PERSPECTIVE

Performance vs. competence in human–machine comparisons

Chaz Firestone[a,1]

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved July 3, 2020 (received for review July 3, 2019)

Does the human mind resemble the machines that can behave like it? Biologically inspired machine-learning systems approach "human-level" accuracy in an astounding variety of domains, and even predict human brain activity—raising the exciting possibility that such systems represent the world like we do. However, even seemingly intelligent machines *fail* in strange and "unhumanlike" ways, threatening their status as models of our minds. How can we know when human–machine behavioral differences reflect deep disparities in their underlying capacities, vs. when such failures are only superficial or peripheral? This article draws on a foundational insight from cognitive science—the distinction between *performance* and *competence*—to encourage "species-fair" comparisons between humans and machines. The performance/competence distinction urges us to consider whether the failure of a system to behave as ideally hypothesized, or the failure of one creature to behave like another, arises not because the system lacks the relevant knowledge or internal capacities ("competence"), but instead because of superficial constraints on demonstrating that knowledge ("performance"). I argue that this distinction has been neglected by research comparing human and machine behavior, and that it should be essential to any such comparison. Focusing on the domain of image classification, I identify three factors contributing to the species-fairness of human–machine comparisons, extracted from recent work that equates such constraints. Species-fair comparisons level the playing field between natural and artificial intelligence, so that we can separate more superficial differences from those that may be deep and enduring.

artificial intelligence | deep learning | perception | cognition | development

Intelligent machines now rival humans on a stunning array of tasks: They can recognize images of objects and faces, answer questions posed in natural language, make strategic decisions under risk and uncertainty, and perform other cognitive feats once thought out of reach for artificial intelligence (AI) (ref. 1 and Fig. 1*A*). These advances support technologies such as automated radiology, machine translation, autonomous vehicles, and more (refs. 2–4; though see refs. 5 and 6). But beyond such practical purposes, machine-learning successes have also been exciting for researchers studying the human mind and brain. Recent work, for example, suggests that intermediate layers of certain artificial neural networks (ANNs) resemble known processing stages in human vision (7–9) and that such models can predict the behavior (10–12) and neural processing of humans and other primates, from large-scale activation of brain regions to the firing patterns of individual neurons (13–19).*

What's at Stake

This unprecedented success raises an exciting possibility: that such systems not only solve the engineering challenges they were designed for, but also meaningfully reproduce aspects of human perception and cognition—a kind of "model organism" for people (20). For example, artificial neural networks—especially those branded as deep neural networks (DNNs)—are now said to have "significant representational similarities to human brains" (7), to potentially "explain many aspects of human cognition" (21), and even to "carve the brain at its joints" (22); see also refs. 12, 23, and 24.

Claims like these are exciting not only for their theoretical significance in cognitive science, engineering,

[a]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218

Author contributions: C.F. wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

[1]Email: chaz@jhu.edu.

First published October 13, 2020.

*This paper is written for an interdisciplinary audience of psychologists, neuroscientists, philosophers, and engineers. Readers already familiar with the performance/competence distinction from developmental psychology and linguistics, or with recent machine-learning successes and failures, could skip any slow-moving sections until *A Performance/Competence Distinction for Machines*.

Fig. 1. (*A*) Machine-learning systems approach (and sometimes surpass) "human-level" benchmarks on a wide array of tasks, especially those involving visual recognition—e.g., identifying traffic signs, recognizing objects and reading text. (*B*) But they also "fail" on carefully chosen inputs that cause bizarre misclassifications. What do such behaviors imply about human–machine similarity?

and philosophy (25–27), but also for opening new avenues for scientific discovery. For example, this possibility could enable a "virtual electrophysiology" (17) in which machines participate in "experiments" that would be unethical or resource intensive in humans—e.g., recording how machine "neurons" respond to experimental stimuli. So too for a virtual neuropsychology exploring "lesions" to machine "brains," or a virtual psychophysics that collects behavioral responses on massive stimulus sets. Although an age of purely virtual neuroscience lies well into the future (if it lies anywhere at all), seeds of this research are planted already (18, 19, 28–30).

Perspective: Distinguish "Performance" from "Competence"

Nearly all of the psychological and neuroscientific promise of these advances rests on a kind of similarity between cognitive processing in humans and the corresponding processes in candidate machine-learning systems. How should we evaluate such similarity? Although this question has a long history, here I aim to shed a new kind of light on it. Both classical and contemporary discussions of such issues tend to focus on machine-learning "successes"—e.g., asking what amazing feat a machine must perform in order to be relevantly humanlike (31–33). By contrast, here I focus on interpreting machine "failures": asking how *differently* a machine must behave from humans in order to be sufficiently *unhumanlike*. Even the most sophisticated machine-learning systems can commit surprisingly odd and alarming errors in perceiving and acting on the world around them (Fig. 1*B*). These strange behaviors pose challenges for incorporating such technologies into our lives (34); but they also threaten the theoretical excitement about humanlike processing in machines, suggesting that glimpses of humanlike success elsewhere must have unhumanlike origins after all (26, 35–38).

However, I suggest that behavioral differences between humans and machines—including even spectacular machine failures—are not always what they seem, and that they must be interpreted in light of the different *constraints* that humans and machines inevitably face (including differences in hardware, speed, task presentation, response modes, and more). In particular, I argue that evaluations of human–machine similarity are unlikely to succeed without incorporating a foundational insight from developmental psychology and psycholinguistics: distinguishing the observed "performance" of a system from its underlying "competence" (39). Drawing on this distinction, I review an emerging literature comparing human and machine behavior—primarily in the domain of image classification—to extract three principles of "fair" human–machine comparisons, and cast these principles into guidelines for future tests: 1) placing human-like constraints on machines, 2) placing machine-like constraints on humans, and 3) "species-specific" task alignment. In each case, I show how adopting these principles has concretely revised earlier thinking about human–machine behavioral differences, and I point to further differences that might be similarly illuminated. Finally, I discuss a challenge problem where machine failures go beyond mere performance constraints. Together, these cases show how distinguishing performance from competence can separate more superficial human–machine differences from those that may be deep and enduring.[†]

The Problem of "Machine Failures"

When machines behave in ways that resemble human performance, they are often described as humanlike or brainlike (43–45). Such claims are especially strong in the domain of visual processing (the primary focus of this paper), where ANNs of various flavors are claimed to "capture the essential characteristics of biological object recognition" (p. 1 in ref. 46) and serve as "the best current model of high-level visual areas in the brain" (p. 1 in ref. 47).

However, a powerful reason to doubt such resemblance arises when the same systems "fail" in ways that humans do not. Such failures pervade many domains of artificial intelligence research (both within and outside the Deep Learning umbrella): Robots that walk with strikingly humanlike gaits may collapse spectacularly if they must also twist a doorknob or hold a box (48); question-answering systems that defeat world-class human trivia players make errors that elementary students might catch [e.g., placing Toronto in America (49)]; machines that easily solve college-level math problems can fail on questions from kindergarten (e.g., 1+1+1+1+1+1+1=?; ref. 50); and in perhaps the most striking class of such failures, machines may be vulnerable to adversarial attacks wherein they embarrassingly misclassify the stimuli around them (Fig. 1*B*).

