This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour

Cameron Buckner

Abstract

The last five years have seen a series of remarkable achievements in deep-neural-network-based artificial intelligence (AI) research, and some modellers have argued that their performance compares favourably to human cognition. Critics, however, have argued that processing in deep neural networks is unlike human cognition for four reasons: they are (i) data-hungry, (ii) brittle, and (iii) inscrutable black boxes that merely (iv) reward-hack rather than learn real solutions to problems. This article rebuts these criticisms by exposing comparative bias within them, in the process extracting some more general lessons that may also be useful for future debates.

1 Introduction

2 Four Popular Criticisms of Deep Learning Research

2.1 Deep learning is too data hungry

2.2 Adversarial examples expose deep learning as a fraud

2.3 Deep neural nets are not interpretable

2.4 Deep neural nets trained by reinforcement learn to 'reward hack' rather than solve problems

3 Purposes, Interests, and Fair Comparisons

4 A Crash Course in Comparative Bias

5 Four Rebuttals

5.1 Human learning involves more trainable exemplars than common sense supposes

5.2 Deep neural net's verdicts on adversarial examples may be correct

5.3 Human decision-making is also opaque

5.4 Humans are also notorious reward-hackers

6 General Lessons

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

1 Introduction

The last five years have seen a series of remarkable achievements in neural-network-based artificial intelligence (AI) research. For example, systems based on deep neural networks (DNNs) can now classify natural images as well as or better than humans, defeat human masters of strategy games as complex as chess, Go, or Starcraft II, navigate autonomous vehicles across thousands of miles of mixed terrain, and compose essays that are often indistinguishable from human writing. In the short history of AI, engineering breakthroughs have swung the pendulum in our theoretical approach to intelligence and rationality—from top-down tactics that emphasize structured representations, explicit, domain-specific knowledge, and rule-based problem solving (Newell and Simon [1976]), to bottom-up methods that locate intelligence in non-representational sensorimotor abilities and skilful coping (Brooks [1991]). The success of DNNs on the kinds of tasks touted by both extremes suggests a revival in the fortunes of connectionist approaches (McClelland *et al.* [1986]; Clark [1989], [2003]; Rogers and McClelland [2014]), a midway position that explains intelligence in terms of the ability of domain-general learning processes to acquire abstract representations of the environment from low-level perceptual input (Botvinick *et al.* [2017]; Hassabis *et al.* [2017]; Buckner [2018]).

However, the DNNs behind these marquee achievements are staggeringly complex and subject to puzzling vulnerabilities, which has led critics to dismiss them as 'black boxes' exhibiting intelligence that is merely ersatz or alien. To cope with this complexity, neural network researchers have suggested that we should engage their behaviour directly with experimental paradigms and data analysis methods derived from the sciences of human and animal behaviour. Such engagement has led neuroscientists to conclude that DNNs are currently the most promising artificial models of perceptual similarity judgements in primates (Khaligh-Razavi and Kriegeskorte

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

[2014]; Lake *et al.* [2015a]; Hong *et al.* [2016]; Kubilius *et al.* [2016]; Yamins and DiCarlo [2016]; Guest and Love [unpublished]). Another area of research aims to extend psychometric methods for intelligence testing in humans to rank the intelligence of artificial computational models (Hernández-Orallo [2017]). Taking the idea that neural networks can be approached with the tools of animal psychology even further, the 'Animal-AI Olympics' has created a testbed application that assesses AI systems on dozens of benchmarks derived from animal cognition research (Crosby *et al.* [2019]; Crosby [2020]). An interdisciplinary coalition of influential scientists has even called for the development of a new scientific field called 'machine behaviour' that would study AI agents in a more contextual and historically informed way, using methods derived from behavioural ecology and ethology (Rahwan *et al.* [2019]).

In short, comparisons between natural and artificial intelligences have never been so varied and ambitious—nor, as we will see below, so fraught. The capacity of DNNs to produce new forms of potentially intelligent behaviour and the development of new methods to evaluate their performance has outpaced our reflection on whether these comparisons are fair or meaningful (Guidotti *et al.* [2019]; Serre [2019]; Zednik [2019]; Zerilli *et al.* [2019]). Moreover, philosophers of science have pointed out that biases plague human evaluation of nonhuman behaviours, and methodological subtlety is required to temper them (Keeley [2004]; Buckner [2013]; Watson [2019]). These difficulties are exacerbated when the other end of the comparison is an artificial system, which are often intended to reproduce only parts or idealized aspects of a cognitive agent (Stinson [2020]). In his defence of his famous imitation game test, Turing himself wrestled with these issues; and commentators have reflected on how to avoid being unwittingly convinced by artificial systems that present the superficial trappings of human-like behaviour (such as human-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

like facial expressions or gestures) without the same underlying competences (Block [1981]; Proudfoot [2011]; Złotowski *et al.* [2015]; Shevlin and Halina [2019]).

This article suggests that this debate about fair comparisons in AI could be expedited by taking the lead from a century of reflection on similar questions in comparative psychology and ethology. While these fields dedicated much effort to developing rigorous empirical methods to avoid anthropomorphism-driven false positives, they have also recently come to grips with the danger of anthropocentrism-driven false negatives. In AI, by contrast, very little of this critical scepticism has yet been directed towards scoring the human behaviours to which AI performance is compared (though for recent exceptions, see Zerilli *et al.* [2019]; Firestone [forthcoming]; Canaan *et al.* [unpublished]).

To illustrate the effect of bias on the evaluation of machine behaviour, Section 2 reviews four popular arguments to the effect that deep learning is fundamentally unlike human learning, all focused on ways in which DNNs allegedly underperform humans. We will see in Sections 3–5 that a bias called 'anthropofabulation' (Buckner [2013])—which scores nonhuman performance against an inflated conception of human competence—threatens the validity of these comparisons. When the same degree of critical scrutiny is directed towards the human side of these comparisons, our minds are also revealed to be black boxes plagued by many of the same vulnerabilities. To sum up, a more apt metaphor for DNNs might be an unflattering if revealing mirror, one which raises new questions about our own intelligence and allows us to see our own blemishes with unprecedented clarity.

2 Four Popular Criticisms of Deep Learning Research

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

This article canvasses and rebuts four criticisms that have been commonly offered against claims that processing in DNNs bears similarity to human cognition: that deep learning is (i) too data-hungry, (ii) vulnerable to adversarial examples, (iii) not interpretable, and (iv) merely reward-hacks rather than learns real solutions to problems. These arguments feature prominently in influential critical reviews of deep learning, such as Lake *et al*. ([2017a]) and/or Marcus ([2018]). To be clear, this is not a complete survey of arguments against the similarity between human cognition and the processing of DNNs. My aim here is not to positively establish a deep similarity between human cognition and DNNs by rebutting all such lines of attack, but rather to redirect attention to the subset of those empirical questions that are more likely to produce fruitful research, and to extract some general lessons about conducting fair comparisons between humans and artificial agents.

Three clarifications on these aims will be useful at the outset (readers wanting to jump straight to the criticisms can skip ahead to Section 2.1). First, though the criticisms and rebuttals discussed here will generalize to many other techniques in machine learning (for a relevant discussion, see Watson [2019]), for ease of exposition we here focus here on deep learning systems, which will be briefly characterized now. DNNs comprise a diverse family of network-based machine learning techniques. As with earlier neural network designs, they consist of layers of simple processing nodes transmitting activation to one another along weighted links, usually intended to model the activity of neurons and synapses at some level of abstraction. In contrast to earlier, shallower neural network architectures, 'deep' neural networks can have anywhere from five to hundreds of layers in-between input and output. Depth itself appears to have profound computational implications; it allows these networks to compose features hierarchically and enjoy ex-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

ponential growth (relative to the number of layers) in their representational capacity and computational power (for a review of evidence for this claim, see Buckner [2019a], Section 2.1).

Such network depth is perhaps the only feature that unites all 'deep' learning systems, and there are many other ways in which their architectures vary. Specifically, they can vary in: the activation functions of their nodes; the connectivity patterns between their layers and number of nodes in each layer (esp. decreasing the numbers in successive layers to impose 'bottlenecks' in processing); their learning rules or training regimes (such as backpropagation, reinforcement, or predictive learning); whether they feature recurrent links connecting later layers back to earlier ones; the use of components or multiple networks to simulate the modulatory effects of memory buffers or attentional control; and the ways in which their processing is tweaked ('regularized') to avoid overfitting spurious correlations in the training set (Schmidhuber [2015]).

To briefly canvass some of the most popular architecture combinations, deep convolutional neural networks (DCNNs) have perhaps featured most prominently in marquee achievements; they leverage a sequence of different activation functions (convolution, pooling, and rectification) to perform hierarchical feature detection, and deploy mostly local connectivity between layers (LeCun *et al.* [2015]; Buckner [2018]). Deep autoencoders impose a bottleneck in the middle of a deep layer hierarchy, with an architecture resembling an 'hourglass' shape with fewer and fewer nodes in the central layers, forcing the network to learn compressed representations that condense categories to their 'gist' (Hinton and Salakhutdinov [2006]). Generative Adversarial Networks (GANs) have also captured the public's attention; they involve tasking a second generative network to fool a primary discriminative network (often a DCNN), with the generative network's nodes performing activation functions akin to the inverse of convolution and pooling ('deconvolution' and 'unpooling') to produce highly detailed and realistic 'deepfakes'

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

and 'adversarial examples' that can pose a security risk to discriminative networks (Goodfellow *et al.* [unpublished]). Variational autoencoders (VAEs) combine features of GANs and deep autoencoders; they attempt to learn hidden relationships between latent variables that could be used to reconstruct its training data (Kingma and Welling [unpublished]). Long short-term memory networks (LSTMs) deploy recurrent connections in memory cells to simulate a memory for context, and can excel at processing complex sequences in input like grammatical structures (Hochreiter and Schmidhuber [1996]). Transformers—the most sophisticated language-production deep learning architecture to date, exhibited in systems like BERT, GPT-2, and GPT-3—modulate relatively homogeneous deep neural networks using a complex form of hierarchical attention to represent multiple channels of complex syntactic and semantic information relevant to predicting word placement in language production and automated translation (Vaswani *et al.* [2017]).

