The Comparative Psychology of Artificial Intelligences

Draft Status

Cameron Buckner

1. Introduction

The last five years have seen a series of remarkable achievements in Artificial Intelligence (AI) research. For example, systems based on Deep Neural Networks (DNNs) can now classify natural images as well or better than humans, defeat human grandmasters in strategy games as complex as chess, Go, or Starcraft II, and navigate autonomous vehicles across thousands of miles of mixed terrain. Like the brain, however, these models have become staggeringly complex, and prominent neural network researchers have suggested that we might come to understand their processing by engaging them with experimental paradigms and data analysis methods derived from human and animal cognitive sciences.

Such engagement has led researchers to conclude that DNNs are currently the best artificial model of perceptual similarity judgments in primates, based on comparisons between model behavior and primate judgments of the similarity of exemplars and electrophysiology measurements obtained from ventral stream visual cortex (Guest and Love 2019; Hong et al. 2016; Khaligh-Razavi and Kriegeskorte 2014; Kubilius, Bracci, and Beeck 2016; Lake et al. 2015; Yamins and DiCarlo 2016). Taking the idea that neural networks can be engaged directly with the tools of psychology even further, the "Animal-AI Olympics" recently proposed by the Leverhulme Centre for the Future of Intelligence has created a testbed application that encodes dozens of benchmark experimental measures of cognitive performance in animals, so that these tests can be systematically applied to artificial systems to determine whether they exhibit a range of animal-level intelligent behaviors (Crosby, Beyret, and Halina 2019).

In short, comparisons between natural and artificial intelligences have never been so varied and ambitious—nor, I shall argue here, so fraught. The capacity of these artificial systems to produce new forms of potentially intelligent behavior has outpaced our reflection on whether these comparisons are fair or meaningful. Aligning the operations of differently-constituted natural intelligences is already a tricky enterprise; even different biological species have different perceptual abilities, motivations, motor capacities,

and neural hardware, so we cannot expect to give them identical psychological tests. Moreover, biases like anthropomorphism and anthropocentrism plague human intuitions about nonhuman behaviors, and methodological subtlety is required to temper them. These difficulties are only exacerbated when the other end of the comparison is an artificial system, for such systems often only aim to reproduce parts of a full cognitive agent. In his defense of his famous imitation game test, Turing himself wrestled with these issues; and commentators have explicitly reflected on how to avoid being unwittingly convinced by artificial systems that present the superficial trappings of human-like behavior (such as human-like facial expressions or gestures) without the same underlying competences (Block 1981; Proudfoot 2011; Shevlin and Halina 2019; Złotowski et al. 2015).

In this paper, I suggest that this debate about fair comparisons in AI could be expedited by taking the lead from 120 years of reflection on similar questions in comparative psychology. While comparative psychology dedicated much effort to developing rigorous empirical methods to avoid anthropomorphism-driven false positives, the field is also recently coming to grips with the danger of anthropocentrism-driven false negatives. In AI, a great deal of critical thought has similarly been put into the skeptical evaluation of artificial system performance, but very little of that critical skepticism has been directed at the selection and scoring of the purported human equivalent. Here, I review three ways in which critics have alleged that DNNs underperform humans and animals. I will argue that a bias called "anthropofabulation" (Buckner 2013)—which scores nonhuman performance against an inflated conception of human competence—threatens the validity of the comparison. I will end by suggesting some lines of research that comparative AI must explore to develop fair and rigorous protocols for such comparisons.

2. Three Criticisms of Deep Learning Research

In this section, I canvass three criticisms that have been commonly offered against inflationary interpretations of DNNs, which can be found in influential critical assessments such as Lake et al. (2017a) and/or Marcus (2018).

A. Deep Learning Is Too Data-Hungry

One of the most common critical refrains is that DNNs require much more training data than humans to produce equivalent performance. It is true that most state-of-the-art DNNs require millions of training exemplars to reach their benchmark 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); its human opponent Lee Sedol, by contrast, 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 seems to allow humans and animals to learn from a single exemplar. For example, Lake et al. (2015) argue that humans can learn to recognize and draw the components of new handwritten characters, even from just a single example (Fig 1.). Skeptics thus wonder whether DNNs will ever be able to learn comparatively rich category information from smaller, more human-like amounts of experience.

