nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-020-00266-y

Understanding adversarial examples requires a theory of artefacts for deep learning

Cameron Buckner

Deep neural networks are currently the most widespread and successful technology in artificial intelligence. However, these systems exhibit bewildering new vulnerabilities: most notably a susceptibility to adversarial examples. Here, I review recent empirical research on adversarial examples that suggests that deep neural networks may be detecting in them features that are predictively useful, though inscrutable to humans. To understand the implications of this research, we should contend with some older philosophical puzzles about scientific reasoning, helping us to determine whether these features are reliable targets of scientific investigation or just the distinctive processing artefacts of deep neural networks.

Deep neural networks (DNNs) have become one of the most important tools in artificial intelligence (AI) and diverse sciences. In AI, DNNs are often said to be capable of superhuman performance, regularly achieving new benchmark scores on standard tests of image recognition. They have defeated human grandmasters in Go[1], a game with a search space once thought beyond the reach of AI. The use of DNNs for scientific data analysis has also promised to help us overcome current limits in scientific knowledge. DNNs have enabled detection of new exoplanets orbiting stars thousands of light years away from Earth[2]; they have been proposed as a tool for analysing the hundreds of petabytes of data CERN generates in its attempt to test the standard model in physics[3]; and on its first attempt, the DNN-based AlphaFold won the Critical Assessment of protein Structure Prediction (CASP) competition, predicting folding outcomes 15% more accurately than expert groups of scientists[4].

When their parameters are reduced to biologically plausible ranges, DNNs have also demonstrated promise as models of human perception in psychology and cognitive neuroscience. With an architecture inspired by the anatomy of mammalian perceptual cortex[5,6], deep convolutional neural networks (DCNNs) are regarded as the best computational models of object recognition and perceptual categorization judgments in primates[7]. Neuroscientists have compared the activity patterns in intermediate layers of a DCNN's hierarchy to firing patterns recorded from implanted electrophysiology arrays in monkey visual cortex; both the networks and the monkeys seem to recover the same kinds of features at comparable depths of their processing hierarchies[8–11]. There has thus been hope that not only do these models replicate the functional input–output patterns observed in primate object recognition and perceptual similarity judgments, but also that they do so by modelling the hierarchical abstraction algorithms implemented in the primate brain[12].

DNNs often excel due to the vast amounts of training data used, which seems to leave them with a critical vulnerability. Specifically, presenting them with unusual data—discovered by further 'adversarial' machine learning methods designed to fool DNNs—can cause them to issue verdicts that look to human observers like bizarre mistakes. A picture of a panda correctly classified by a DNN can be modified in a way that is imperceptible to humans, but afterwards causes the same network to label it as containing a gibbon[13]; automated vehicles might drive past carefully vandalized stop signs that their recognition systems classify as yield signs[14]; and a female researcher—when accessorized with some 'adversarial glasses'—was repeatedly labelled by state-of-the-art facial detection software as a male co-author[15] (Fig. 1). These findings have curbed the enthusiasm with which some researchers regard DNNs, suggesting that their performance is brittle and cannot be trusted. Despite appearing to derive highly structured, sophisticated knowledge from their large training sets, the discoverers of adversarial examples worried that DNNs merely construct "a Potemkin village that works well on naturally occurring data, but is exposed as fake when one visits points in space that do not have a high probability"[13]. A debate has thus recently developed over whether the patterns that DNNs detect in adversarial examples are 'real' patterns in the signal source or 'fake' conglomerations of noise.

Here, I argue that understanding the implications of adversarial examples requires exploring a third possibility: that at least some of these patterns are artefacts. Artefacts can contain predictive information about target signals that may not be available through other means; but until we understand their origins, they can easily be misinterpreted. Thus, there are presently both costs in simply discarding these patterns and dangers in using them naively; and responding to them wisely requires developing a theory of DNNs' distinctive artefacts.