As a second introductory clarification, we consider three other prominent criticisms that readers might be anticipating, in order to set them aside for the remainder of the article. Specifically, this article will not engage with claims that (a) DNNs cannot create new compositional representations on-the-fly, (b) strategies learned by DNNs do not transfer well to radically different tasks or stimuli, and (c) that DNNs cannot learn to distinguish causal relationships from mere correlations. Whether current or future DNN architectures can achieve such compositionality, radical transfer, and causal inference remain open empirical questions (Lake [2014]; Battaglia *et al.* [unpublished]; Russin *et al.* [unpublished]), ones that will hopefully receive more attention in future research. The ability to learn and reason about causal relationships in particular might be thought a distinguishing feature of human cognition and a key goal for more human-like AI (Penn and Povinelli [2007a]; Hespos and VanMarle [2012]; Pearl [2019]). Granted, most neural

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

networks are not trained to diagnose causal relationships, and many humans confuse correlation for causation (Lassiter *et al.* [2002]). When neural networks are trained to diagnose causal relationships, they have shown some successes, especially generative architectures like variational autoencoders (Kusner *et al.* [unpublished]; Zhang *et al.* [2019]) and models that use deep reinforcement learning (Zhu *et al.* [unpublished]). That said, comparative biases will surely affect these debates too, and we may hope that the four rebuttals canvassed here will suggest how to mitigate them when they do.

Finally, in what follows, we will not here discuss linguistic behaviour or cognition. The likeliest default position is that compositional recursive grammar is a uniquely human capacity amongst animals, and some classical criticisms of the neural network approach take this to be essential for intelligent behaviour (Fodor and Pylyshyn [1988]). Furthermore, this capacity is engaged by many classic assessments of artificial intelligence like the Turing Test, and deep learning models—especially massive transformers like GPT-3—have recently achieved impressive results on tasks like automated translation, question answering, and text production. However, this capacity is closely related to the other three that we have already set aside, and the way that the brain enables linguistic production remains contentious in developmental linguistics and cognitive neuroscience (Fitch [2014]; Scott-Phillips *et al.* [2015]; Berwick and Chomsky [2017]; Moore [2017]). Again, the goal of this article is not to positively establish that DNNs are intelligent by rebutting all comers, so we leave the question of whether current or future DNN architectures can implement compositional recursive grammar open (Russin *et al.* [unpublished]; though see Lake [2019]). The kinds of biases that will be described for perceptual decision-making and strategy game-play also appear in the linguistic domain (including the Turing test), so this may seem an odd omission given the article's aims. The reason for it is simply that the

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

evaluation of linguistic behaviour from deep learning systems (especially transformers like GPT-3) deserves its own specialized article- (or book-)length treatment, whereas issues of comparative bias are already complex enough in the simpler systems and applications to occupy us here.

With these clarifications in place, we now proceed to review the four popular criticisms that will be considered here.

2.1 Deep learning is too data-hungry

One of the most common critical refrains is that DNNs require far more training data than humans to achieve equivalent performance. The standard methods of training image-labelling DNNs, for example, involves supervised backpropagation learning on the ImageNet database, which contains 14 million images that are hand-annotated with labels from more than 20,000 object categories. To consider another example, AlphaGo's networks were trained on over 160,000 stored Go games recorded from human grandmaster play, and then further trained by playing millions of games against iteratively stronger versions of itself (over 100 million matches in total); by contrast, AlphaGo's human opponent Lee Sedol could not have played more than 50,000 matches in his entire life. In the human case, critics emphasize the phenomena of 'fast mapping' and 'one-shot learning', which seem to allow humans and animals to learn from a single exemplar. For example, Lake *et al*. ([2015b]) argue that humans can learn to recognize and draw the components of new handwritten characters, even from just a single example (Fig. 1). Sceptics thus wonder whether DNNs will ever be able to learn comparatively rich category information from smaller, more human-like amounts of experience.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

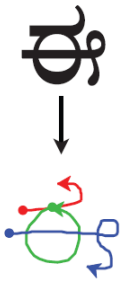
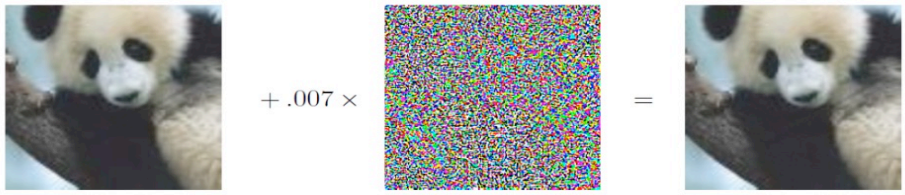


Figure 1. The decomposition of a novel handwritten figure into three individual pen strokes, which humans can purportedly learn from a single exemplar (reproduced from Lake et al. [2015b]).

2.2 Adversarial examples expose deep learning as a fraud

'Adversarial examples' are unusual stimuli that are generated by one 'adversarial' DNN to fool another. The original adversarial examples were 'perturbed images' that were created by a generative adversarial network (GAN) by slightly modifying an easily classifiable exemplar in a way that was imperceptible to humans, but which could cause dramatic misclassification by DNNs targeted for attack (Goodfellow *et al.* [unpublished]; and see Fig. 2). Perturbation methods most commonly modify many pixels across an entire image, but they can be as focused as a single-pixel attack (Su *et al.* [2019]). The pixel vectors used to perturb images are usually discovered by training the adversarial DNN on a discriminative DNN's response to specific images, but some methods can also create 'universal perturbations' that disrupt classifiers on any natural image (Moosavi-Dezfooli *et al.* [2017]).

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.



'panda' 'nematode' 'gibbon'

57.7% confidence 8.2% confidence 99.3% confidence

Figure 2. An adversarial perturbed image, reproduced from (Goodfellow *et al.* [unpublished]). After the 'panda' image was modified slightly by the addition of a small noise vector (itself classified with low confidence as a nematode), it was classified as a gibbon with high confidence, despite the modification being imperceptible to humans.

It was soon discovered that many perturbation attacks can be disrupted with simple pre-processing techniques, such as systematic geometric transformations of images like rotation, re-scaling, smoothing, and/or de-noising (a family of interventions called 'feature squeezing'—Xu *et al*. [2017]). A reasonable interpretation of this phenomenon is that DNNs are vulnerable to image perturbations because their perceptual acuity is too keen; the attack exploits their sensitivity to precise pixel locations across an entire image, so it can be disrupted by slightly altering the pixel locations across the entire input image.

However, another family of adversarial example generation methods—involving the creation or discovery of 'rubbish images' that are supposed to be meaningless to humans but confidently classified by DNNs—were found to be more resistant to such default countermeasures (Nguyen *et al.* [2015]). Subsequent research has found that these (and other) adversarial examples exhibit many counterintuitive properties: they can transfer with (incorrect) labels to other DNNs with different architectures and training sets, they are difficult to distinguish from real exemplars using pre-processing methods, and they can be created without 'god's-eye' access to model parameters or training data. Rather than being an easily overcome quirk of particular models or training sets, they appear to highlight a core characteristic of current DNN methods.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Much of the interest in adversarial examples derives from the assumption that humans do not see them as DNNs do. For practical purposes, this would entail that hackers and other malicious agents could use adversarial examples to fool automated vision systems—for example, by placing a decal on a stop sign that caused an automated vehicle to classify it as a speed limit sign (Eykholt *et al.* [2018])—and human observers might not know that anything was awry until it was too late. For modelling purposes, however, they might also show that despite categorizing naturally occurring images as well or better than human adults, DNNs do not really acquire the same kind of category knowledge that humans do—perhaps instead building 'a Potemkin village that works well on naturally occurring data, but is exposed as fake when one visits points in [data] space that do not have a high probability' (Goodfellow *et al.* [unpublished]).

2.3 Deep neural nets are not interpretable

Another common lament holds that DNNs are 'black boxes' that are not 'interpretable' (Lipton [unpublished]) or not 'sufficiently transparent' (Marcus [2018]). State-of-the-art DNNs can contain hundreds of layers and billions of individual parameters, making it difficult to understand the significance of specific aspects of their internal processing. However, key questions in this charge remain unanswered (Zednik [2019]), such as: What kind of interpretability needs to be provided, to whom should the interpretation be provided, what is the purpose of interpretability, and how would we know whether we had succeeded in providing it? At any rate, these concerns should only be counted against deep learning models if some obvious alternative systems perform better on them. While DNNs are often compared to linear models (which are—probably incorrectly—thought to be more interpretable), usually the comparison class is adult humans. Recent governmental initiatives such as DARPA's eXplainable AI (XAI) challenge (Fig. 3) and

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

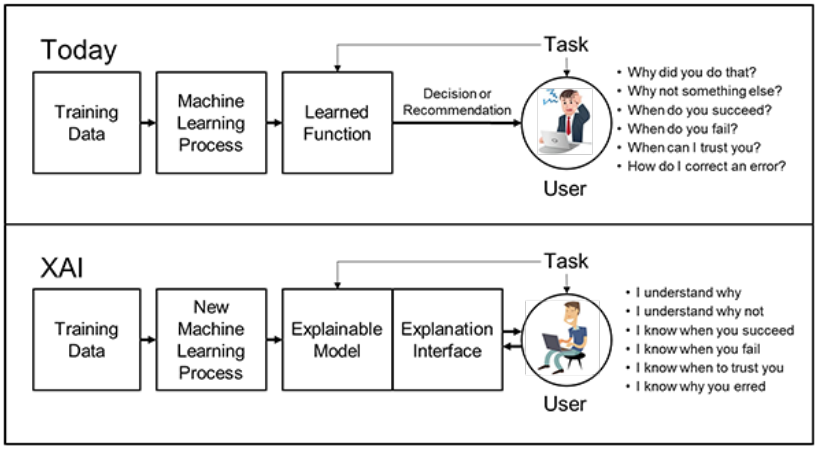


Figure 3. The DARPA XAI concept; figure created by DARPA for public release (Turek [unpublished]).

the EU's General Data Protection Regulation—which provides users with a 'right to explanation' for decisions made by algorithms that operate on their data—have quickened the challenge and provided it with some practical goals, if not always conceptual clarity (Goodman and Flaxman [2017]; Turek [unpublished]).

2.4 Deep neural nets trained by reinforcement learn to 'reward hack' rather than solve problems

Many of the most impressive achievements by DNNs highlighted above were produced by reinforcement learning (for an overview of this area, see Sutton and Barto [2018]). This method trains networks using a general reward signal that is designed by the network's programmers and tells the network whether it succeeded or failed on its last decision. Many of the high-profile achievements of DNNs involved games like Go, chess, or Starcraft II because game score provides an easily quantifiable reward signal. In other areas of research such as artificial locomotion, creating an effective reward signal is more difficult. Many reward an agent for simply moving forward in an artificial environment, perhaps with minimal energy expenditure by its digital

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

avatar. For example, one deep reinforcement model trained in the 'Half-Cheetah' testbed environment—in which models learn to move an idealized, two-dimensional cheetah avatar forward by manipulating several points of freedom in its two legs—learned that it could locomote the cheetah by falling forward and then flailing the legs in the air so as to flop the avatar forward on its back (Irpan [unpublished]). In another widely shared blog post written by the research group OpenAI, the researchers recount how their DNN learned to play the boat racing Atari game 'Coast Runners' by endlessly turn the boat in tight, off-course circles without ever completing the race, because doing so allowed it to continually collect replenishing 'turbo' bonus widgets that provided a rapid, never-ending boost to its game score (Amodei and Clark [unpublished]). Critics worry that these examples show that the models lack the 'common sense' that humans would bring to bear on these tasks, and that the solutions they learn are brittle 'reward hacks' that optimize the reinforcement signal without any real understanding of the problems they are trained to solve.