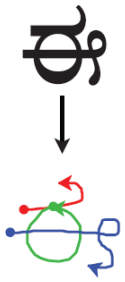

Fig. 1 The hierarchical decomposition of a novel handwritten figure decomposed into three individual pen strokes, which humans can purportedly learn from a single exemplar (reproduced from Lake et al. 2015).

B. Adversarial Examples Expose Deep Learning as a Fraud

Adversarial examples were originally defined as "perturbed images" created by slightly modifying an easily-classifiable exemplar in a way that was imperceptible to humans, but which could cause dramatic misclassification by DNNs (Goodfellow, Shlens, & Szegedy, 2015—Fig. 2). A related line of research has discovered numerous forms of "rubbish images" that are supposedly meaningless to humans but are confidently classified as members of particular categories by DNNs (Nguyen, Yosinski, and Clune 2015).

Subsequent research has found that such adversarial examples have 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 robust property of current DNN methods.

Much of the interest of 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 them 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—and human observers might not know that anything was awry until it was too late. For modeling 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, Shlens, & Szegedy, 2015).

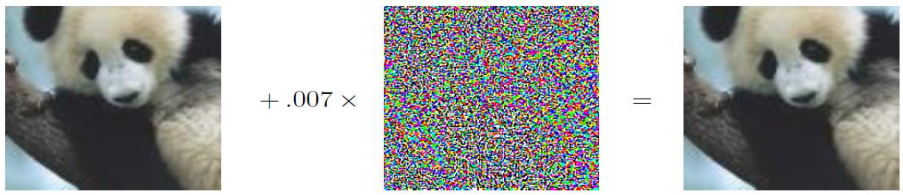

Figure 2. An adversarial "perturbed image," reproduced from Goodfellow et al. (2014). After the "panda" image was modified slightly by the addition of a small noise vector (itself classified with low confidence as a nematode), it was classified as a gibbon with high confidence, despite the modification being imperceptible to humans.

C. DNNs Are Not Interpretable

Another common criticism holds that DNNs are "black boxes" which are not "interpretable" (Lipton 2016) or not "sufficiently transparent" (Marcus 2018). State-of-the-art DNNs can contain hundreds of layers and millions of individual parameters, making it difficult to understand the significance of their internal processing. However, much in this charge remains underspecified (Zednik 2019). What kind of

interpretability needs to be provided, and to whom? 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—said 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 the EU's General Data Protection Regulation (which provides users with a "right to explanation" for decisions made by algorithms) have quickened the challenge and provided it with some practical goals, if not always conceptual clarity (Goodman & Flaxman, 2017; Gunning, 2017).

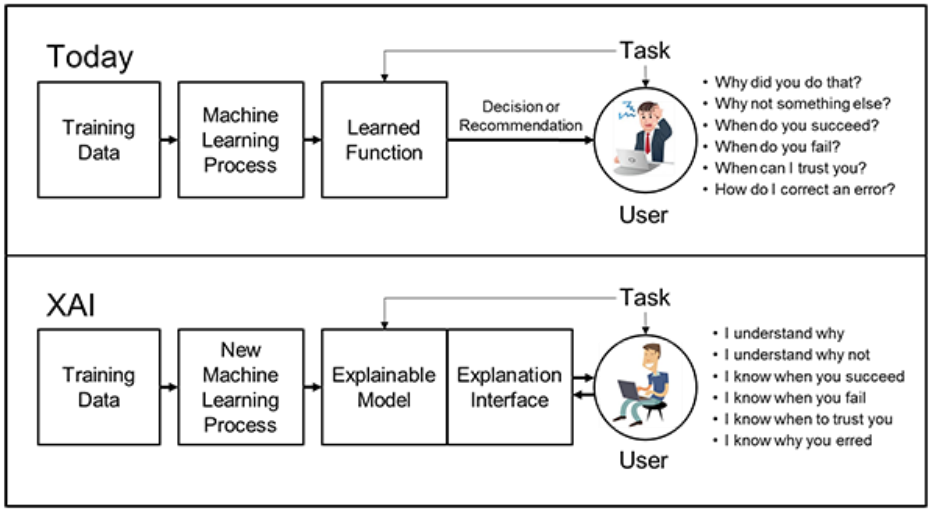

Figure 3. The DARPA XAI concept (from Gunning 2017).