Adversarial examples are inputs that "fool" machines into giving strange and inappropriate classifications of images, speech, or text (51). For example, carefully crafted noise images that otherwise look like meaningless visual patterns may be

[†]Importantly, the question of humanlike processing in machines is distinct from whether such systems are truly "intelligent" or "minded" in some deeper sense (27, 33, 40), which can be asked independently of their similarity to humans. It also differs from the engineering goal of actually building such machines (26), which is separable from evaluation. Although all such questions may benefit from contact with cognitive science, my focus here is on *comparing* cognitive processes in humans to analogous processes in machine-learning systems, especially as such processes manifest in the "output behavior" of such systems. Indeed, throughout this discussion, I'll speak generally of "human–machine similarity" (or dissimilarity), since the factors I consider are intended to generalize beyond any one case study (e.g., image classification by people vs. by AlexNet) to other architectures, training regimes, and tasks.

classified as a "crossword puzzle" or "armadillo" (52). Similarly, ordinary images that would normally be classified accurately can be slightly perturbed to completely change machine classifications—e.g., from "orange" to "power drill" (53) or "daisy" to "jaguar" (54). ANNs may also diverge from humans on images with conflicting cues, as when a cat's shape is rendered in the texture of elephant skin; many ANNs say "elephant", whereas humans know not to (55, 56). Machines can even be fooled by natural images depicting unusual scenarios or views, such as an overturned school bus on a wintry road (classification: "snowplow"), or a honeycomb-patterned umbrella (classification: "chainlink fence") (57).

What Do Human–Machine Differences Mean? Machine failures don't just differ from ordinary human behaviors; they seem completely inaccurate and even bizarre from a human point of view. And they appear to undermine claims of human–machine similarity, suggesting "an astonishing difference in the information processing of humans and machines" (p. 1 in ref. 35) or "a disconnect between human and computer vision" (p. 1 in ref. 58), and showing how "deep learning systems do not construct human-like intermediate representations" (p. 35 in ref. 59).

But must these failures have such bleak implications? The mere observation that two systems behave differently—even giving extremely divergent responses to the same stimuli—needn't by itself refute more general claims of shared capacities and representations. After all, two *people* may give radically different responses to a stimulus without a correspondingly radical difference in the architecture of their brains: The same insect can inspire fear in one person and fascination in another; the same math puzzle can frustrate one person and captivate another; the same dress can even appear black and blue to one person and white and gold to another (60). If none of those differences entails a fundamental cognitive or neural disparity, what should we conclude from human–*machine* differences?

How to Compare Minds: Insights from Development

Within the study of intelligent systems, one research area has a special expertise at comparing the capacities of different minds. Developmental psychology—along with its neighbor, comparative psychology—frequently explores problems rather like the human–machine comparisons we are considering here. For example, they may ask how similar the mind of an infant is to that of a mature adult; how a healthy mind differs from a disordered one; or how a human mind compares to the mind of a chimpanzee, crow, or octopus. In doing so, these fields have established methodological and conceptual tools for ensuring fair comparisons between intelligences. One tool in particular has been essential to such comparisons, but almost never appears in the scientific literature comparing humans and machines (refs. 26, 36–38, 41, 45, and 61–64; cf. refs. 27 and 65): distinguishing performance from competence.

Internal Knowledge vs. External Expression. Cognitive science traditionally distinguishes what a system *knows* (competence) from what it *does* (performance). Competence is a system's underlying knowledge: the internal rules and states that ultimately explain a given capacity, often in idealized terms. Performance, by contrast, is the application or use of such competences: how the system actually behaves when prompted to express its knowledge.

The insight motivating this distinction is that intelligent creatures often know more than their behavior may indicate, because of "performance constraints" that get in the way. Consider some examples from linguistics, where the performance/competence distinction was introduced by Chomsky (39). Even native English speakers occasionally misspeak (e.g., accidently saying "I swimmed yesterday") or stumble over long, complex sentences—as when we easily process "The dog barked" or "The dog who bit the cat barked" but perhaps struggle with "Sherry yelped when the dog who bit the cat that Jay's son Ross fed barked." Why? The reason for these errors is not that we don't know how to conjugate "swim" or don't understand how phrasal structure works; rather, it's that humans are subject to performance constraints that are distinct from their underlying linguistic capacities, such as limited working memory or attention. If there's not enough room in one's head to juggle all the linguistic balls that long or convoluted sentences present, then one won't easily process those sentences—but only because of practical (even superficial) constraints on memory. So even if human listeners could understand arbitrarily long sentences in principle—i.e., even if their linguistic knowledge per se wouldn't prevent such understanding—they may not do so in practice because of other limitations. (By contrast, other failures really do reflect missing knowledge: When monolingual English speakers don't understand sentences spoken in Mandarin, they not only fail to behave like Mandarin speakers but also lack the relevant linguistic knowledge.)

Accommodating Performance Constraints

The performance/competence distinction is more than a theoretical insight: It has directly motivated new empirical research that has revolutionized our understanding of what various creatures know. And it can assist here too, by enriching human–machine comparisons and making them more interpretable. The next section outlines three "guidelines" for doing so; but first, it is worth seeing how concretely such insights have assisted elsewhere, by radically reinterpreting—or actively confirming—apparent differences between adult humans and other minds.

When Superficial Differences Hide Deep Similarities. A central question in developmental psychology concerns the origins of human knowledge: how much of what we know as sophisticated adults is rooted in capacities present in infancy, vs. how much comes from culture, instruction, and experience. For example, adults have the capacity for physical understanding: We know that unsupported objects fall, that solid objects cannot pass through one another, and that most things continue to exist even when out of view. Do infants share this knowledge?

Observe infants' natural behavior and you would be tempted to conclude not. After all, babies drop their favorite toys without appearing to worry what gravity will do to them, and they seem genuinely surprised by games like "peek-a-boo," in which adults cover their faces as if disappearing. These behaviors almost resemble the embarrassing machine failures reviewed earlier, in that it is hard to imagine a mature adult behaving in these strange ways. But does this mean infants *don't know* that unsupported objects fall? Perhaps not; perhaps infants drop their toys simply because they have poor motor control—a mere performance constraint that needn't imply a missing competence.

For this reason, psychologists who study infant cognition actively accommodate such constraints—e.g., by measuring infants' understanding in ways that don't require fine motor control. And indeed, when infants are instead shown "puppet shows" in which objects either rest stably on a surface or magically float while unsupported, infants stare measurably longer at unsupported floating objects, as if surprised by a violation of their expectations (66). In other words, infants knew about gravity all along; they just

failed to display that knowledge in their natural behavior, because of other limitations on their performance.