3 Purposes, Interests, and Fair Comparisons

There are many reasons why we might want to compare different kinds of agents in terms of their intelligence, rationality, or other mental abilities. For one, such comparisons can serve metaphysical goals: we may want to learn about the different ways that intelligence can be realized in nature or artefacts—as has been traditionally explored in the literature on 'multiple realizability' (Polger and Shapiro [2016]). Second, such investigations can serve semantic projects, by helping us clarify these concepts, which are often vaguely defined or equivocal (Akagi [2018]; Miracchi [2019]). Third, they can serve more practical goals: we may be interested in the epistemic, ethical, or legal status of other kinds of agents, and the possession of specific mental abili-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

ties may be relevant to those statuses (Allen [2006]; Andrews *et al.* [unpublished]). Fourth, they can be used for scientific modelling purposes in human cognitive psychology, to better understand and explain how intelligent behaviour is produced in our own case, by engineering systems based on different hypotheses and comparing their performance to human behaviour or their structure to that of the human brain (Stinson [2020]). Fifth, comparisons can serve a variety of engineering or medical projects: we may want to establish the suitability of artificial models to predict the results of medical interventions on human brains before conducting human trials, or as alternatives for human labour in a variety of different applications (Hassabis *et al.* [2017]).

Though some of these purposes have been more frequently discussed in the context of animal psychology, they will become increasingly relevant to artificial intelligence as our computational models are able to successfully replicate more and more aspects of human and animal behaviour. Though this list is not exhaustive, we can already see that there are many competing pressures underlying such comparisons, even and especially when the goals of the comparisons are not made explicit. All of the aims, however, should focus on the degree of relevant underlying similarity that holds between the two systems to determine whether they succeed. From a philosophy of science perspective, we should accept that these models often only need to reproduce parts or idealized aspects of these phenomena to serve their purpose; as Stinson ([2020]) puts it, quoting Winsberg ([2010]), the right relationship is often something far more complicated and subtle than 'mere mimicry'.

So, how similar, or in what way, must a DNN's processing be to a target system or mental ability to serve as an artificial implementation of it that is useful for these purposes? Obviously, some aspects of a DNN's implementation will be irrelevant to all of these goals; we should not fault artificial systems because they require external electricity sources to perform their pro-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

cessing any more than we should reward them for being able to function better than humans in low-oxygen environments. One way to pose this question emphasizes the traditional distinction in cognitive science between competence and performance (Firestone [forthcoming]); artificial models should engage the same underlying competence that humans do when performing some task, but do not need to reproduce all the performance factors. One concern about this strategy, however, is that competences can be construed in different ways, inviting evaluative differences to masquerade as empirical ones. To review a topically relevant example, one diagnosis of the famous disagreement between classicists Fodor and Pylyshyn ([1988]) and the connectionists (such as Smolensky [1988]) is that Fodor and Pylyshyn were only interested in a particular sort of explanation of compositionality and systematicity, whereas the connectionists were interested in many other phenomena that were better (or only) explained by connectionist representations (Matthews [1994]). Differences of explanatory interest are common in debates in cognitive science, which perhaps explains why they are often difficult to resolve by empirical means (for another case study of such an impasse in comparative social psychology, see Penn and Povinelli [2007b]; Call and Tomasello [2008]; Buckner [2013]).

One should be wary that one has been invited to such a masquerade whenever critics argue that only systems meeting certain restrictive criteria count as 'genuine', 'real', 'strong', or 'bona fide' examples of mental capacities like intelligence, learning, rationality, cognition, and so on. Despite the formal ornament gilding these critiques, these adjectives are not natural kind terms with empirical content; they are rather baldly honorific, and their evaluative criteria can be stipulated arbitrarily from one moment to the next to suit the critics' whims. Indeed, such honorifics are beginning to show up in the appraisals of deep learning critics such as (Bringsjord *et al.* [2018]), who allege that deep neural networks are not capable of 'real' learning—which these

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

authors hold is instantiated only in cases where agents can provide demonstrations for what they have learned adverting to formal definitions of key terms involved, as (to use an example suggested by an anonymous reviewer) a math student may produce in a proof of the fundamental theorem of algebra. This benchmark produces the surprising verdict that children do not really learn how to walk, talk, or recognize objects, when it is sensible to suppose that artificial intelligence should aim to solve these basic competences on the road to more ambitious ones. A diagnosis of this debate is that these critics are only interested in a special kind of learning that is paradigmatically instantiated in mathematical education, but which is hardly as central to other characteristically human cognitive competences as they suppose.

While it is usually otiose to belabour such matters of taste or terminology, there are some practical disadvantages to indulging such restrictionism when it comes to such general terms as 'intelligence', 'learning', 'rationality', or 'cognition' (Akagi [2018]). First, there is worry that such critiques would confine AI to blind alleys that had already been explored in earlier stages of research. Starting out by attempting to build systems that can solve pinnacle human achievements using declarative knowledge derived from human verbal justifications has repeatedly produced fragile systems that can mimic human behaviour only in limited applications involving pre-digested input for which they were explicitly programmed, but which can do little else, and whose behaviour fails to generalize to situations even slightly outside of their programming (Hofstadter [1985]; Brooks [1991]). Though IBM's DeepBlue defeated world champion Garry Kasparov in chess in 1997—perhaps the highest-profile achievement of this top-down approach to AI—it would have to be completely reprogrammed to play another game. Reinforcement-learning-based DNNs, by contrast, have by now shown an impressive ability to learn their own

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

solutions to dozens of different games without changing their algorithms (Mnih *et al.* [2015]; Silver *et al.* [2018]; Lyre [2020]).

Second, such stipulations can close off questions that ought to be settled by empirical rather than terminological methods (Allen [2017]; Ramsey [2017]). For example, even if mathematical cognition were our primary interest, empirical investigation of mathematical demonstration shows that low-level perceptual and pattern-matching abilities are more involved in the reliable manifestation of these competences in typical math students than we would have presumed from the armchair (Landy *et al.* [2014]). And finally, reliance on such honorifics has a way of leading to constantly shifting goalposts; every time an animal or artificial system satisfies a previously specified benchmark, the critic can simply endorse a yet more restrictive interpretation of 'real' or 'genuine' and push the borderline ever-closer to the uppermost limits of human performance—and possibly even beyond. For example, these interests led the same critics to conclude controversially that human cognition is hypercomputational, without providing any empirical evidence that humans reliably hypercompute or ethological investigation into the conditions in which they do so that would be required to conduct fair comparisons (Chalmers [1995]; Bringsjord and Arkoudas [2004]; Davis [2004]; Govindarajulu and Bringsjord [2012]).

4 A Crash Course on Comparative Bias

In this section, we extract a general lesson that can help us avoid these pitfalls by looking to other sciences that have faced similar pressures. Comparative psychology and cognitive ethology have struggled to fairly align different kinds of intelligences for more than a century, and have by now come to appreciate that human researchers are vulnerable to systematic biases that can distort such comparisons by causing us to rush to judgement without properly evaluating the rel-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

evant underlying similarities. To counter these biases, the study of machine behaviour should adopt similar methodological correctives, such as Morgan's canon and Hume's dictum (Buckner [2013]; Rahwan *et al.* [2019]). One bias that has already been well studied by philosophy of comparative psychology and artificial intelligence is anthropomorphism (de Waal [2000]; Wynne [2004]; Proudfoot [2011]). A sizeable literature in comparative psychology explores correctives for anthropomorphism and their proper application (Sober [1998]; Karin-D'Arcy [2005]; Buckner [2017]). On the other hand, there are also a variety of anthropocentric biases that can thumb down the scales against nonhumans. Anthropocentrism can cause us to assume that only behaviours with the superficial trappings of human performance are valuable or intelligent—such as supposing that only animals that navigate by sight could possess cognitive mapping, when bats or dolphins might create maps of their environment using echolocation. Semantic anthropocentrism is usually a mistake, but not always; in cases where traits really are uniquely human—as again is probably the case with semantically compositional language with recursive grammar (Fitch [2010]; Berwick and Chomsky [2017])—semantic anthropocentrism may be unavoidable.

One form of anthropocentrism is guaranteed to be a mistake, however: the bias of 'anthropofabulation' (Buckner [2013]). Anthropofabulation combines semantic anthropocentrism with an exaggerated view of human cognitive performance. Anthropofabulation results from an empirically uninformed picture of human cognitive processing derived from introspection or cultural traditions. Common sense in some cultures tells us that our thought processes are rational—derived from a dispassionate processing of the situation, a direct introspective access to our actual beliefs and motivations, and independence from subtle environmental scaffolding, historical associations, or emotional reactions. A great deal of human social psychology and philosophy of

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

psychology, however, has cast this picture of human cognition into doubt (Nisbett and Ross [1980]; Kahneman and Frederick [2002]; Samuels *et al.* [2002]; Carruthers [2011]).

In practice, anthropofabulation has caused sceptics to compare human and animal performance in situations that are crucially disanalogous, such as when humans are tested with conspecifics but chimpanzees with heterospecifics, humans tested in a known caregiver's lap while chimpanzees are tested with strangers behind Plexiglas, or humans are tested on culturally familiar stimuli while chimpanzees are tested on unfamiliar artificial stimuli (Boesch [2007]). Anthropofabulation's rosy vision of human cognition causes us to implicitly assume that human performance could not possibly depend upon such environmental scaffolding, leading us to overlook or downplay these disanalogies. While these disanalogies are generally now seen as mistakes in comparative psychology, we are only beginning to appreciate their analogues in artificial intelligence (Zerilli *et al.* [2019]; Firestone [forthcoming]; Canaan *et al.* [unpublished]). The remainder of the article argues that critics of DNNs are similarly evaluated in unfairly disanalogous situations or by assessing penalties to DNNs for factors that apply equally well to adult human cognition. Once the anthropofabulation in these critiques is exposed, they no longer clearly support the conclusion that deep learning systems and human brains are performing fundamentally different kinds of processing—and indeed, might teach us hard lessons about our own cognition as well.

