (Eds.), *The RoutledgeHandbook of the Computational Mind*.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. OUP Oxford.

Chirimuuta, M. (2017). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*.

Damian, R. I., Spengler, M., & Roberts, B. W. (2017). Whose job will be taken over by a computer? The role of personality in predicting job computerizability over the lifespan. *European Journal of Personality*, *31*(3), 291–310.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

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

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), Nd Web*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-w

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

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).

Lin, P. (2018). The moral gray space of AI decisions. *The Ethical Machine*. Retrieved from https://ai.shorensteincenter.org/ideas/2018/12/1/the-moral-gray-space-of-ai-decisions-6sc59

Marcus, G. (2018). Deep learning: A critical appraisal. *ArXiv Preprint ArXiv:1801.00631*.

Michaelian, K. (2016). *Mental time travel: Episodic memory and our knowledge of the personal past*. MIT Press.

Minsky, M., & Papert, S. (1969). Perceptrons.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

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).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

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

Shapiro, L. (2010). *Embodied cognition*. Routledge.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Graepel, T. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144.

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), 354–359.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

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

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6199

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, koray, & Wierstra, D. (2016). Matching Networks for One Shot Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3630–3638). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf

Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese*. Retrieved from http://www.springerlink.com/index/141166V581398375.pdf

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *ArXiv Preprint ArXiv:1611.03530*.

Zhou, Z., & Firestone, C. (2018). Taking a machine's perspective: Human deciphering of adversarial images. *ArXiv Preprint ArXiv:1809.04120*.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, *559*(7714), 324. https://doi.org/10.1038/d41586-018-05707-8