When Superficial Differences Are Deep Ones Too. Distinguishing performance from competence can also "confirm" differences between minds. For example, another enduring question is whether other animals share humans' capacity for language. Human children begin speaking English after prolonged exposure to English speech; but chimpanzees raised near English-speaking humans do not. Does this mean chimpanzees cannot learn language? Again, perhaps not; perhaps chimpanzees just have the wrong vocal tract for English phonemes, but do have the underlying linguistic capacity.

To accommodate such limitations, an ambitious project attempted to teach a chimpanzee American Sign Language (ASL) instead of a spoken language (67). This chimpanzee—affectionately named "Nim Chimpsky"—was raised by human parents, lived with human children, and was instructed in ASL like a young human signer. However, despite learning some vocabulary words and short phrases, Nim completely failed to master ASL: His vocabulary was far smaller than experience-matched human peers, and more centrally he failed to acquire ASL's syntax—the rules for embedding and modification that allow competent signers to build rich clauses with generative power. So here an apparent difference was confirmed: Since accommodating performance constraints still failed to produce humanlike behavior, we might feel safer concluding that chimpanzees lack the competence for language.

Fair Comparisons. These examples illustrate what may be called *species-fair* comparisons between minds (68). A creature may well "know what we know" and yet still fail to "do what we do" when placed in similar circumstances. And though the precise theoretical status of the performance/competence distinction is a subject of ongoing debate, its utility in the above cases is clear: Allowing other minds to demonstrate their knowledge requires accommodating their performance constraints, so that their success or failure won't depend on those constraints. Indeed, such insights have been so indispensable for comparing humans to other creatures that we should import them into comparisons between humans and machines.

A Performance/Competence Distinction for Machines

Like infants and chimpanzees, machine-learning systems are serious candidates for sharing at least some of our cognitive capacities. But also like infants and chimpanzees, human and machine performance is distinguishable from their underlying competence (27). A familiar way this distinction can arise is when humans and machines perform similarly but for different underlying reasons—as in longstanding critiques of accuracy benchmarks as bases for comparison. But a less familiar way this distinction can arise is when humans and machines perform *differently* despite *similar* underlying competences, as a result of differing performance constraints. Biological and artificial systems are constrained in different ways: They operate on hardware of differing speed and capacity, have different modes of issuing behavioral responses, and even accept sensory input differently in the first place. For example, human vision is sharp and colored at the fovea but blurry and nearly colorblind in the periphery, whereas machine-vision systems process digital images that may be uniformly resolved. And so humans and machines that view the "same" photograph of a scene (or hear the same speech sample, or touch the same object) may process "different" internal images, because of different constraints on receiving input. Conversely, machines are often constrained to a pool of limited response options (e.g., choosing among ImageNet labels), whereas humans typically have more freedom in their classification decisions. These and other differing constraints could cause humans and machines to perform differently even if their internal competences were similar. How should we accommodate them?

Fig. 2. Intelligent systems may fail to express their knowledge in tests that don't accommodate their particular performance constraints. Image credit: Victoria Dimitrova (artist); inspired by Hans Traxler.

Species-Fair Human–Machine Comparisons: Three Factors. The rest of this discussion explores three factors contributing to the species-fairness of human–machine comparisons. Importantly, each factor discussed below is empirically anchored, having been extracted from a concrete "case study" of how adopting it led researchers to reinterpret human–machine behavioral differences (primarily in machine vision). Although most of these cases recast human–machine differences into similarities (as with infants' physical knowledge), I conclude by discussing a challenge problem that has eluded certain machine-learning architectures in ways that don't reflect mere performance constraints (as with chimpanzees' failed language acquisition). When a difference remains even after suitably fair comparisons, this suggests even more strongly that the difference is deep and meaningful after all.

1. Limit Machines Like Humans

Could human performance constraints account for differences in human and machine behavior? One way to find out—our first factor of three—is to actively burden machines with human limitations and ask if the relevant behavioral differences attenuate or even disappear.

How This Has Helped: A Case Study. Researchers exercised by adversarial examples often highlight how the tiniest perturbations can alter machine classifications: Even "small and almost imperceptible" changes (p. 1 in ref. 35) that are "invisible to the human eye" (64) can cause woeful misclassifications, which is interpreted as "an astonishing difference in the information processing of humans and machines" (p. 1 in ref. 35). However, as just noted, the image details a machine can process are limited only by the resolution of the digital file it takes as input—whereas humans view such images on physical displays that may not preserve the perturbation[‡], using

[‡]Indeed, some adversarial perturbations are too small to flip a single bit on an 8-bit-color display (e.g., a VGA-connected monitor or projector), leaving them completely physically absent when viewed on such displays.

limited-acuity sensors (their eyes) that distort the images further. In that case, being invisible to the human eye might be a reason not to infer an astonishing difference in "information processing" after all. (When one is interested in information processing, one is surely interested in processing done by the mind, not the eye.) Indeed, for all we know, humans too could detect the regularities captured by some of these perturbations (69), if only we were permitted to see them.

One recent study tested this possibility directly. To explore how optical or physiological limitations contribute to divergent responses to certain classes of adversarial images, Elsayed et al. (70) asked what would happen if an adversarial attack had to fool machine-vision systems that viewed images through a humanlike "eye" (with a standard architecture thereafter). In other words, instead of fooling standard convolutional neural networks (CNNs) such as AlexNet or GoogLeNet, these researchers required the adversarial perturbation to retain its fooling powers even after passing through a model of the retina. Can such attacks still succeed? And what do they look like?

The remarkable answer is that such attacks fool human vision too. For example, to alter a cat image so that multiple retinally constrained CNNs call it a dog, the required perturbation ends up adding features that look doglike to machines *and* humans (Fig. 3*A*)—so much so that humans classify it as a dog under brief exposures. Similar results arise for adversarial perturbations confined to small but high-contrast (and so highly visible) patches (such as a daisy-to-jaguar image that acquires a jaguar-like patch in its corner) (54) or when an attack must fool many different machine-recognition systems in real-world settings—as with a 3D "turtle" model that is misclassified as a "jigsaw puzzle" after acquiring a puzzle-like texture on its shell (53, 71). Intriguingly, this suggests that the more general the attack (e.g., fooling many systems in many conditions), the more sensitive humans may be to its result.

Importantly, what happened in this case study was not that a machine behaved in more humanlike ways because human behaviors were "programmed into it"; that would be less interesting. Rather, placing a superficial limitation on a peripheral processing stage attenuated a difference previously attributed to more central processes. Of course, this needn't mean that all human insensitivity to adversarial attacks will be explained in this way (72), but rather that approaches like this can reveal which behavioral differences have more superficial explanations and which have deeper origins. And although Elsayed et al. (70) do not use the language of performance and competence, their work perfectly embodies that insight—namely, that fair comparisons must equate constraints.