3. A Crash Course in Comparative Bias

Comparative psychology has come to appreciate that human researchers are vulnerable to systematic biases when comparing human and nonhuman intelligence. One bias which has already been well-studied by philosophy of science is anthropomorphism. Understood as an error, anthropomorphism is the tendency of humans to attribute human-like psychological characteristics to nonhumans on the basis of insufficient empirical evidence. In comparative psychology, a methodological corrective that has been recommended since the discipline's inception is Morgan's Canon, which states that "in no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower on the scale of psychological evolution and development" (Morgan 1903). A sizeable

literature in comparative psychology explores the Canon's proper interpretation (Karin-D'Arcy 2005; Sober 1998). In practice, the set of "lower" processes usually includes reflexes, instincts, and simple associative learning mechanisms, and the goal of an experiment is to devise a situation which could be solved only by using cognitive information-processing or rational insight, and not by the use of any of these less-flexible alternatives (Buckner 2017b).

On the other hand, there are also a variety of anthropocentric biases which can push down the scales against nonhumans. For example, semantic anthropocentrism is the tendency to tie the criteria for the possession of some cognitive capacity to a distinctively human expression of the trait, without adequate theoretical justification. In other words, it causes us to assume that only the distinctive way humans exhibit a trait is valuable or worthwhile—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. Even more pernicious, however, is the bias of "anthropofabulation" (Buckner 2013), which combines semantic anthropocentrism with confabulation about human cognitive performance. Anthropofabulation results from an uncritical and inflated picture of human cognitive processing derived from introspection or cultural traditions. Common sense tells us that our thought processes are always rational—derived from a dispassionate processing of the situation, a direct introspective access to our true beliefs and motivations, and independence from subtle environmental scaffolding, historical associations, or emotional reactions. A great deal of human social psychology and philosophy of psychology, however, has cast this picture of human cognition into doubt (Carruthers 2011; Kahneman and Frederick 2002; Nisbett and Ross 1980; Samuels, Stich, and Bishop 2002).

In practice, anthropofabulation has caused skeptics to compare human and animal performance in situations which 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 causes us to implicitly assume that human performance could not possibly depend upon such environmental scaffolding, and leading us to

trivialize these disanalogies. The worry is that critics of DNNs have similarly drawn upon the rosy picture of human cognition presented to us by introspection and common sense, testing humans and DNNs in unfairly disanalogous situations or assessing penalties to DNNs for factors that apply equally to rational human cognition.

4. Three Rebuttals

A. 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. Taking these factors into account, however, would score orders of magnitude more equivalently trainable exemplars to this prolonged interaction.

It is difficult to decide exactly which features of human perceptual learning are relevant to the comparison, but we can review some results in the neighborhood. Studies of motion-picture perception has suggested that human vision has a frame rate of about 10-12 images per second (below which we cannot perceive motion as continuous). While it might take 200-400 ms for us to become consciously aware of a perceptual stimulus, attentional shifting to a new stimulus begins in as little as 20 ms, and category structure can be implicitly influenced by nonconscious exposures to stimuli as brief as 1 ms (Kunst-Wilson and Zajonc 1980; Schacter 1987). Moreover, perceptual memories may be repeatedly reconsolidated by theta rhythm 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 cortex on novel experiences synthesized from combinations or transformations of previous training information—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.

Neither is this merely idle nit-picking; neural network models that attempt to replicate these aspects of human learning can make much 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 or positions of the same object can be treated as thousands of independent training instances (Luc et al. 2017). When models are supplemented with "episodic replay" buffers that are inspired by declarative memory faculties in mammals, a DNN's performance can continue to benefit from repeatedly replaying exposure to the same training instances numerous times (Blundell et al. 2016; Vinyals et al. 2016; Mnih et al. 2015). Predictive, "self-supervised" models—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 seen as the future of the field by DNN pioneers like LeCun (2018). There is little evidence that the efficiency benefits from such biologically-inspired innovations of this sort have already plateaued.