5 Four Rebuttals

This final section illustrates and rebuts anthropofabulation in the four criticisms of deep learning on which we focused above. This story is as much or more about humans than about machines; indeed, the story's moral is that artificial intelligence researchers need to draw less upon intro-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

spection and more on an unbiased, empirically grounded appraisals of human intelligence—warts and all—to fairly evaluate machine behaviour. In many cases, when we do this systematically, we will find that the machines have not been given the same kinds of tasks or provided with the same kind of training as the humans, even when it is possible to do so (Firestone [forthcoming]).

5.1 Human learning involves more trainable exemplars than common sense supposes

One way that anthropofabulation might bias us against DNNs is by causing us to undercount the number of trainable instances that should be scored to adult human performance. Two factors are often neglected in counting the number of exemplars that humans should be scored as having been exposed to in learning: 1) that many different vantages of the same object can provide distinct training exemplars for cortical learning, and 2) that offline memory consolidation during sleep and daydreaming can replay the same exemplars—and even simulated novel exemplars generated from those same experiences—many thousands of times in offline repetitions. Ignoring these factors, common sense might score an infant's ten-minute interaction with a new toy as a single exemplar.

It is difficult to decide exactly which features of human perceptual learning are relevant to the comparison in order to devise a proper accounting system for humans, but we can review some results in the neighbourhood. Studies of motion-picture perception have suggested that human vision has a frame rate of about ten to twelve images per second (below this rate, we cannot perceive motion as continuous). We can also consider how long it takes us to become consciously aware of or be affected by a stimulus; while it takes 200-400 ms for us to become consciously aware of a perceptual stimulus, attentional shifting to a new stimulus begins in as little

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

as twenty milliseconds, and category structure can be implicitly influenced by nonconscious exposures to stimuli as brief as one millisecond (Kunst-Wilson and Zajonc [1980]; Schacter [1987]; Murphy and Zajonc [1993]). Moreover, perceptual memories may be repeatedly reconsolidated by theta rhythm in the medial temporal lobes during sleep and daydreaming many times over a period of months and years (Stickgold [2005]; Walker and Stickgold [2010]). We also know that in mammals, these consolidation exposures can train the cortex on novel experiences synthesized from combinations or transformations of previous training information—as revealed by cell recordings that show rats mentally exploring novel maze routes during sleep that they never actually traversed when awake (Gupta *et al.* [2010]). Taking all these factors into account, an infant's ten-minute interaction with a new toy might be fairly scored as providing tens of thousands of trainable exemplars, rather than a single one, as common sense might suppose. In this sense, Herbert Simon's classic quip that 'everything of interest in cognition happens above the 100-millisecond level' is classic anthropofabulation, focusing attention on only the introspectively available surface features of human categorization while ignoring a vast iceberg below (Hofstadter [1985]).[1]

Neither is this merely idle nit-picking; neural network models that attempt to replicate these nonconscious aspects of human learning can make more efficient use of smaller, more human-like training sets. For example, when deep learning models are trained on successive frames of video rather than static exemplars, many different vantage points on the same object can be treated as independent training instances that improve model performance (Luc *et al.* [2017]; Lotter *et al.* [unpublished]; Orhan *et al.* [unpublished]). When DNNs are supplemented with 'episodic

[1] Granted, useful information can be obtained from the first-person perspective; Ericsson and Simon ([1984]) emphasized speak-aloud protocols, which can provide useful information about the information attended to by a subject, but which are quite different than the kind of rationalization considered in Section 5.3.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

replay' buffers that are inspired by declarative memory faculties in mammals, a network's performance can continue to benefit from repeatedly replaying exposure to the same training instances numerous times (Mnih *et al.* [2015]; Blundell *et al.* [2016]; Vinyals *et al.* [2016]). Predictive, 'self-supervised' networks—which attempt to learn by predicting the future from the past, the past from the present, occluded aspects of objects from the seen aspects, and so on—are championed as the future of the field by DNN pioneers like LeCun ([2018]). There is little evidence that the efficiency gains that can be obtained from such biologically inspired innovations have already plateaued.

Still, critics hold that this all falls short of the kind of one-shot learning of novel digits and their construction emphasized by some critics, which has purportedly been modelled in some Bayesian systems (see Section 2.1 above). While numerous DNN systems produce one-shot or even zero-shot learning on related tasks (Socher *et al.* [2013]; Rezende *et al.* [2016]; Brown *et al.* [unpublished]), critics note that they do so only with extensive pre-training. Nevertheless, there remain significant questions about the fairness of this response. Humans are capable of such one-shot learning only after extensive practice in recognizing and generating a variety of different handwritten figures, experience that has occurred outside the purview of any laboratory experiment. The Bayesian programs that are purported to model this one-shot learning must incorporate significant amounts of high-level knowledge and representational structures that are manually encoded by their programmers (Botvinick *et al.* [2017]). These Bayesian modellers on some occasions profess agnosticism as to the origins of this knowledge, and on others wave their hands at genetically programmed innate mechanisms (Lake *et al.* [2017b], p. 53). Such specific forms of knowledge are not plausibly encoded directly in the genome, however, which likely only contains enough storage space to specify very general wiring principles of the sort that already

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

make DNNs especially good at things like translation invariance and which were inspired by neuroanatomical observations (Zador [2019]). In short, until the cognitive provenance of this knowledge is accounted for in humans—specifically, until we know the nature and number of training exposures adult humans require to scaffold such one-shot learning, and how their genetic scaffolding expresses itself in the human brain—these concerns cannot fairly be scored against DNNs in this debate.

5.2 Deep neural net's verdicts on adversarial examples may be correct

Recent investigations have challenged the assumption that a DNN's take on adversarial examples is really so alien to human perception. One still-controversial way to challenge this assumption is by using perturbation methods to produce artificial stimuli that can fool humans (Elsayed *et al.* [2018]). Even more interestingly, however, Zhou and Firestone ([2019]) showed that humans can easily 'adopt the machine perspective' and, when forced to choose between a predetermined list of candidate labels, predict a DNN's labels for rubbish images with high accuracy (Fig. 5). These authors suggest that the behaviour of DNNs in these cases that initially appeared to be an error might have been due to the fact that during training and testing, the DNNs were always forced to choose amongst a list of candidate labels, even when images were very different from previously classified exemplars. Humans, by contrast, can typically reject stimuli as unusual or ambiguous.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

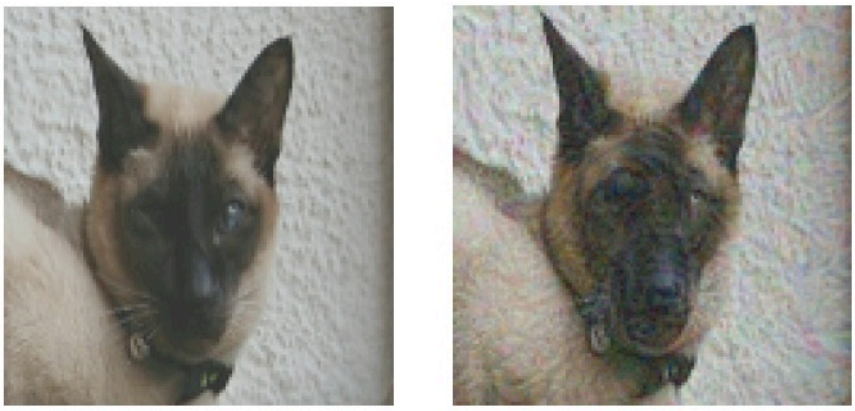


Figure 4. A perturbed image that can purportedly fool human subjects, with the original image of a cat on the left, and the perturbed image (often classified as a dog) on the right. Image reproduced from (Elsayed *et al*. [2018]).

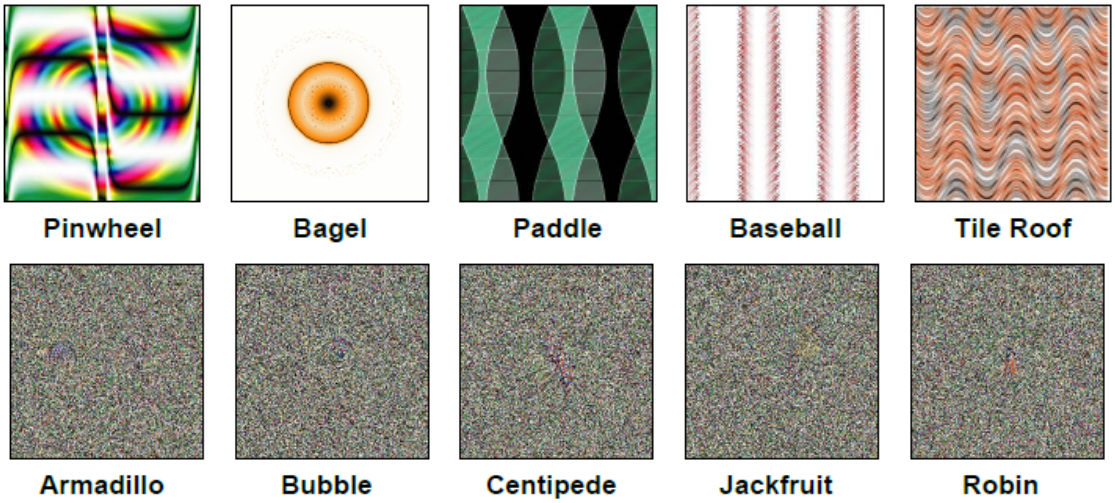


Figure 5. Examples of two different types of rubbish images tested by Zhou and Firestone ([2019]) with preferred DNN labels. In a forced-choice task, humans were able to guess a DNN's preferred labels for these images with high accuracy. (Image reproduced from Zhou and Firestone [2019]).

This difference marks a crucial disanalogy in many comparisons between natural and artificial judgements on adversarial examples, a difference that may be obscured by anthropofabulation. Specifically, Zhou and Firestone's results suggest that DNNs do appear to capture some aspects of lower-level perceptual categorization in humans; many rubbish images do look like

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

members of the purportedly incorrect label class, even if humans do not ultimately think that they look like they are members of that class (the way an intrinsically meaningless inkblot in a Rorschach test may look like a duck without looking like it is a duck). DNNs may thus be correctly delivering human perceptual similarity judgements, but not yet have the resources to draw a distinction between an exemplar superficially resembling something and actually looking like a member of the class.[2] This kind of distinction is difficult for even human children and adult chimpanzees to master (Flavell *et al.* [1983]; Krachun *et al.* [2016]), and the DNN modellers did not even attempt to train their networks to perform this kind of discrimination. Perhaps it remains an open question how to model the latter kind of judgement in DNNs (Smith [2019]), but currently available comparisons do not yet demonstrate that a DNN's processing is hopelessly alien to human perception.