How It Could Help. Might other behavioral differences be illuminated by human performance constraints? It is easy to imagine so. For example, many machine-vision systems are intolerant to image distortions: If CNNs are trained on clean images but tested on noisy images, they perform far below humans at test (73). But here too, if machines were burdened with humanlike visual acuity and so could barely represent the high-frequency features in the training set (i.e., the features most distorted by this sort of noise), they may be less sensitive to the patterns that later mislead them (74). Indeed, recent work finds that giving CNNs a humanlike fovea (75) or a hidden layer simulating V1 (76) improves robustness to distortions and other perturbations (including adversarial examples).

Imposing humanlike limitations could also apply beyond vision. For example, systems trained through deep reinforcement learning now defeat the best human players not only at Chess and Go but also more complex and dynamic videogames. In doing so, machines often use strategies that human experts find unhumanlike and "almost nonsensical" (77). However, machine players often have abilities and knowledge that humans couldn't have, including a constant complete map view, machine-level calculation speed, and superhuman response times—all of which could contribute to strange-seeming gameplay. Indeed, it has been suggested that a "level playing field" equating such constraints could produce more humanlike strategies (78).

2. Limit Humans Like Machines

Humans face limitations that machines do not, but machines also face limitations that humans do not. For example, machine-vision systems are typically "passive" learners that cannot explore their environment, acquire new data, or compose new responses. Could such constraints explain any human–machine behavioral differences? One way to find out could be to grant curiosity-like abilities to machines and ask whether they behave more humanlike; but of course this is a major engineering challenge (79). A different approach—and one whose technology is more easily available—is to place machine-like constraints on humans and ask whether they come to *behave like machines*.

How This Has Helped: A Case Study. Machine-vision systems are often trained on curated datasets of labeled images, such as MNIST, CIFAR10, or ImageNet, which arose from challenge

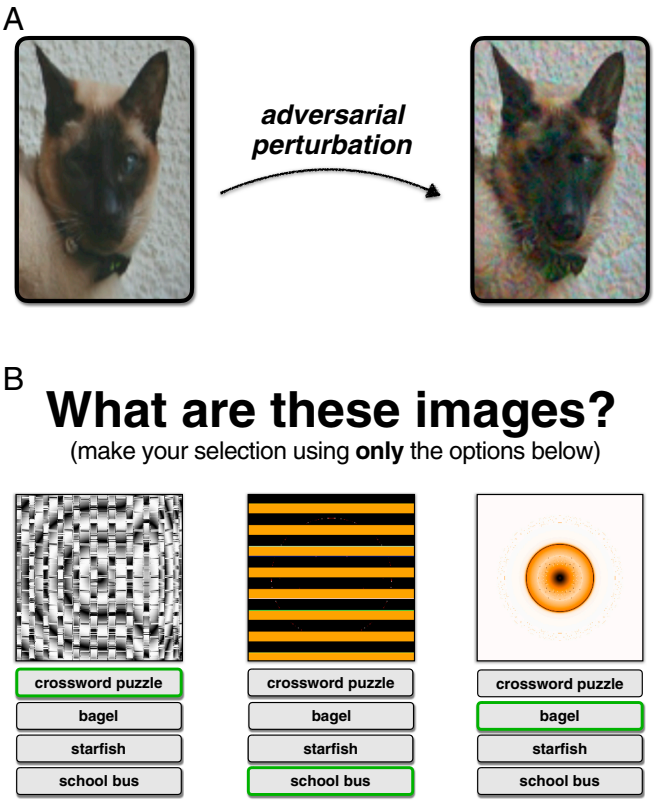

Fig. 3. (*A*) Most adversarial perturbations are too small for humans to see. But when a cat is perturbed to appear like a dog to a CNN with a prepended "retina," the image acquires features that appear doglike to humans, too. (*B*) Machines label these odd patterns as familiar objects. But if you had to choose from a limited set of labels, what would you pick? Under such constraints, humans become more likely to agree with machine (mis)classifications.

problems in computer vision. Although such datasets are very rich, a consequence of such training regimes is that the machine's resulting "vocabulary" is typically limited to those labels in the dataset—often an extremely eclectic array of specific terms such as "crossword puzzle," "chihuahua," "school bus," or "meatloaf." Every image a system like AlexNet (42) sees can receive only labels from this set; but this constraint doesn't exist for humans, who can describe images however they like. So, even if machines trained on such datasets could (internally) "see" images in humanlike ways, they may nevertheless give (externally) unhumanlike classifications if their output is constrained in ways that humans are not.

One study asked whether such constraints might contribute to human–machine differences on peculiar "noise" images (52), which seem meaningless to humans but are confidently recognized as familiar objects by ImageNet-trained CNNs (e.g., Fig. 3*B*'s crossword puzzle). These images have been described as "totally unrecognizable to human eyes" (52); and indeed, this behavior (initially) seems completely unreplicable in humans. For example, you might describe this image as "some vertical bars with black, white, and gray segments, atop a swirling black and white background"—but certainly not a crossword puzzle. However, AlexNet isn't even permitted to answer that way; it can only decide which label fits best.

To explore the role of such constraints, Zhou and Firestone (80) asked humans to classify such noise images, but with the machine-like constraint of choosing from labels available to the relevant CNNs. For example, humans saw the crossword-puzzle image, and beneath it appeared the label "crossword puzzle" alongside one or more ImageNet labels randomly drawn from the imageset ["school bus," "bagel," "starfish," etc. (cf. ref. 81)]. Intriguingly, humans in these conditions tended to choose labels that agreed with CNN classifications better than would be expected by chance—just as you might when looking at Fig. 3*B*. Indeed, like AlexNet, you might even be "confident" in your selection of crossword puzzle; not confident that the image really is a crossword puzzle, but rather confident that crossword puzzle is the best label given the options, which is roughly what AlexNet does in choosing it too (i.e., in making a softmax decision over activations for each label). In that case, strange machine classifications of at least some of these images are perhaps not so odd after all: Since the machines aren't permitted to invent or compose new labels (or say "looks like a crossword puzzle but isn't one"), the best they can do is compute a similarity judgment and output a label better than any other—just as humans might under similar constraints.

How It Could Help. Placing machine-like constraints on humans might illuminate other behavioral differences. For example, CNNs have purely feedforward architectures; in humans, this may better map onto fast, early visual processes than slower, flexible cognitive ones (82). In that case, additional machine-like constraints for humans may be those that load more on vision than higher cognition [e.g., brief image presentations (12, 70)], which may promote more "machine-like" answers (e.g., becoming more likely to say "golf ball" in Fig. 1*B*). Related approaches could apply beyond vision, including auditory misclassification (83).