Still, critics hold that this all falls short of the kind of one-shot learning of e.g. novel digits and their construction emphasized by some critics, which has purportedly been modeled by Bayesian systems. While numerous DNN systems produce one-shot or even zero-shot learning on related tasks, (Socher et al. 2013; Rezende et al. 2016), Bayesian modelers complain about their need for extensive pre-training. Nevertheless, there remain significant questions about the fairness of this comparison. Humans are capable of such one-shot learning only after extensive practice in recognizing and generating a variety of different handwritten figures. The Bayesian models which are purported to replicate one-shot learning, moreover, must incorporate significant amounts of high-level knowledge and representational structures that are manually-encoded by their programmers (Botvinick et al. 2017). These Bayesian modelers 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, 53). However, until the cognitive provenance of this knowledge—

specifically, the nature and number of training exposures adult humans require to scaffold such one-shot learning— is accounted for in humans, these concerns cannot fairly be scored against DNNs in this debate.

B. We cannot assume that DNNs' verdicts on adversarial examples are incorrect

Recent investigations have challenged the assumption that a DNN's take on adversarial examples is really so alien to human perception, either by producing perturbed images that fool humans (Elsayed et al., 2018--Fig. 4) or by showing that humans can easily "adopt the machine perspective" and, when forced to choose between a preset list of candidate labels, predict a DNN's labels for rubbish images with high accuracy (Zhou and Firestone 2019--and see Fig. 5). These authors suggest that the behavior of DNNs in these cases might be due to the fact that they are always forced to choose amongst a list of candidate labels, even when images are very different from previously-classified exemplars. This is a crucial disanalogy in many comparisons between natural and artificial judgments on adversarial examples which invites anthropofabulation. Specifically, DNNs do appear to capture some aspects of lower-level perceptual categorization in human, such that many rubbish images do *look like* members of the purportedly incorrect label class in some sense, even if humans do not ultimately think that they *actually are* members of that class. DNNs may thus be correctly delivering human perceptual similarity judgments, but not yet have the resources to draw a distinction between what something *looks like* and what it *really is*. Perhaps it remains an open question whether and how to model the latter kind of judgment in DNNs, but the comparison would not yet demonstrate that a DNN's processing is hopelessly alien to human perception. In the terminology of dual system theory (Kahneman and Frederick 2002), we would be comparing a System I intuitive judgments from DNNs to a System II reflective judgment in humans—another improper alignment for comparison.

Even more recently, commentators have begun explicitly calling out the anthropocentrism of the literature on adversarial examples, suggesting that vulnerability to adversarial examples may be a feature and not a bug of DNNs (Ilyas et al. 2019). These authors note that while it is sometimes possible to modify training sets so that they do not contain the features that cause DNNs and humans to classify perturbed images differently, these features generalize well and are predictively- valid in natural data sets, and humans may only fail to detect them due to inferior perceptual acuity. While this does not diminish the practical

significance of the phenomenon (because it can be exploited by malicious agents), it raises philosophical questions as to which features ought to be relevant to assessing intelligence in categorization tasks. Should AIs only detect features that are "projectible" and "natural" for humans due to idiosyncratic features of our perceptual endowment or evolutionary history (Goodman, 1955; Quine, 1969)? Or should they respond to any features which are objectively valid and predictively useful? When the latter can produce more accurate results on natural images, it is difficult to justify the assessment that only the former are relevant to intelligence without simply relying on flat-footed anthropocentrism.