Even more recently, commentators have begun explicitly calling out the foundational anthropocentrism of the debate over adversarial examples, by questioning whether the verdicts DNNs issue on these unusual stimuli should be considered mistaken or unintelligent in the first place. A ground-breaking series of empirical studies by Ilyas *et al*. ([unpublished]) recently suggested that vulnerability to adversarial examples may be a feature and not a bug of DNNs. These authors discovered two surprising things: first, that when DNNs were trained exclusively on a diet of adversarial examples, their classification behaviour transferred well to novel natural images, and second, that when their training sets were altered to remove the features that caused them to be susceptible to adversarial examples, their discrimination performance on natural stimuli was also significantly diminished. Combined, these two findings suggest that the features to which DNNs

[2] One residual concern here, pointed out by an anonymous reviewer, is that the kinds of errors made by these networks may evince what Watson ([2019]) calls 'myopia', or a tendency to ignore structural relationships that seem obvious to humans.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

respond in adversarial examples are well-generalizing aspects of naturally occurring data: they are predictively valid in naturally distributed data, and humans may only fail to deploy them in their own categorizations due to comparatively inferior perceptual or cognitive acuity. While this does not diminish the practical significance of the phenomenon as a security threat, it raises philosophical questions as to which features ought to be relevant to assessing intelligence in categorization tasks.

These questions may soon become especially pressing, for the detection of such features may have enabled some DNNs to make dramatic leaps beyond the limits of human intuition on problems characterized by high complexity and holistic nonlinear interactions—such as the prediction of stable end states for folding proteins, a problem on which the DNN-based AlphaFold system recently outperformed human modellers who had devoted their professional lives to solving this kind of task (AlQuraishi [2019]). Perhaps the DNNs can discover intricate, high-frequency 'interaction fingerprints'—similar in form to the features that cause them to be vulnerable to adversarial perturbations—that point the way to new discoveries in disease diagnosis and drug development, but which are beyond human ken (Gainza *et al.* [2019]). It is difficult to justify the conclusion that science should eschew such features without simply relying on a flat-footed form of anthropocentrism; and pragmatic philosophers of science would have little grounds for turning down this more fecund future science, even if its course is driven by inscrutable DNNs. If these categorizations are not necessarily blunders, then the ability of deep nets to detect the features on which they are based should no more be counted against their candidacy for intelligence than the ability of Einstein to see things others did not in the equations describing gravity and black holes. Though we have here raised more questions than we have answered, we can already reject the common, anthropofabulous conclusion that the DNNs' verdicts on adversarial examples expose

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

them as exhibiting merely ersatz intelligence; from there, we must leave the full investigation of adversarial examples and their implications to other work (Buckner [unpublished]).

5.3 Human decision-making is also opaque

As noted above and in several critical analyses, the interpretability challenge conflates several different concerns that are probably best separated. To make a start at disentangling them, the distinction between explanatory rationality and justificatory rationality may be useful here (Buckner [2019b]). Questions of explanatory rationality concern the causal history of agent's decision-making in terms of its internal reasons for acting—that is, the evidence or grounds that it acted upon when producing the output that it did in that situation. In the XAI challenge, for example, the questions 'Why did the model do that?', 'Why not something else?', and 'How do I correct an error?' concern dimensions of explanatory rationality. Justificatory rationality, on the other hand, involves the correctness or trustworthiness of the model's decisions, which may or may not cite causally determinative factors. In the XAI challenge, this covers the questions, 'When do you succeed or fail?', 'When can I trust you?', and especially 'Why was that the correct thing to do?'. A key concern here is that we should not expect a single approach to the interpretability challenge to simultaneously address both dimensions of rationality; it is possible that causal explanations of the nets' behaviours may not cite factors that provide intelligible justifications to humans, and justifications may not cite causally determinative factors. Anthropofabulation causes us to conflate these two kinds of concern, however, because common sense supposes that the justifications humans produce through introspection have direct, non-inferential access to the causal antecedents of the behaviours so justified. However, a significant amount of cognitive science suggests that this picture of human introspection is mistaken.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

To provide some examples, one of the reasons that people have supposed the internal processing of DNNs to be opaque is that popular visualization methods that have been developed to determine the representational functions of their hidden nodes have produced strange, chimerical images. Activity maximization is perhaps the most popular method; it tweaks input images using further machine learning until they maximally activate some particular node in a DNN's internal layers. This is supposed to show us the feature that node detects in input images when it activates. A widely circulated paper from Google's AI research group noted that their popular Inception network seemed to detect a variety of chimerical features in images, such as 'pig-snails', 'admiral-dogs', and 'camel-birds' which resemble no intuitively available features in conscious human perception (Mordvintsev *et al.* [unpublished]; and see Fig. 6).

However, activity maximization is a new visualization technique that is poorly understood and very unlike introspection in humans; directly comparing introspectible features to its results is like comparing apples to resequenced orange DNA. There is little reason to suppose that we have the ability to introspectively generate images that maximally activate particular neurons somewhere in our visual cortex. It is also likely that representation in visual cortex is highly distributed across many neurons, so individual neurons in primate brains probably lack intelligible representational functions to begin with (Plaut and McClelland [2010]). In fact, when activity maximization is applied to neurons in a live monkey's brain, the synthesized images are similarly chimerical (Ponce *et al.* [2019]; and see Fig. 7). In short, these methods may have some useful role in addressing explanatory questions—telling us why, causally, the DNN (or monkey) reacted in that way to that exemplar; but we should not expect the images produced by these methods—either in DNNs or biological brains—to provide intuitively interpretable justifications.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Figure 6. Results of running an activity maximization algorithm on a picture of clouds in a trained-up version of Google's Inception image-classifying DNN. Reproduced from (Mordvintsev *et al*. [2015]).

Figure 7. Results of running an activity maximization algorithm on an electrode implanted to detect the firing rate of a live monkey neuron, reproduced from (Ponce *et al*. [2019]).

On the side of justificatory rationality, methods have been designed to generate justifications for DNN behaviour that humans find intuitively satisfying, but they have been criticized for failing to highlight causally determinative factors. Many of these methods rely on producing verbal justifications for a network's decisions that are the result of further machine learning. For example, the 'AI Rationalization' system collects a series of verbal justifications from humans while playing the Atari game 'Frogger', and then uses further machine learning to correlate those verbal justifications with cases where a DNN made similar decisions in similar circumstances (Ehsan *et al.* [2018]; and see Fig. 8). The system can then deliver those justifications to human

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

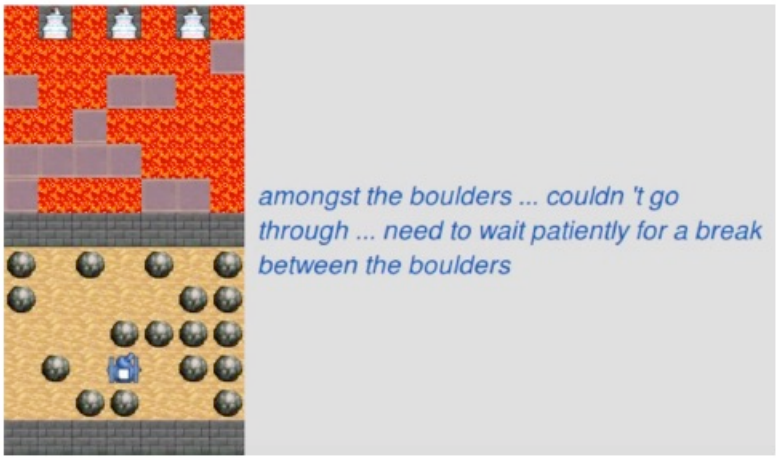


Figure 8. The 'Rationalizing Robot' from (Ehsan *et al*. [2018]) providing an example rationalization of its decisions.

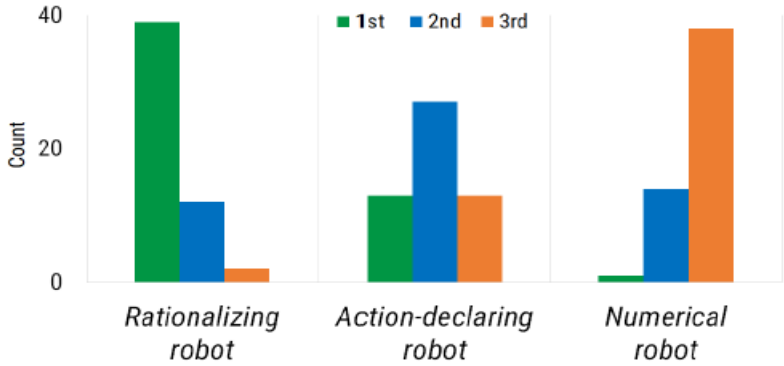


Figure 9. Favourability rank-orderings from human subjects who were asked to rank their preferences for three different policies, as reported by Ehsan *et al*. ([2018]). The 'Rationalizing robot' provided the human-correlated justifications for its actions; the 'action-declaring robot' simply stated the action it was going to perform as it did it; and the numerical robot provided its calculated confidence values for the actions it had just performed (which is perhaps the most causally accurate explanation for the robot's decision-making).

observers to support its decisions after they have been made. The researchers who developed this system obtained user-satisfaction ratings from three different justification policies. Human subjects reported finding the human-derived rationalizations more satisfying than more causally accurate alternatives (in fact, the more causally accurate a justification was, the less subjects liked it; see Fig. 9). The authors conceded that there is no direct causal link in this case between the features that actually caused the system to make the decision and the features cited in the verbal

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

justification. However, these authors note that social psychology similarly finds a disconnect between human rationalizations and the factors that actually caused the actions so rationalized. In fact, the best empirical theories of these systems in humans construe them as interpretive and inferential, generated post-hoc to promote social acceptance, coherent self-identity, positive self-esteem, and future-oriented control rather than out of a concern for backward-looking causal accuracy. This conclusion derives from many different lines of evidence (for reviews, see Carruthers [2011]; Cushman [2018]).

For one, there is research from split-brain patients, who have had the connections between their brain hemispheres severed (often to mitigate seizures—and similar symptoms can be caused by stroke, tumours, or arterial ruptures). Such patients cannot integrate visual information obtained only by one hemisphere of the brain with verbal justifications generated by the other; as a result, an instruction (such as 'get up and walk') can be visually presented to the right hemisphere (via the left eye), causing the patients to initiate an appropriate behaviour (Gazzaniga [2000]). The patients can then be asked to explain their behaviour, and their left hemispheres (which are responsible for most of the linguistic processing) can use contextual information to produce justifications that are plausible but completely confabulated (such as 'I wanted to go into the kitchen to get a Coke'). For another, the phenomena of choice blindness further demonstrates that even neurotypical individuals can readily confabulate plausible justifications for choices they did not actually make, justifications that could not possibly be causally accurate but are indistinguishable from normal cases of introspection (Johansson *et al.* [2006]). In a choice blindness experimental design, subjects are asked to make a choice, then distracted, and finally given an option other than the one they actually selected. When they are then asked to justify having

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

selected this option, most subjects readily do so, often without any awareness that the item they were provided is different from the one they actually chose.