Similarly, image-classifying machines typically produce rank-ordered lists of their favored labels, making such responses richer than the single labels humans are usually asked for (56). But humans could easily produce machine-like lists too. Indeed, since machine-learning benchmarks often consider it a success if the correct label is among the machine's top few choices, one could similarly ask if the machine's label is among the human's top choices, even if their first-choice labels differ. Consider again the teapot/golf-ball image, for which many CNNs answer "golf ball" but humans answer "teapot"; how concerning is this discrepancy? If two people examine Michelangelo's *David*, one might say "man" while the other says "marble" without this implying a radical difference in their vision—especially if one would have said the other's first choice as its second. Similarly, "golf ball" would presumably be high on humans' lists too. A fuller inventory of human responses might thus reveal more similarity than was initially apparent.

3. Species-Specific Task Alignment

The previous factors "equate" constraints on humans and machines, to make their testing conditions more similar. But sometimes equating such factors is impossible or undesirable, and instead it is better to "accommodate" constraints—even when this means choosing different tasks for different creatures.

Consider an example from cross-species work in psychology. Many studies ask how different organisms process reward, including which reward-learning mechanisms are shared by rodents, human children, and adults. But different organisms value different stimuli, and so the rewards used in rodent studies (e.g., water, alcohol, or cocaine) differ from those used for children (e.g., colorful stickers or plush toys), which in turn differ from adult rewards (e.g., monetary payments or course credit). Clearly, one should not literally equate these details; using the same stimuli across these species (e.g., dollars in humans and dollars in rodents) would surely probe different processes. Does this apply to machines too?

How This Has Helped: A Case Study. It has recently been suggested that human vision fundamentally differs from machine vision in that human vision is "atomic" in ways that machine vision is not (84). When humans must classify images from only a small cropped patch, they exhibit near-ceiling accuracy until some critical patch size, after which they are near floor. For example, humans can recognize an airplane from a patch showing its cockpit and wheel; but any smaller patch causes accuracy to drop steeply and discontinuously. By contrast, CNN performance on the same images and patches simply decays gradually, without any discrete inflection. This result led Ullman et al. (84) to conclude that human recognition differs from machine recognition in a fundamental way: Humans rely on discrete features to classify images ("atoms of recognition"), whereas CNNs do not.

However, these superficially similar human and machine tasks might be more different than they seem. In particular, both humans and machines in the minimal-patch studies saw patches selected for humans; the patches were derived from human psychophysics experiments, and then those patches were shown to humans and machines. What if, instead, machines chose their own patches? Funke et al. (85) ran just this test, generating minimal patches from a "machine psychophysics" experiment. When tested on machine-selected patches, CNNs showed the same sharp dropoff in accuracy. In other words, aligning tasks required different stimuli: human-selected patches for humans and machine-selected patches for machines. In this aligned setting, both showed atomic recognition patterns (86).

Of course, the fact that different patches were needed in the first place suggests *some* kind of human–machine difference—i.e., a difference in the features that each system finds diagnostic. But that difference seems relatively less significant—and more likely to change—than the deeper claim that human recognition is a fundamentally different kind of process than machine recognition. After all, a child's preference for stickers over money differs

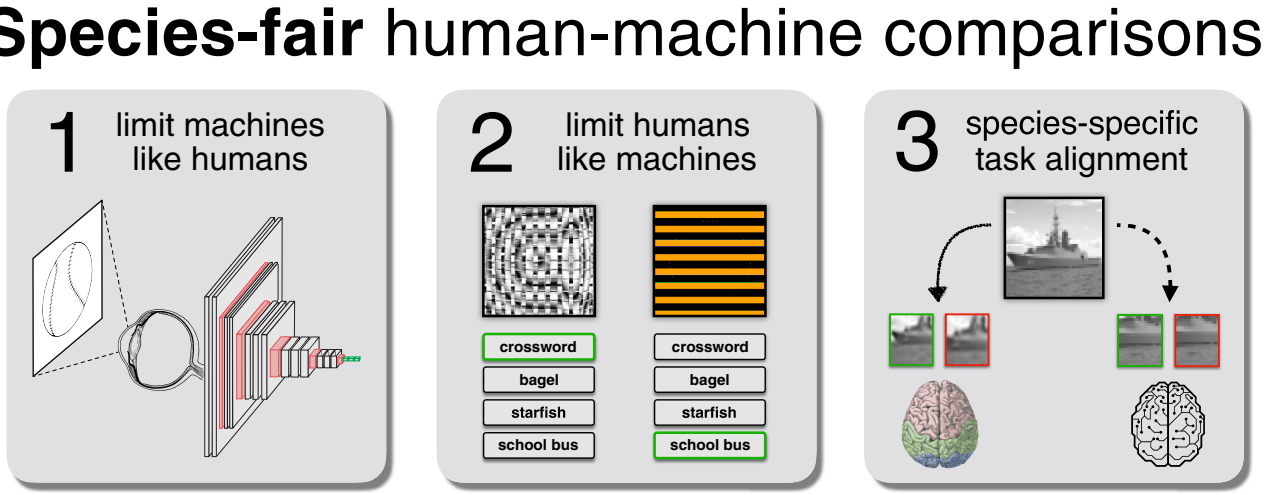

Fig. 4. Principles of "species-fair" human-machine comparisons. 1) Placing humanlike constraints on machines, such as filtering visual input through a humanlike "retina." 2) Placing machine-like constraints on humans, such as limited response options or brief exposures. 3) Aligning tasks in specific-specific ways: In this case study, natural images were distilled into "atomic" minimal pairs, such that a small patch is recognizable (green) but a slightly different patch isn't (red); although it initially seemed that this applied only to humans (*Left*), recent work shows that optimizing patches for machine vision produces a similarly atomic pattern (*Right*). Future human–machine comparisons may incorporate one or more of these accommodations.

from an adult's preference for money over stickers; but this isn't a particularly central difference, especially compared to the more foundational process of learning the rewards that those stimuli represent. Similarly here then: Even if a human and a machine prioritize different features, the processing they do over those features may share an atomic signature.

How It Could Help. Could failed alignment explain other human–machine behavioral differences? A growing literature asks whether machine-recognition systems see humanlike visual illusions; but the results are mixed, with CNNs showing some illusions but not others (87). While this could imply deep architectural differences, another possibility concerns alignment: As Ward (87) notes, machines might "misunderstand" their task (e.g., which features should be reported), as even humans sometimes do (88). More generally, object classification alone may not lead a system to share our "perceptual goals" (although other training regimes might).

Related work has explored CNNs' apparent lack of a shape bias. If an object has the texture of an elephant but the shape of a cat (Fig. 1*B*), humans classify it as a cat with elephant skin; but most CNNs behave oppositely, answering "elephant." These and similar patterns have suggested "a crucial divergence between artificial visual systems and biological visual processes" (p. 2 in ref. 56). However, perhaps CNNs *would* classify based on shape, but their particular training environment never demanded it. To find out, Geirhos et al. (55) used style transfer techniques to produce a whole training set in which one object is rendered in the style of another. This training set makes it impossible to classify on texture alone, since the various category instances (e.g., the many "cat" images in the set) share only their shapes, not their textures. Remarkably, these conditions caused CNNs to acquire a shape bias on cat/elephant-style images, calling such images cats instead of elephants. Similarly, Hermann et al. (89) show that simply augmenting training images with naturalistic noise (e.g., color distortions or blur) produces a shape bias as well. In other words, it wasn't that CNNs couldn't give shape-based classifications of images; they just didn't employ that humanlike strategy until their environment invited it.