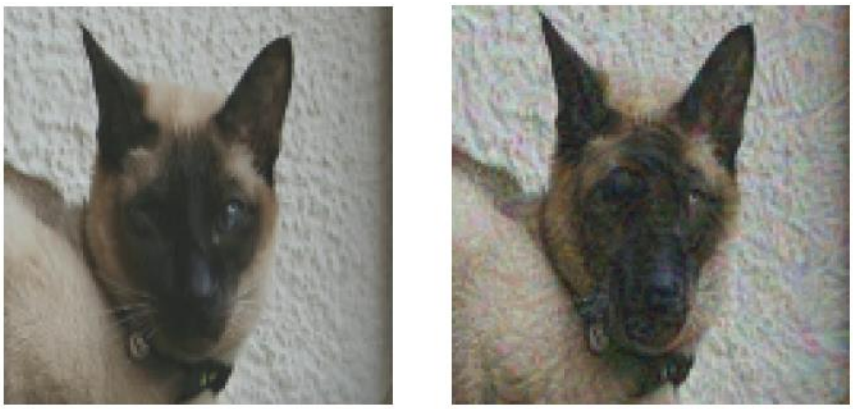

Fig 4. A perturbed image that can purportedly fool human subjects, with the original image of a cat on the left, and the perturbed image (often classified as a dog) on the right. (Image reproduced from Elsayed et al. 2018).

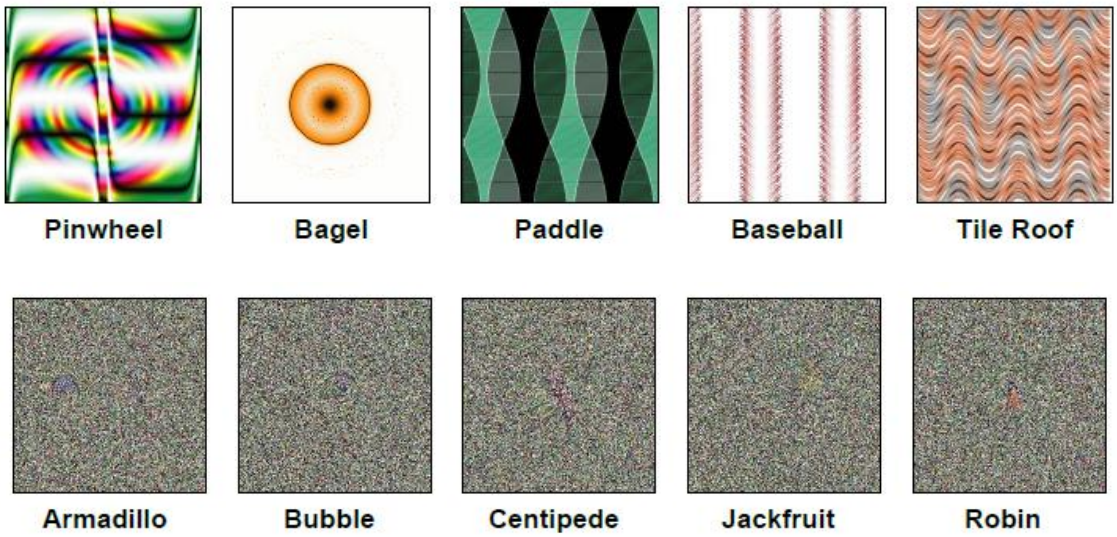

Figure 5. Examples of two different types of rubbish images tested by Zhou & Firestone (2019) with preferred DNN labels. In a forced-choice task, humans were able to guess a DNN's preferred labels for these images with high accuracy.

C. Human decision-making is also opaque

As noted above and in several critical analyses, the interpretability challenge conflates several different concerns which are probably best separated. To make a start at disentangling them, I suggest that a distinction between explanatory rationality and justificatory rationality may be useful here (Buckner 2017a). Explanatory rationality concerns a causal explanation of agent's internal reasons for acting—explaining *which factors* caused the agent to produce the output that it did in that situation, rather than some other output. 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, concerns 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?" and "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; but anthropofabulation causes us to conflate them, because common sense supposes that the justifications we produce through introspection have direct, non-inferential access to the causal antecedents which produced the behaviors so justified. However, a significant amount of cognitive science suggests that this picture is mistaken. Once the two dimensions of rational evaluation are separated, humans and DNNs compare favorably on interpretability.