There are many other sources of evidence impugning the causal accuracy of human introspective justifications, and illustrating the readiness with which we confabulate when we lack causally accurate information. While there may be good reasons to demand more from deep learning systems than we could expect from humans, at present we are merely considering fair comparisons. In that respect, so long as we do not conflate explanatory and justificatory rationality, it does not seem that DNNs have a fundamental problem with interpretability that is not also exhibited by human minds.

5.4 Humans are also notorious reward-hackers

The final criticism to rebut is the concern that DNNs trained by reinforcement signals merely learn to 'reward hack' rather than learn real solutions to the problems on which they are trained. The response here is to note that humans are also notorious reward-hackers when placed in badly designed environments. One of the most obvious and directly comparable situations involves humans playing video games that are 'imbalanced' in their reward structure. This is a very common concern in online roleplaying games that offer many different routes to advance one's character. In these games, experience points and in-game currency are typically obtained by defeating foes or completing skilful actions. Game designers work exhaustively to create a homeostatic economy of experience and currency within the game; different methods to obtain these resources should all be perceived as roughly as difficult, time-consuming, and enjoyable as one another to support diverse routes to advance within the game. Proper balancing enhances playa-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

bility and perceived fairness, in order to keep players coming back for further character enhancement.

Humans, however, are highly adept at discovering the most efficient ways to obtain resources within a competitive game, and even slight imbalances will be found if present. These opportunities are often called 'exploits'. Game exploits are discussed and shared in online message boards, and tens of thousands of game players can quickly flock to repeating an exploit for days on end. Viewed with the same kind of detachment as the OpenAI's endlessly spinning boat, these human behaviours look just as pathological. One exploit in the game Fallout 4, for example, involves repeatedly building and disassembling tens of thousands of copper statues of a baseball player (which provides a small boost to experience) for twelve hours straight until they fill an entire abandoned town. For present purposes, the important point is that we do not conclude that these players have fundamentally misunderstood the point of their activities. Instead, we conclude that the game environment is badly designed, and the human players are ingenious at seeking out and taking advantage of these imbalances. The solution to an exploit is not to lecture players about lacking 'genuine' rationality or 'real' learning; it is to patch the game to change its reward structure, to restore balance to its reinforcement ecosystem—which can sometimes take teams of experienced programmers dozens of patches to achieve through trial-and-error.

Indeed, the difficulty of engineering artificial environments in which humans do not reward hack can make us wonder how the problem was ever solved in natural environments. The glib answer is that we did not solve it; natural selection solved it, by applying a trial-and-error approach to millions upon millions of our striving and starving ancestors. There is little reason to suppose that the product of this tinkering is a simple, transparent set of learning principles that could be captured in symbolic, rule-based form. It is, instead, a highly complex physical body

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

whose kinematics makes some motions more natural than others, a nervous system that reads its status in real time via a set of rich, multi-modal sensory inputs, a set of specialized sensory input organs that are more receptive to certain stimuli than others, and a highly constrained brain whose operation can be subtly modified by a symphony of hormones, neurotransmitters, and neuromodulators whose levels are dynamically controlled by these bodily inputs. Thus, rather than seeking out the optimal, intuitively satisfying Bayesian meta-learning rule, biologically inspired progress in reinforcement learning is more likely to be achieved by evolutionary search algorithms exploring combinations of bodily parameters for richer, more multi-dimensional reinforcement learning. DNN researchers should be trying to supplement models with additional and more multi-dimensional reward signals like fatigue, digestion, anxiety, surprise, tissue damage, emotional reactions, and social cues like accolade or embarrassment, rather than monolithic new learning rules or innate domain-specific knowledge.

Even worse, there is little reason to think that reinforcement learning in humans is as successful as anthropofabulation might have us believe. In sociology and psychology, there is an entire research area investigating the ways that simple, quantified evaluation systems distort human decision-making (Merry [2016]; Nguyen [2020]). Obvious examples occur when more natural reward systems—like food or social cues—become co-opted by new, more readily available stimuli. Simple examples of this phenomenon involve normally self-limiting reinforcers like sugar or alcohol suddenly becoming available in purer forms and unlimited amounts, leading to pathologically unhealthy behaviour. Subtler are the cases where reward policies are co-opted by entirely new kinds of stimuli that decouple the reward signal from the goals that evolution tweaked them to indicate. These hijackings can be good or bad; whether the ability of artificial sweeteners to decouple sweetness from caloric content is a good thing depends upon the balance

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

of relevant dietary science. Those working in this area, however, worry especially about cases where symbolic or numerical stimuli are used as proxies for more difficult-to-assess rewards. Examples of such proxies are endemic in modern life: credit scores, Fitbit counts, social media likes, grade point averages, h-index, university rankings, and so on. One does not need to look hard to find many examples where whole organizations or societies pathologically chase the maximization of reward proxies, often to the detriment of the more basic goals that they were initially designed to track.

In short, reward-hacking is not just some curious problem that confronts badly designed DNNs and their 'alien' ability to game a reward signal; it is a characteristically human pathology that plagues our own ability to play video games, succeed in business or academia, and generally not render the world unliveable. Anthropofabulation suggests that humans have some uncanny innate ability to flexibly pursue intrinsically valuable goals in highly diverse environments; but a fair appraisal of modern life would suggest that humanity is not currently doing so well at this particular balancing act. Perhaps for both humans and DNNs, the needed solution is to improve the structure of the environments in which we learn, rather than to fault the learning agents that seek solutions within them.

6 General Lessons

The goal of this article has been to advocate for fairer comparisons between DNN and human behaviour along four of the most popular criteria deployed by sceptics to argue that the kind of processing that occurs in DNNs is fundamentally different from human cognition, and explore general morals that can be applied to more productive future debates. Section 5 argued that unbiased assessments would score humans similarly to DNNs along all four criteria. The assumption

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

that humans are not vulnerable to these criticisms is not supported by empirical data, perhaps instead propped up by the bias of anthropofabulation. Where modelling human cognition is our goal, we should not aim to create DNNs that learn only from very few training examples without significant scaffolding, are immune to adversarial examples, use decision-making mechanisms that are completely transparent, or fail to exploit reward imbalances when placed in poorly designed environments.

Some unifying threads of the preceding discussion can now be drawn out as promising topics for future research. Critical discussions about artificial intelligence should feature more explicit reflection on how to properly align human and machine performance when conducting comparisons, especially by bringing in empirical research from human psychology, neuroscience, and biology. In particular, we need to be sure that our evaluation of human behaviour comes from a sceptical appraisal of empirical data, undistorted by the rose-tinted hue of anthropofabulation. In some cases, the relevant empirical work on humans remains inchoate; in particular, we need more research on the provenance of scaffolded learning in humans and on the implications of adversarial examples in perceptual psychology. And finally, we need more research on how to properly structure bodies and environments so as to obviate pathological reward-hacking behaviour in both humans and artificial agents. Regarding DNNs not as black boxes but rather as unflattering mirrors might help us accept hard lessons about ourselves, and in so doing take steps toward addressing some of the most pressing problems of our day.

Acknowledgements

I thank many for offering comments on previous versions of this work, especially Colin Allen, David Chalmers, Chaz Firestone, James Garson, Hans-Joachim Greif, Marta Halina, Paul Hum-

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

phries, Corey Maley, Carlos Zednik, three anonymous reviewers, and audiences at the University of Cambridge, the Technical University of Munich, the M. D. Anderson Cancer Center, the University of Virginia, the University of Cincinnati, the University of Vienna, New York University, OpenAI, and the Philosophy of Science Association. This research was kindly supported by a visiting fellowship from the Leverhulme Centre for the Future of Intelligence and by the National Science Foundation (grant no. 2020585).

Department of Philosophy
The University of Houston
Houston, USA
cjbuckner@uh.edu

References

Akagi, M. [2018]: 'Rethinking the Problem of Cognition', *Synthese*, **195**, pp. 3547–70.

Allen, C. [2006]: 'Ethics and the Science of Animal Minds', *Theoretical Medicine and Bioethics*, **27**, pp. 375–94.

Allen, C. [2017]: 'On (Not) Defining Cognition', *Synthese*, **194**, pp. 4233–49.

AlQuraishi, M. [2019]: 'AlphaFold at CASP13', *Bioinformatics*, **35**, pp. 4862–5.

Amodei, D. and Clark, J. [unpublished]: 'Faulty Reward Functions in the Wild', *OpenAI*, available at <openai.com/blog/faulty-reward-functions/>.

Andrews, K., Comstock, G., Crozier, G., Donaldson, S., Fenton, A., John, T., Johnson, L. S. M., Jones, R., Kymlicka, W. and Meynell, L. [unpublished]: 'The Philosophers' Brief on Chimpanzee Personhood', available at <philpapers.org/rec/ANDTPB-5>.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A. and Faulkner, R. [unpublished]: 'Relational Inductive Biases, Deep Learning, and Graph Networks', available at <arxiv.org/abs/1806.01261>.

Berwick, R. C. and Chomsky, N. [2017]: 'Why Only Us: Recent Questions and Answers', *Journal of Neurolinguistics*, **43**, pp. 166–77.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Block, N. [1981]: 'Psychologism and Behaviorism', *The Philosophical Review*, **90**, pp. 5–43.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D. and Hassabis, D. [2016]: 'Model-Free Episodic Control', available at <arxiv.org/abs/1606.04460>.

Boesch, C. [2007]: 'What Makes Us Human (*Homo Sapiens*)? The Challenge of Cognitive Cross-Species Comparison', *Journal of Comparative Psychology*, **121**, p. 227.

Botvinick, M., Barrett, D. G., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., Lillicrap, T., Modayil, J., Mohamed, S. and Rabinowitz, N. C. [2017]: 'Building Machines That Learn and Think for Themselves', *Behavioral and Brain Sciences*, **40**, p. 253.

Bringsjord, S. and Arkoudas, K. [2004]: 'The Modal Argument for Hypercomputing Minds', *Theoretical Computer Science*, **317**, pp. 167–90.

Bringsjord, S., Govindarajulu, N. S., Banerjee, S. and Hummel, J. [2018]: 'Do Machine-Learning Machines Learn?', in V. C. Müller (*ed.*), *Philosophy and Theory of Artificial Intelligence 2017*, Cham: Springer, pp. 136–57.