What Species-Fair Comparisons Look Like

The previous sections extracted three principles that have concretely improved the informativeness of human–machine comparisons (Fig. 4). These principles have thus been implemented by previous work; how could future work follow suit? Although one might consider these principles "ingredients" in a "recipe" for fair comparisons, it isn't necessary—and may not be possible—for any one test to implement them all. Instead, the most fruitful way to engage these principles may be to treat them like psychologists already do: as tools for comparing intelligent systems. When developmental psychologists wish to read deeply into adult–infant behavioral differences, the field's standards are such that researchers must consider performance constraints in making their case (whether by running more experiments to accommodate such constraints, or by arguing that the constraints don't relevantly differ). My suggestion is simply that human–machine differences follow this standard. In other words, machines should occupy the same conceptual role as chimpanzees, infants, neurological patients, etc., when compared to healthy human adults.

Saliently, however, such factors are rarely considered by contemporary human–machine comparisons. For example, one influential approach, Brain-Score, proposes a scheme for evaluating behavioral and neural similarity to humans; machines whose outputs don't match human benchmarks are scored as less "brain-like" (45). Although such scoring systems are certainly valuable, they rarely account for the practical obstacles that can prevent other systems from behaving like humans. Indeed, a small human child would almost certainly earn a low behavioral Brain-Score and thus be evaluated as not very brain-like—despite having a real human brain! So too for other frameworks, including the creative "Animal-AI Olympics" (61) that proposes two "playgrounds" (a virtual one and a real-life copy) in which to compare animal and machine performance at finding food, manipulating objects, or avoiding danger. This project aims to "keep the comparison to the animal

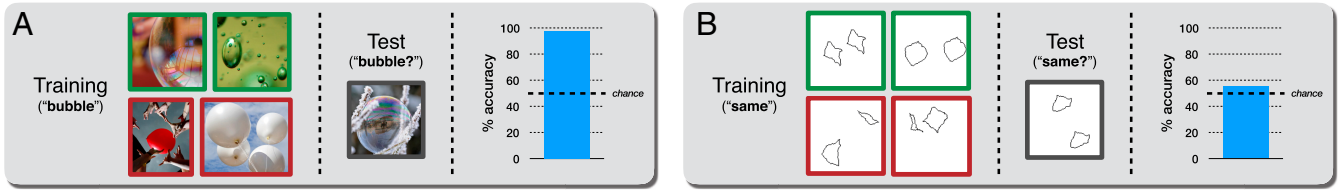

Fig. 5. (*A*) CNNs can classify natural images into hundreds of categories (e.g., bubble), generalizing over views, lighting, and other factors. (*B*) But they struggle with basic abstract relations between objects (such as two objects being the "same"), even from extremely simple images.

case as close as possible"; but differential success will be hard to interpret without explicitly considering performance constraints. This isn't to say that such frameworks aren't useful—they certainly are (especially concerning patterns of errors) (90)—but rather that failing to separate performance from competence may produce misleading results that could be avoided by accommodating such constraints.

Species-Fair Comparisons with Differences?

Considering performance constraints can also *confirm* differences between systems, as shown by the case of chimpanzees' failed language acquisition. Are there any such examples for machines?

A Case Study: Same vs. Different. Consider Fig. 5*A*. After training on natural images (e.g., bubbles and balloons), CNNs can classify new instances with high accuracy, generalizing over views, lighting, and other factors. What about Fig. 5*B*? These images are from the Synthetic Visual Reasoning Test (SVRT) (91), which presents abstract relations such as "inside vs. outside," "same vs. different," "rotated vs. mirrored," etc. Such relations are easily learned by humans; for example, the test image in Fig. 5*B* clearly involves two shapes being the "same." What about machines?

Recent work shows that CNNs are surprisingly stumped by such problems; in particular, by same/different relations. Kim et al. (92) trained nine different CNNs on SVRT problems; although the networks easily solved many of them (e.g., whether two objects are touching, or whether one object is enclosed by another), all models were at or near chance for same/different judgments—despite this being trivial for humans. Kim et al. (p. 10 in ref. 92) suggest that this "demonstrates feedforward neural networks' fundamental inability to efficiently and robustly learn visual relations", marking a "key dissimilarity between current deep network models and various aspects of visual cognition."

How far should we read into this result? Does it truly reflect a deep difference (rather than a superficial one)? The factors we have considered give us the tools to say yes. First, this differential success doesn't seem attributable to human performance constraints; after all, people do better here than CNNs, such that adding human limitations seems unhelpful. [Indeed, Kim et al. (92) explicitly rule out acuity as an explanation.] Second, there can be no concern about output constraints as with misclassified "noise" images (where machines, but not humans, have limited labels); here, there are only two labels (same/different), and they are the same labels for humans and machines. Finally, species-specific (mis)alignment seems not to blame; the task and images are so straightforward, and they were not tailored to humans in any evident way. Indeed, other creatures (including ducklings, pigeons, and insects) can complete similar same–different tasks (93); apparently, then, there is "cross-species" transfer. Yet, for one same/different problem where humans require six instances, "the best performing CNN model for this problem could not get significantly above chance from 1 million training examples" (p. 10 in ref. 92).

These CNN architectures are thus like Nim Chimpsky, who failed to acquire language despite care to exclude more superficial explanations. So just as that species-fair comparison licenses more confidence about how humans differ from other primates, this example should license similar conclusions about humans and these machine-learning implementations. And although it is conceivable that some other test could reveal efficient same/different abstraction in such architectures (just as Nim might one day learn language by some even more accommodating test), the space of superficial alternative explanations has narrowed considerably—and the performance/competence distinction allows us to say so.[§]

What Species-Fair Comparisons Show

Machines are not people, and there remain many reasons to doubt that today's leading AI systems resemble our minds—in terms of how they learn (94), what they can do (26), and even what they are in the first place (36). Even for vision, many core functions remain beyond the reach of the most advanced Deep Learning systems (38, 95). But if ever such systems do achieve humanlike perception and cognition, how will we know? I have suggested that it will not be sufficient to ask if machines reproduce humanlike behaviors—not only for the familiar reason that similar behaviors can arise from different underlying processes ("performance without competence"), but also for the less familiar reason that different behaviors can arise from similar underlying processes ("competence without performance").