To provide some examples, one of the reasons that people have supposed that the internal processing of DNNs is opaque is that the visualization methods which have deployed to determine the representational functions of internal processing nodes in DNNs have sometimes produced strange, chimerical images. Activity maximization is one popular method, which relies on tweaking images using further machine learning until they maximally activate some particular node in a DNN's internal layers. This is supposed to show us what 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, Olah, and Tyka 2015--Fig 6).

However, activity maximization is a new visualization technique that is very unlike introspection or metacognition in humans; directly comparing introspectible features to its results is like comparing apples to resequenced orange DNA. There is little reason to think that the results of the two should be similar; and 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--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 satisfying justifications.

Figure 6. Results of running an activity maximization algorithm on a picture of clouds in a trained-up version of Googles Inception image-classifying DNN. 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, from Ponce et al. (2019).

On the side of justificatory rationality, methods have been designed to generate intuitively satisfying justifications for DNN behavior, but they have been criticized for failing to highlight causally-determinative factors. 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 when a DNN made similar decisions in similar circumstances (Ehsan et al. 2018--Fig 8). The system can then deliver those verbal justifications to human interpreters to support its decisions while they are being made, and human subjects find those rationalizations more satisfying than more causally-accurate alternatives. The authors conceded that there is no direct causal link in this case between the features which actually caused the system to make the decision and the features cited in the verbal justification. However, they also note that social psychology often finds a disconnect between human rationalizations and the factors which actually cause human actions. In fact, the best empirical theories of these systems in humans construe them as interpretive and inferential, generated mainly to promote social acceptance, coherent self-identity, and positive self-esteem rather than out of a concern for causal accuracy (Carruthers 2011). In short, so long as we do not conflate explanatory and justificatory rationality, it is less clear that DNNs have a fundamental problem with interpretability that is not also exhibited by human minds.

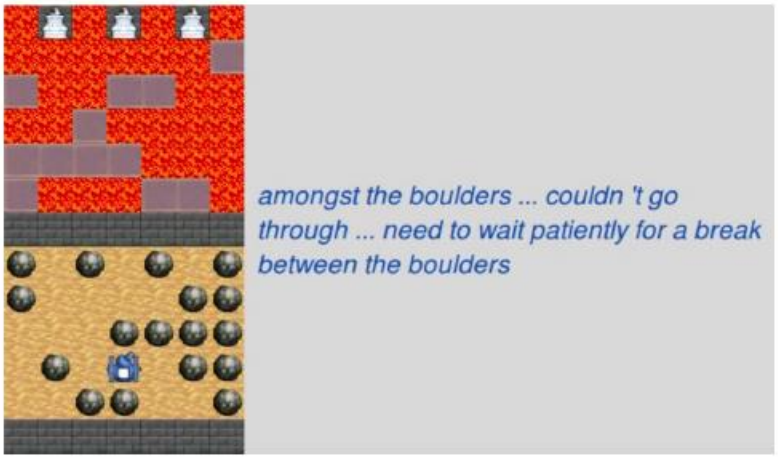

Figure 8. The "Rationalizing Robot" from Ehsan et al. (2018) providing an example rationalization of its decisions.

5. General Lessons

Some unifying threads of the previous discussion can now be highlighted as topics for future research. First, we need more explicit discussion of how to properly align machine and human performance when conducting comparisons. Second, we need to be more rigorous in scoring the equivalent human performance, especially by removing the rose-tinted glasses of anthropofabulation. Third and finally, we need

to better understand the provenance of scaffolded learning in adult humans—and the manually-encoded knowledge or representational primitives in competitor models which are designed to capture this scaffolding. Only by addressing these issues can we have a thriving and meaningful comparative psychology of AI.

References

Block, Ned. 1981. "Psychologism and Behaviorism." *The Philosophical Review* 90 (1): 5–43.

Blundell, Charles, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. 2016. "Model-Free Episodic Control." *arXiv Preprint arXiv:1606.04460*.

Boesch, Christophe. 2007. "What Makes Us Human (Homo Sapiens)? The Challenge of Cognitive Cross-Species Comparison." *Journal of Comparative Psychology* 121 (3): 227.