Brooks, R. A. [1991]: 'Intelligence without Representation', *Artificial Intelligence*, **47**, pp. 139–59.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A. *et al*. [unpublished]: 'Language Models Are Few-Shot Learners', available at <arxiv.org/abs/2005.14165>.

Buckner, C. [2013]: 'Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons', *Biology and Philosophy*, **28**, pp. 853–71.

Buckner, C. [2017]: 'Understanding Associative and Cognitive Explanations in Comparative Psychology', in K. Andrews and J. Beck (*eds*), *The Routledge Handbook of Philosophy of Animal Minds*, London: Routledge.

Buckner, C. [2018]: 'Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks', *Synthese*, **195**, pp. 5339–72.

Buckner, C. [2019a]: 'Deep Learning: A Philosophical Introduction', *Philosophy Compass*, **14**, p. e12625.

Buckner, C. [2019b]: 'Rational Inference: The Lowest Bounds', *Philosophy and Phenomenological Research*, **98**, pp. 697–724.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Buckner, C. [unpublished]: 'Adversarial Examples and the Deeper Riddle of Induction: The Need for a Theory of Artifacts in Deep Learning', available at <arxiv.org/abs/2003.11917>.

Call, J. and Tomasello, M. [2008]: 'Does the Chimpanzee Have a Theory of Mind? 30 Years Later', *Trends in Cognitive Sciences*, **12**, pp. 187–92.

Canaan, R., Salge, C., Togelius, J. and Nealen, A. [unpublished]: 'Leveling the Playing Field: Fairness in AI versus Human Game Benchmarks', available at <arxiv.org/abs/1903.07008>.

Carruthers, P. [2011]: 'The Opacity of Mind: An Integrative Theory of Self-knowledge', Oxford: Oxford University Press.

Chalmers, D. J. [1995]: 'Minds, Machines, and Mathematics', *Psyche*, **2**, pp. 117–18.

Clark, A. [1989]: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, MA: MIT Press.

Clark, A. [2003]: 'Artificial Intelligence and the Many Faces of Reason', in S. P. Stich and T. A. Warfield (*eds*), *The Blackwell Guide to the Philosophy of Mind*, Malden, MA: Blackwell, p. 309–21.

Crosby, M., Beyret, B. and Halina, M. [2019]: 'The Animal-AI Olympics', *Nature Machine Intelligence*, **1**, p. 257.

Crosby, M. [2020]: 'Building Thinking Machines by Solving Animal Cognition Tasks', *Minds and Machines*, **30**, pp. 589–615.

Cushman, F. [2018]: 'Rationalization Is Rational', *Behavioral and Brain Sciences*, **43**, p. 28.

Davis, M. [2004]: 'The Myth of Hypercomputation', in C. Teuscher (*ed.*), *Alan Turing: Life and Legacy of a Great Thinker*, Berlin: Springer, pp. 195–211.

de Waal, F. [2000]: 'Anthropomorphism and Anthropodenial: Consistency in Our Thinking about Humans and Other Animals', *Philosophical Topics*, **27**, pp. 255–80.

Ehsan, U., Harrison, B., Chan, L. and Riedl, M. O. [2018]: 'Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations', **2018**, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, pp. 81–87.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I. and Sohl-Dickstein, J. [2018]: 'Adversarial Examples That Fool Both Computer Vision and Time-Limited Humans', in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett (*eds*), *Advances in Neural Information Processing Systems*, **31**, pp. 3910–20.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Ericsson, K. A. and Simon, H. A. [1984]: *Protocol Analysis: Verbal Reports as Data*, Cambridge, MA: MIT Press.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D. [2018]: 'Robust Physical-World Attacks on Deep Learning Visual Classification', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2018**, pp. 1625–34.

Firestone, C. [forthcoming]: 'Performance vs. Competence in Human-AI Comparisons', *Proceedings of the National Academy of Sciences*.

Fitch, W. T. [2010]: 'The Evolution of Language', Cambridge: Cambridge University Press.

Fitch, W. T. [2014]: 'Toward a Computational Framework for Cognitive Biology: Unifying Approaches from Cognitive Neuroscience and Comparative Cognition', *Physics of Life Reviews*, **11**, pp. 329–64.

Flavell, J. H., Flavell, E. R. and Green, F. L. [1983]: 'Development of the Appearance-Reality Distinction', *Cognitive Psychology*, **15**, pp. 95–120.

Fodor, J. A. and Pylyshyn, Z. W. [1988]: 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition*, **28**, pp. 3–71.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M. and Correia, B. E. [2019]: 'Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning', *Nature Methods*, pp. 1–9.

Gazzaniga, M. S. [2000]: 'Cerebral Specialization and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?', *Brain*, **123**, pp. 1293–326.

Goodfellow, I., Shlens, J. and Szegedy, C. [unpublished]: 'Explaining and Harnessing Adversarial Examples', available at <arxiv.org/abs/1412.6572>.

Goodman, B. and Flaxman, S. [2017]: 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"', *AI Magazine*, **38**, pp. 50–57.

Govindarajulu, N. S. and Bringsjord, S. [2012]: 'The Myth of "The Myth of Hypercomputation"', *Parallel Processing Letters*, **22**, p. 1240012.

Guest, O. and Love, B. [unpublished]: 'Levels of Representation in a Deep Learning Model of Categorization', available at <www.biorxiv.org/content/10.1101/626374v1>.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. [2019]: 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys*, **51**, p. 93.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S. and Redish, A. D. [2010]: 'Hippocampal Replay Is Not a Simple Function of Experience', *Neuron*, **65**, pp. 695–705.

Hassabis, D., Kumaran, D., Summerfield, C. and Botvinick, M. [2017]: 'Neuroscience-Inspired Artificial Intelligence', *Neuron*, **95**, pp. 245–58.

Hernández-Orallo, J. [2017]: 'The Measure of All Minds: Evaluating Natural and Artificial Intelligence', Cambridge: Cambridge University Press.

Hespos, S. J. and VanMarle, K. [2012]: 'Physics for Infants: Characterizing the Origins of Knowledge about Objects, Substances, and Number', *Wiley Interdisciplinary Reviews: Cognitive Science*, **3**, pp. 19–27.

Hinton, G. E. and Salakhutdinov, R. R. [2006]: 'Reducing the Dimensionality of Data with Neural Networks', *Science*, **313**, pp. 504–07.

Hochreiter, S. and Schmidhuber, J. [1996]: 'LSTM Can Solve Hard Long Time Lag Problems', *Advances in Neural Information Processing Systems*, **1996**, pp. 473–79.

Hofstadter, D. R. [1985]: 'Waking up from the Boolean Dream, or Subcognition as Computation', *Metamagical Themas: Questing for the Essence of Mind and Pattern*, pp. 631–65.

Hong, H., Yamins, D. L., Majaj, N. J. and DiCarlo, J. J. [2016]: 'Explicit Information for Category-Orthogonal Object Properties Increases along the Ventral Stream', *Nature Neuroscience*, **19**, pp. 613–22.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A. [unpublished]: 'Adversarial Examples Are Not Bugs, They Are Features', available at <arxiv.org/abs/1905.02175>.

Irpan, A. [unpublished]: 'Deep Reinforcement Learning Doesn't Work Yet', *Sorta Insightful*, available at <www.alexirpan.com/2018/02/14/rl-hard.html>.

Johansson, P., Hall, L., Sikström, S., Tärning, B. and Lind, A. [2006]: 'How Something Can Be Said about Telling More than We Can Know: On Choice Blindness and Introspection', *Consciousness and Cognition*, **15**, pp. 673–92.

Kahneman, D. and Frederick, S. [2002]: 'Representativeness Revisited: Attribute Substitution in Intuitive Judgment', in T. Gilovich, D. Griffin and D. Kahneman (*eds*), *Heuristics and Biases: The Psychology of Intuitive Judgment*, pp. 49–81.

Karin-D'Arcy, M. [2005]: 'The Modern Role of Morgan's Canon in Comparative Psychology', *International Journal of Comparative Psychology*, **18**, p. 3.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Keeley, B. [2004]: 'Anthropomorphism, Primatomorphism, Mammalomorphism: Understanding Cross-Species Comparisons', *Biology and Philosophy*, **19**, pp. 521–40.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. [2014]: 'Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation', *PLoS Computational Biology*, **10**, p. e1003915.

Kingma, D. P. and Welling, M. [unpublished]: 'Auto-encoding Variational Bayes', available at <arxiv.org/abs/1312.6114>.

Krachun, C., Lurz, R., Russell, J. L. and Hopkins, W. D. [2016]: 'Smoke and Mirrors: Testing the Scope of Chimpanzees' Appearance–Reality Understanding', *Cognition*, **150**, pp. 53–67.

Kubilius, J., Bracci, S. and Beeck, H. P. O. de [2016]: 'Deep Neural Networks as a Computational Model for Human Shape Sensitivity', *PLOS Computational Biology*, **12**, p. e1004896.

Kunst-Wilson, W. R. and Zajonc, R. B. [1980]: 'Affective Discrimination of Stimuli That Cannot Be Recognized', *Science*, **207**, pp. 557–58.

Kusner, M. J., Paige, B. and Hernández-Lobato, J. M. [unpublished]: 'Grammar Variational Autoencoder', available at <arxiv.org/abs/1703.01925>.

Lake, B. M. [2014]: 'Towards More Human-like Concept Learning in Machines: Compositionality, Causality, and Learning-to-Learn', PhD Thesis, Massachusetts Institute of Technology.

Lake, B. M., Zaremba, W., Fergus, R. and Gureckis, T. M. [2015a]: 'Deep Neural Networks Predict Category Typicality Ratings for Images', *Proceedings of the 37th Annual Cognitive Science Society*, **2015**, available at <cims.nyu.edu/~brenden/LakeZarembaEtAl2015CogSci.pdf>.

Lake, B. M., Salakhutdinov, R. and Tenenbaum, J. B. [2015b]: 'Human-Level Concept Learning through Probabilistic Program Induction', *Science*, **350**, pp. 1332–38.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B. and Gershman, S. J. [2017a]: 'Building Machines That Learn and Think like People', *Behavioral and Brain Sciences*, **40**, p. 253.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B. and Gershman, S. J. [2017b]: 'Ingredients of Intelligence: From Classic Debates to an Engineering Roadmap', *Behavioral and Brain Sciences*, **40**, p. 281.

Lake, B. M. [2019]: 'Compositional Generalization through Meta Sequence-to-Sequence Learning', *Advances in Neural Information Processing Systems*, **32**, pp. 9791–801.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Landy, D., Allen, C. and Zednik, C. [2014]: 'A Perceptual Account of Symbolic Reasoning', *Frontiers in Psychology*, **5**, p. 275.