The Value of Behavior. Species-fair tests offer a distinct kind of evidence from other human–machine comparisons. For example, studying machine representations "neurophysiologically" [e.g., visualizing representations in intermediate layers (96)] might seem to sidestep the performance/competence issue altogether. But even these valuable approaches can be misled by performance constraints—especially constraints on input. For example, visualization techniques on adversarial images led to conclusions that "human perception and DNN representations are clearly at odds with one another" (p. 3 in ref. 97)—which, while literally true, may not explain why such different representations arise in the first place. Indeed, it is telling that much progress in understanding such errors has come from "behavioral" studies (in humans and machines) (69, 70, 74, 80). A mixed strategy is likely best, with behavioral comparisons as essential components.

Other Developmental Lessons. Distinguishing performance from competence is not the only insight that development can

[§]Indeed, Kim et al. (92) suggest that getting feedforward networks to succeed efficiently requires fundamentally transforming them to implement Gestalt-like principles of perceptual organization, in ways that highlight the absence of segmented object representations in such networks. (See also ref. 85 for an architectural solution involving [inefficient] brute-force methods.)

offer. Others include avoiding machine anthropomorphization (40) and even human "anthropofabulation" (98)—a bias to inflate human intelligence in comparisons with other creatures. When we scoff at the errors of animals or machines, we may react this way because we imagine humans to be free of such errors. But in doing so we may be imagining humans at our best—e.g., a skilled actor in favorable conditions—and comparing such idealized abilities to animals or machines in unfavorable conditions. Importantly, such biases might lead to unfair comparisons that underestimate machine capacities (65).

Conclusion

Artificial intelligence research is increasingly informed by cognitive science and developmental psychology. The innate machinery within humans, and the rich environments we encounter, are well worth considering in the broader quest for humanlike machine intelligence. But these fields offer more than data and hypotheses about what makes humans smart; they also offer tools for identifying and comparing intelligence in other creatures. One tool that has been indispensable is to consider the performance constraints that various creatures face. Human–machine comparisons demand the same care.

Acknowledgments

For helpful discussion and/or comments on previous drafts, the author thanks Cameron Buckner, Lisa Feigenson, Steven Gross, Justin Halberda, Katherine Hermann, Christopher Honey, Tal Linzen, Pat Little, Jorge Morales, Ian Phillips, Thomas Serre, Kate Storrs, Brad Wyble, Sami Yousif, and Alan Yuille.

1 Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
2 P. Lakhani, B. Sundaram, Deep learning at chest radiography. *Radiology* **284**, 574–582 (2017).
3 D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 (19 May 2016).
4 Z. Zhu *et al.*, "Traffic-sign detection and classification in the wild" in *Computer Vision and Pattern Recognition* (2016).
5 J. Buolamwini, T. Gebru, "Gender shades" in *Proceedings of the First Conference on Fairness, Accountability and Transparency* (Association for Computing Machinery, New York, NY, 2018), pp. 77–91.
6 T. de Vried, I. Misra, C. Wang, L. van der Maaten, "Does object recognition work for everyone?" in *Computer Vision and Pattern Recognition* (2019).
7 R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
8 U. Güçlü, M. A. J. van Gerven, Deep Neural Networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
9 D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
10 M. Kümmerer, L. Theis, M. Bethge, "Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet" in *International Conference on Learning Representations* (2014).
11 T. P. O'Connell, M. M. Chun, Predicting eye movement patterns from fMRI responses to natural scenes. *Nat. Commun.* **9**, 5159 (2018).
12 J. Kubilius, S. Bracci, H. P. O. de Beeck, Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12**, e1004896 (2016).
13 M. F. Bonner, R. A. Epstein, Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Comput. Biol.* **14**, e1006111 (2018).
14 C. F. Cadieu *et al.*, Deep Neural Networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
15 S. M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
16 A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644 (2018).
17 D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
18 P. Bashivan, K. Kar, J. J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
19 D. A. Pospisil, A. Pasupathy, W. Bair, 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *eLife* **7**, e38242 (2018).
20 H. S. Scholte, Fantastic DNimals and where to find them. *Neuroimage* **180**, 112–113 (2018).
21 K. M. Jozwik, N. Kriegeskorte, K. R. Storrs, M. Mur, Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* **8**, 1726 (2017).
22 M. A. Bertolero, D. S. Bassett, Deep Neural Networks carve the brain at its joints. arXiv:2002.08891 (9 September 2020).
23 N. Kriegeskorte, Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
24 N. J. Majaj, D. G. Pelli, Deep learning—Using machine learning to study biological vision. *J. Vis.* **18**, 2 (2018).
25 C. Buckner, Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese* **195**, 5339–5372 (2018).
26 B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
27 L. Miracchi, A competence framework for artificial intelligence research. *Philos. Psychol.* **32**, 589–634 (2019).
28 C. R. Ponce *et al.*, Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009 (2019).
29 B. RichardWebster, S. Anthony, W. Scheirer, Psyphy: A psychophysics-driven evaluation framework for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2280–2286 (2018).
30 S. Ritter, D. G. T. Barrett, A. Santoro, M. M. Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study" in *34th International Conference on Machine Learning* (2017).
31 A. M. Turing, Computing machinery and intelligence. *Mind* **59**, 433–460 (1950).
32 D. Geman, S. Geman, N. Hallonquist, L. Younes, Visual Turing test for computer vision systems. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3618–3623 (2015).
33 F. Chollet, On the measure of intelligence. arXiv:1911.01547 (25 November 2019).
34 S. G. Finlayson *et al.*, Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
35 W. Brendel *et al.*, Adversarial vision challenge. arXiv:1808.01976 (6 December 2018).
36 G. Marcus, Deep learning: A critical appraisal. arXiv:1801.00631 (2 January 2018).
37 S. Stabinger, A. Rodríguez-Sánchez, J. Piater, "25 years of CNNs: Can we compare to human abstraction capabilities?" in *International Conference on Artificial Neural Networks* (Springer, 2016), pp. 380–387.
38 T. Serre, Deep learning: The good, the bad, and the ugly. *Ann. Rev. Vis. Sci.* **5**, 399–426 (2019).
39 N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965).
40 H. Shevlin, M. Halina, Apply rich psychological terms in AI with care. *Nat. Mach. Intell.* **1**, 165–167 (2019).
41 I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).
42 A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems* (2012).