Botvinick, Matthew, David GT Barrett, Peter Battaglia, Nando de Freitas, Darshan Kumaran, Joel Z. Leibo, Timothy Lillicrap, Joseph Modayil, Shakir Mohamed, and Neil C. Rabinowitz. 2017. "Building Machines That Learn and Think for Themselves." *Behavioral and Brain Sciences* 40.

Buckner, Cameron. 2013. "Morgan's Canon, Meet Hume's Dictum: Avoiding Anthropofabulation in Cross-Species Comparisons." *Biology & Philosophy* 28 (5): 853–71.

———. 2017a. "Rational Inference: The Lowest Bounds." *Philosophy and Phenomenological Research*, no. Early View.

———. 2017b. "Understanding Associative and Cognitive Explanations in Comparative Psychology." In *The Routledge Handbook of Philosophy of Animal Minds*, edited by Kristin Andrews and Jacob Beck. London: Routledge University Press.

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

Crosby, Matthew, Benjamin Beyret, and Marta Halina. 2019. "The Animal-AI Olympics." *Nature Machine Intelligence* 1 (5): 257. https://doi.org/10.1038/s42256-019-0050-3.

Ehsan, Upol, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 81–87. ACM.

Elsayed, Gamaleldin F., Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. "Adversarial Examples That Fool Both Human and Computer Vision." *arXiv Preprint arXiv:1802.08195*.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. *Explaining and Harnessing Adversarial Examples*. ICLR 2015.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'right to Explanation.'" *AI Magazine* 38 (3): 50–57.

Goodman, Nelson. 1955. *Fact, Fiction and Forecast*. Vol. 6. Harvard University Press.

Guest, Olivia, and Bradley Love. 2019. "Levels of Representation in a Deep Learning Model of Categorization | bioRxiv." 2019. https://www.biorxiv.org/content/10.1101/626374v1?fbclid=IwAR0ieBRWC7mRMt6FR_CaqwZtxM6U2CMZZXA7SzZgCs6CZ3L0cRAY9Jj91EY.

Gunning, David. 2017. "Explainable Artificial Intelligence (xai)." *Defense Advanced Research Projects Agency (DARPA), Nd Web*.

Gupta, Anoopum S., Matthijs AA van der Meer, David S. Touretzky, and A. David Redish. 2010. "Hippocampal Replay Is Not a Simple Function of Experience." *Neuron* 65 (5): 695–705.

Hong, Ha, Daniel LK Yamins, Najib J. Majaj, and James J. DiCarlo. 2016. "Explicit Information for Category-Orthogonal Object Properties Increases along the Ventral Stream." *Nature Neuroscience* 19 (4): 613–22.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." *arXiv Preprint arXiv:1905.02175*.