Lassiter, G. D., Geers, A. L., Munhall, P. J., Ploutz-Snyder, R. J. and Breitenbecher, D. L. [2002]: 'Illusory Causation: Why It Occurs', *Psychological Science*, **13**, pp. 299–305.

LeCun, Y., Bengio, Y. and Hinton, G. [2015]: 'Deep Learning', *Nature*, **521**, pp. 436–44.

LeCun, Y. [2018]: 'The Power and Limits of Deep Learning', *Research-Technology Management*, **61**, pp. 22–7.

Lipton, Z. C. [unpublished]: 'The Mythos of Model Interpretability', available at <arxiv.org/abs/1606.03490>.

Lotter, W., Kreiman, G. and Cox, D. [unpublished]: 'Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning', available at <arxiv.org/abs/1605.08104>.

Luc, P., Neverova, N., Couprie, C., Verbeek, J. and LeCun, Y. [2017]: 'Predicting Deeper into the Future of Semantic Segmentation', *Proceedings of the IEEE International Conference on Computer Vision*, **2017**, pp. 648–57.

Lyre, H. [2020]: 'The State Space of Artificial Intelligence', *Minds and Machines*, **30**, pp. 325–47.

Marcus, G. [2018]: 'Deep Learning: A Critical Appraisal', available at <arxiv.org/abs/1801.00631>.

Matthews, R. J. [1994]: 'Three-Concept Monte: Explanation, Implementation, and Systematicity', *Synthese*, **101**, pp. 347–63.

McClelland, J. L., Rumelhart, D. E. and Group, P. R. [1986]: 'Parallel Distributed Processing', *Explorations in the Microstructure of Cognition*, **2**, pp. 216–71.

Merry, S. E. [2016]: *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*, Chicago, IL: University of Chicago Press.

Miracchi, L. [2019]: 'A Competence Framework for Artificial Intelligence Research', *Philosophical Psychology*, 32, pp. 588–633.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K. and Ostrovski, G. [2015]: 'Human-Level Control through Deep Reinforcement Learning', *Nature*, **518**, p. 529.

Moore, R. [2017]: 'The Evolution of Syntactic Structure', *Biology and Philosophy*, **32**, pp. 599–613.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O. and Frossard, P. [2017]: 'Universal Adversarial Perturbations', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2017**, pp. 86–94.

Mordvintsev, A., Olah, C. and Tyka, M. [unpublished]: 'Inceptionism: Going Deeper into Neural Networks', *Google AI Blog*, available at <ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>.

Murphy, S. T. and Zajonc, R. B. [1993]: 'Affect, Cognition, and Awareness: Affective Priming with Optimal and Suboptimal Stimulus Exposures', *Journal of Personality and Social Psychology*, **64**, p. 723.

Newell, A. and Simon, H. A. [1976]: 'Computer Science as Empirical Inquiry: Symbols and Search', *Communications of the ACM*, **19**, pp. 113–26.

Nguyen, A., Yosinski, J. and Clune, J. [2015]: 'Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images', *IEEE Conference on Computer Vision and Pattern Recognition*, **2015**, pp. 427–36.

Nguyen, T. [2020]: *Games: Agency as Art*, Oxford: Oxford University Press.

Nisbett, R. and Ross, L. [1980]: *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.

Orhan, A. E., Gupta, V. V. and Lake, B. M. [unpublished]: 'Self-supervised Learning through the Eyes of a Child', available at <arxiv.org/abs/2007.16189>.

Pearl, J. [2019]: 'The Seven Tools of Causal Inference, with Reflections on Machine Learning', *Communications of the ACM*, **62**, pp. 54–60.

Penn, D. C. and Povinelli, D. J. [2007a]: 'Causal Cognition in Human and Nonhuman Animals: A Comparative, Critical Review', *Annual Review of Psychology*, **58**, pp. 97–118.

Penn, D. C. and Povinelli, D. J. [2007b]: 'On the Lack of Evidence That Non-human Animals Possess Anything Remotely Resembling a "Theory of Mind"', *Philosophical Transactions of the Royal Society of London: Series B*, **362**, pp. 731–44.

Plaut, D. C. and McClelland, J. L. [2010]: 'Locating Object Knowledge in the Brain: Comment on Bowers's (2009) Attempt to Revive the Grandmother Cell Hypothesis', *Psychological Review*, **117**, pp. 284–88.

Polger, T. W. and Shapiro, L. A. [2016]: *The Multiple Realization Book*, Oxford: Oxford University Press.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G. and Livingstone, M. S. [2019]: 'Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences', *Cell*, **177**, pp. 999–1009.

Proudfoot, D. [2011]: 'Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game', *Artificial Intelligence*, **175**, pp. 950–57.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B. and Wellman, M. [2019]: 'Machine Behaviour', *Nature*, **568**, pp. 477–86.

Ramsey, W. [2017]: 'Must Cognition Be Representational?', *Synthese*, **194**, pp. 4197–214.

Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K. and Wierstra, D. [2016]: 'One-Shot Generalization in Deep Generative Models', available at <arxiv.org/abs/1603.05106>.

Rogers, T. T. and McClelland, J. L. [2014]: 'Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition', *Cognitive Science*, **38**, pp. 1024–77.

Russin, J., Jo, J., O'Reilly, R. C. and Bengio, Y. [unpublished]: 'Compositional Generalization in a Deep Seq2seq Model by Separating Syntax and Semantics', available at <arxiv.org/abs/1904.09708>.

Samuels, R., Stich, S. and Bishop, M. [2002]: 'Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear', in R. Elio (*ed.*), *Common Sense, Reasoning, and Rationality*, Oxford: Oxford University Press, pp. 236–68.

Schacter, D. L. [1987]: 'Implicit Memory: History and Current Status', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **13**, p. 501.

Schmidhuber, J. [2015]: 'Deep Learning in Neural Networks: An Overview', *Neural Networks*, **61**, pp. 85–117.

Scott-Phillips, T. C., Cartmill, E. A., Crockford, C., Gärdenfors, P., Gómez, J. C., Luef, E. M., Pika, S., Moore, R., Rendall, D., Townsend, S. W. and others [2015]: 'Nonhuman Primate Communication, Pragmatics, and the Origins of Language', *Current Anthropology*, **56**, pp. 58–80.

Serre, T. [2019]: 'Deep Learning: The Good, the Bad, and the Ugly', *Annual Review of Vision Science*, **5**, pp. 399–426.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Shevlin, H. and Halina, M. [2019]: 'Apply Rich Psychological Terms in AI with Care', *Nature Machine Intelligence*, **1**, p. 165.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D. and Graepel, T. [2018]: 'A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-play', *Science*, **362**, pp. 1140–44.

Smith, B. C. [2019]: *The Promise of Artificial Intelligence: Reckoning and Judgment*, Cambridge, MA: MIT Press.

Smolensky, P. [1988]: 'On the Proper Treatment of Connectionism', in A. Goldman (*ed.*), *Behavioral and Brain Sciences*, **11**, pp. 1–74.

Sober, E. [1998]: 'Morgan's Canon', in D. Cummins and C. Allen (*eds*), *The Evolution of Mind*, New York: Oxford University Press, pp. 224–42.

Socher, R., Ganjoo, M., Manning, C. D. and Ng, A. [2013]: 'Zero-Shot Learning through Cross-Modal Transfer', *Advances in Neural Information Processing Systems*, **26**, pp. 935–43.

Stickgold, R. [2005]: 'Sleep-Dependent Memory Consolidation', *Nature*, **437**, p. 1272.

Stinson, C. [2020]: 'From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence', *Philosophy of Science*, **87**, pp. 590–611.

Su, J., Vargas, D. V. and Sakurai, K. [2019]: 'One Pixel Attack for Fooling Deep Neural Networks', *IEEE Transactions on Evolutionary Computation*, **25**, pp. 828–41.

Sutton, R. S. and Barto, A. G. [2018]: *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press.

Turek, M. [unpublished]: 'Explainable Artificial Intelligence', available at <www.darpa.mil/program/explainable-artificial-intelligence>.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. [2017]: 'Attention Is All You Need', *Advances in Neural Information Processing Systems*, **30**, pp. 5998–6008.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. and Wierstra, D. [2016]: 'Matching Networks for One Shot Learning', in D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (*eds*), *Advances in Neural Information Processing Systems*, **29**, pp. 3630–38.

Walker, M. P. and Stickgold, R. [2010]: 'Overnight Alchemy: Sleep-Dependent Memory Evolution', *Nature Reviews: Neuroscience*, **11**, p. 218.

This is the author's accepted manuscript without copyediting, formatting, or final corrections. It will be published in its final form in an upcoming issue of The British Journal for the Philosophy of Science, published by The University of Chicago Press on behalf of The British Society for the Philosophy of Science. Include the DOI when citing or quoting: https://doi.org/10.1086/714960 Copyright 2021 The British Society for the Philosophy of Science.

Watson, D. [2019]: 'The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence', *Minds and Machines*, **29**, pp. 417–40.

Winsberg, E. [2010]: 'Science in the Age of Computer Simulation', Chicago, IL: University of Chicago Press.

Wynne, C. D. L. [2004]: 'The Perils of Anthropomorphism', *Nature*, **428**, p. 606.

Xu, W., Evans, D. and Qi, Y. [unpublished]: 'Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks', available at <arxiv.org/abs/1704.01155>.

Yamins, D. L. and DiCarlo, J. J. [2016]: 'Using Goal-Driven Deep Learning Models to Understand Sensory Cortex', *Nature Neuroscience*, **19**, p. 356.

Zador, A. M. [2019]: 'A Critique of Pure Learning and What Artificial Neural Networks Can Learn from Animal Brains', *Nature Communications*, **10**, pp. 1–7.

Zednik, C. [2019]: 'Solving the Black Box Problem: A General-Purpose Recipe for Explainable Artificial Intelligence', *Philosophy and Technology*, pp. 1–24.

Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. [2019]: 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?', *Philosophy and Technology*, **32**, pp. 661–83.

Zhang, M., Jiang, S., Cui, Z., Garnett, R. and Chen, Y. [2019]: 'D-VAE: A Variational Autoencoder for Directed Acyclic Graphs', *Advances in Neural Information Processing Systems*, **32**, pp. 1588–600.

Zhou, Z. and Firestone, C. [2019]: 'Humans Can Decipher Adversarial Images', *Nature Communications*, **10**, p. 1334.

Zhu, S., Ng, I. and Chen, Z. [unpublished]: 'Causal Discovery with Reinforcement Learning', available at <arxiv.org/abs/1906.04477>.

Złotowski, J., Proudfoot, D., Yogeeswaran, K. and Bartneck, C. [2015]: 'Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction', *International Journal of Social Robotics*, **7**, pp. 347–60.