43 M. Kempka, M. Wydmuch, G. Runc, J. Toczek, W. Jaśkowski, "ViZDoom: A Doom-based AI research platform for visual reinforcement learning" in *IEEE Conference on Computational Intelligence and Games* (2016), pp. 1–8.
44 A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, "Meta-Learning with memory-augmented neural networks" in *International Conference on Machine Learning* (PMLR, 2016).
45 M. Schrimpf *et al.*, Brain-Score: Which artificial neural network for object recognition is most brain-like? bioRxiv:407007 (5 September 2018).
46 I. Kuzovkin *et al.*, Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Commun. Biol.* **1**, 107 (2018).
47 K. R. Storrs, T. C. Kietzmann, A. Walther, J. Mehrer, N. Kriegeskorte, Diverse deep neural networks all predict human IT well, after training and fitting. bioRxiv: 082743 (8 May 2020).
48 E. Guizzo, The hard lessons of DARPA's robotics challenge. *IEEE Spectrum* **52**, 11–13 (2015).
49 J. Markoff, Computer wins on 'Jeopardy!': Trivial, it's not. *NY Times*, 16 February 2011, Science section.
50 D. Saxton, E. Grefenstette, F. Hill, P. Kohli, Analysing mathematical reasoning abilities of neural models. arXiv:1904.01557 (2 April 2019).
51 C. Szegedy *et al.*, Intriguing properties of neural networks. arXiv:1312.6199 (19 February 2014).
52 A. Nguyen, J. Yosinski, J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images" in *Computer Vision and Pattern Recognition* (2015).
53 A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, "Synthesizing robust adversarial examples" in *35th International Conference on Machine Learning* (2018), vol. **80**, pp. 284–293.
54 D. Karmon, D. Zoran, Y. Goldberg, LaVAN: Localized and visible adversarial noise. arXiv:1801.02608 (1 March 2018).
55 R. Geirhos *et al.*, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" in *International Conference on Learning Representations* (2019).
56 N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
57 M. A. Alcorn *et al.*, Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. arXiv:1811.11553 (18 April 2019).
58 D. Hendrycks, K. Gimpel, Visible progress on adversarial images. arXiv:1608.00530 (1 August 2016).
59 K. D. Forbus, D. Gentner, Evidence from machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
60 R. Lafer-Sousa, K. L. Hermann, B. R. Conway, Striking individual differences in color perception uncovered by 'the dress' photograph. *Curr. Biol.* **25**, R545–R546 (2015).
61 M. Crosby, B. Beyret, M. Halina, Animal-AI Olympics. *Nat. Mach. Intell.* **1**, 257 (2019).
62 R. A. Jacobs, C. J. Bates, Comparing the visual representations and performance of humans and Deep Neural Networks. *Curr. Dir. Psychol. Sci.* **28**, 34–39 (2018).
63 G. Lindsay, Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cognit. Neurosci.* **39**, 1–15 (2020).
64 R. Geirhos *et al.*, Shortcut learning in deep neural networks. arXiv. 2004.07780 (16 April 2020).
65 C. Buckner, The comparative psychology of artificial intelligences. *Philsci. Archive*:16128 (20 June 2019).
66 A. Needham, R. L. Baillargeon, Intuitions about support in 4.5-month-old infants. *Cognition* **47**, 121–148 (1993).
67 H. S. Terrace, L. A. Petitto, R. J. Sanders, T. G. Bever, Can an ape create a sentence? *Science* **206**, 891–902 (1979).
68 C. Allen, M. Bekoff, *Species of Mind* (MIT Press, 1997).
69 A. Ilyas *et al.*, Adversarial examples are not bugs, they are features. arXiv:1905.02175 (12 August 2019).
70 G. Elsayed *et al*, "Adversarial examples that fool both computer vision and time-limited humans" in *Advances in Neural Information Processing Systems* (2018), pp. 3910–3920.
71 T. Brown, D. Mane, A. Roy, M. Abadi, J. Gilmer, Adversarial patch. arxiv:1712.09665 (17 May 2018).
72 E. Kim, J. Rego, Y. Watkins, G. T. Kenyon, "Modeling biological immunity to adversarial examples" in *Computer Vision and Pattern Recognition* (2020).
73 S. Dodge, L. Karam, "Human and deep learning recognition performance under visual distortions" in *International Conference on Computer Communication and Networks* (Institute of Electrical and Electronics Engineers Inc., 2017).
74 H. Wang, X. Wu, P. Yin, E. P. Xing, High frequency component helps explain the generalization of convolutional neural networks. arXiv:1905.13545 (24 March 2020).
75 A. Deza, T. Konkle, Emergent properties of foveated perceptual systems. arXiv:2006.07991 (14 June 2020).
76 J. Dapello *et al.*, Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. bioRxiv:154542 (17 June 2020).
77 M. Gault, OpenAI is beating humans at 'Dota 2' because it's basically cheating. *Vice*, 17 August 2018.
78 R. Canaan, C. Salge, J. Togelius, A. Nealen, Leveling the playing field: Fairness in AI versus human game benchmarks. arXiv:1903.07008 (29 August 2019).
79 D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, "Curiosity-driven exploration by self-supervised prediction" in *Computer Vision and Pattern Recognition* (2017).
80 Z. Zhou, C. Firestone, Humans can decipher adversarial images. *Nat. Commun.* **10**, 1334 (2019).
81 M. Dujmović, G. Malhotra, J. Bowers, What do adversarial images tell us about human vision? *eLife*, **9**:e55978.
82 T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6424–6429 (2007).
83 M. A. Lepori, C. Firestone, Can you hear me *now*? Sensitive comparisons of human and machine perception. arXiv:2003.12362 (27 March 2020).
84 S. Ullman, L. Assif, E. Fetaya, D. Harari, Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U..S.A.* **113**, 2744–2749 (2016).
85 C. M. Funke *et al.*, The notorious difficulty of comparing human and machine perception. arXiv:2004.09406 (20 April 2020).
86 S. Srivastava, G. Ben-Yosef, X. Boix, Minimal images in Deep Neural Networks: Fragile object recognition in natural images. arXiv:1902.03227 (8 February 2019).
87 E. J. Ward, Exploring perceptual illusions in deep neural networks. bioRxiv:687905 (2 July 2019).
88 I. Phillips, Naive realism and the science of (some) illusions. *Philos. Top.* **44**, 353–380 (2016).
89 K. L. Hermann, T. Chen, S. Kornblith, The origins and prevalence of texture bias in convolutional neural networks. arXiv:1911.09071 (29 June 2020).
90 R. Rajalingham *et al.*, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
91 F. Fleuret *et al.*, Comparing machines and humans on a visual categorization test. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 17621–17625 (2011).
92 J. Kim, M. Ricci, T. Serre, Not-So-CLEVR: Learning same–different relations strains feedforward neural networks. *Interface Focus* **8**, 20180011 (2018).
93 A. Martinho, A. Kacelnik, Ducklings imprint on the relational concept of "same or different". *Science* **353**, 286–288 (2016).
94 A. M. Zador, A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 3770 (2019).
95 A. L. Yuille, C. Liu, Deep Nets: What have they ever done for vision? arXiv:1805.04025 (11 January 2019).
96 J. Yosinski, J. Clune, T. Fuchs, H. Lipson, "Understanding neural networks through deep visualization" in *ICML Workshop on Deep Learning* (2015).
97 S. Sabour, Y. Cao, F. Faghri, D. J. Fleet, Adversarial manipulation of deep representations. arXiv:1511.05122 (4 March 2016).
98 C. Buckner, Morgan's Canon, meet Hume's Dictum: Avoiding anthropofabulation in cross-species comparisons. *Biol. Philos.* **28**, 853–871 (2013).