Kahneman, Daniel, and Shane Frederick. 2002. "Representativeness Revisited: Attribute Substitution in Intuitive Judgment." *Heuristics and Biases: The Psychology of Intuitive Judgment* 49: 81.
Karin-D'Arcy, M. 2005. "The Modern Role of Morgan's Canon in Comparative Psychology." *International Journal of Comparative Psychology* 18 (3).
Khaligh-Razavi, Seyed-Mahdi, and Nikolaus Kriegeskorte. 2014. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation." *PLoS Computational Biology* 10 (11). https://doi.org/10.1371/journal.pcbi.1003915.
Kubilius, Jonas, Stefania Bracci, and Hans P. Op de Beeck. 2016. "Deep Neural Networks as a Computational Model for Human Shape Sensitivity." *PLOS Computational Biology* 12 (4): e1004896. https://doi.org/10.1371/journal.pcbi.1004896.
Kunst-Wilson, William R., and Robert B. Zajonc. 1980. "Affective Discrimination of Stimuli That Cannot Be Recognized." *Science* 207 (4430): 557–58.
Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. "Human-Level Concept Learning through Probabilistic Program Induction." *Science* 350 (6266): 1332–38.
Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017a. "Building Machines That Learn and Think like People." *Behavioral and Brain Sciences* 40.
———. 2017b. "Ingredients of Intelligence: From Classic Debates to an Engineering Roadmap." *Behavioral and Brain Sciences* 40.
Lake, Brenden M., Wojciech Zaremba, Rob Fergus, and Todd M. Gureckis. 2015. "Deep Neural Networks Predict Category Typicality Ratings for Images." In *Proceedings of the 37th Annual Cognitive Science Society*.
LeCun, Yann. 2018. "The Power and Limits of Deep Learning." *Research-Technology Management* 61 (6): 22–27. https://doi.org/10.1080/08956308.2018.1516928.
Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *arXiv:1606.03490 [cs, Stat]*, June. http://arxiv.org/abs/1606.03490.
Luc, Pauline, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. 2017. "Predicting Deeper into the Future of Semantic Segmentation." In *Proceedings of the IEEE International Conference on Computer Vision*, 648–57.
Marcus, Gary. 2018. "Deep Learning: A Critical Appraisal." *arXiv:1801.00631 [cs, Stat]*, January. http://arxiv.org/abs/1801.00631.
Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, and Georg Ostrovski. 2015. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518 (7540): 529.
Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. "Inceptionism: Going Deeper into Neural Networks." *Google Research Blog. Retrieved June* 20: 14.
Morgan, Conwy Lloyd. 1903. *An Introduction to Comparative Psychology*. London: W. Scott.
Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks Are Easily Fooled: High Confidence... - Google Scholar." *IEEE Conference on Computer Vision and Pattern Recognition*, 427–36.
Nisbett, RE, and L Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs: Prentice-Hall. http://www.getcited.org/pub/101972193.
Ponce, Carlos R., Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. 2019. "Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences." *Cell* 177 (4): 999–1009.e10. https://doi.org/10.1016/j.cell.2019.04.005.
Proudfoot, Diane. 2011. "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game." *Artificial Intelligence* 175 (5-6): 950–57.
Quine, WV. 1969. "Natural Kinds." In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, 5–23. Dordrecht: Riedel.
Rezende, Danilo Jimenez, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. 2016. "One-Shot Generalization in Deep Generative Models." *arXiv Preprint arXiv:1603.05106*.
Samuels, Richard, Stephen Stich, and Michael Bishop. 2002. "Ending the Rationality Wars: How to Make Disputes About Human Rationality Disappear." In *Common Sense, Reasoning and Rationality*, edited by Renee Elio, 236–68. Oxford University Press.

Schacter, Daniel L. 1987. "Implicit Memory: History and Current Status." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13 (3): 501.
Shevlin, Henry, and Marta Halina. 2019. "Apply Rich Psychological Terms in AI with Care." *Nature Machine Intelligence* 1 (4): 165.
Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton. 2017. "Mastering the Game of Go without Human Knowledge." *Nature* 550 (7676): 354–59.
Sober, Elliott. 1998. "Morgan's Canon." In *The Evolution of Mind*, edited by Denise Cummins and Colin Allen, 224–42. New York, NY: Oxford University Press.
Socher, Richard, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. 2013. "Zero-Shot Learning through Cross-Modal Transfer." In *Advances in Neural Information Processing Systems*, 935–43.
Stickgold, Robert. 2005. "Sleep-Dependent Memory Consolidation." *Nature* 437 (7063): 1272.
Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. "Matching Networks for One Shot Learning." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 3630–38. Curran Associates, Inc. http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf.
Walker, Matthew P., and Robert Stickgold. 2010. "Overnight Alchemy: Sleep-Dependent Memory Evolution." *Nature Reviews. Neuroscience* 11 (3): 218. https://doi.org/10.1038/nrn2762-c1.
Yamins, Daniel LK, and James J. DiCarlo. 2016. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." *Nature Neuroscience* 19 (3): 356.
Zednik, Carlos. 2019. "Solving the Black Box Problem: A General-Purpose Recipe for Explainable Artificial Intelligence." *arXiv:1903.04361 [cs]*, March. http://arxiv.org/abs/1903.04361.
Zhou, Zhenglong, and Chaz Firestone. 2019. "Humans Can Decipher Adversarial Images." *Nature Communications* 10 (1): 1334. https://doi.org/10.1038/s41467-019-08931-6.
Złotowski, Jakub, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. "Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction." *International Journal of Social Robotics* 7 (3): 347–60. https://doi.org/10.1007/s12369-014-0267-6.