Adversarial examples

After their discovery, research revealed that many early intuitions about adversarial examples were incorrect[16]. For one, researchers found that an adversarial example created to fool one network was often assigned the same, seemingly incorrect label by other DNNs with different parameters and training sets—which is difficult to explain on the hypothesis that adversarial susceptibility is due to overfitting noise[13,17]. A related hypothesis held that systems were vulnerable to adversarial attacks because their sensitivity to input details was too acute, and so disrupting fine details through transformations like de-noising or rotation could defeat such attacks. While this worked against some adversarial attacks[18]—especially on early 'perturbed images', such as the panda/gibbon image—researchers soon discovered more resilient methods that rely on generating nonsense images or adding apparently meaningless swatches to normal images[19]. These so-called rubbish images can overcome such countermeasures and can be deployed in real-world

Department of Philosophy, The University of Houston, Houston, TX, USA. e-mail: cjbuckner@uh.edu

Fig. 1 | An 'impersonation' attack using 'adversarial glasses'. a,**b**, While wearing these patterned glasses, a female researcher (Sruti Bhagavatula, **a**) was categorized by a facial-detection DNN as a male co-author (Mahmood Sharif, **b**) 88% of the time in their experiments. Figure reproduced with permission from ref. [15], ACM.

settings, as with the stop-to-yield decals or adversarial glasses discussed above.

More recently, research has begun to suggest that vulnerability to adversarial attacks might not be so unusual, after all. One study by Elsayed et al.[20] devised a method to create perturbed images that could fool time-limited humans. Another study by Zhou and Firestone[21] presented human subjects with a series of adversarial examples generated by a variety of different methods. When human subjects were asked to select from a list which labels they thought computer systems were most likely to assign to the adversarial examples, they were able to guess a DNN's preferred labels at rates well above chance. As a result, these authors hypothesize that DNNs are indeed successfully modelling perceptual similarity judgments in humans; they were simply never trained to tell the difference between what something looks like and what it looks like it is (as a reviewer suggested, the way a cloud may look like a dog without looking like it is a dog).

A series of experiments by Ilyas et al.[22] purported to even further redeem DNNs' decisions on adversarial examples. In the first set of experiments, the researchers trained a DCNN to label images in the standard way, and then created a set of adversarial examples that were effective against this network. They then trained a second DCNN on a training set consisting entirely of these adversarial examples—with their seemingly incorrect labels. They then tested this second network on natural images that had been held out from the entire process, and found that the second network was able to reliably produce the correct labels for these images—despite never having seen a single correctly labelled image during its training. In the second set of experiments, they created another artificial dataset that was designed to be free of the 'non-robust' features that could be manipulated by adversarial attacks, by training a DCNN to resist such attacks, and then creating a new dataset using the activation values of this 'robustified' model's final layer. When other networks were trained on this robustified dataset, they were less vulnerable to adversarial attacks; but interestingly their accuracy in classifying held-out naturally distributed data was also correspondingly diminished.

Ilyas et al. argue that these combined results show that the non-robust features inserted into images by adversarial attacks are present in naturally distributed data and carry useful information about the target labels that human observers regard as mistakes. Ilyas et al. argue that this explains why adversarial examples created for one network can fool others with different architectures and training sets: these non-robust features are part of what they call the "inherent geometry of the data", even if humans cannot see them. A special issue of the online computer science journal *Distill* was devoted to evaluating this claim; even those who were sceptical of Ilyas et al.'s conclusion largely replicated and extended their experiments. However, several commentators pushed back on the ambitious interpretation of their results. In particular, Wallace[23] argued that the effects they demonstrated are merely a special case of model distillation (in which information from one model 'leaks' into another, because the incorrect labels for the adversarial examples were derived from a classifier trained on an initial, correctly labelled dataset). Ilyas et al. concede this point, but argue that it does not challenge their central claim, for the only features that could have been distilled from the model trained on adversarial examples are the non-robust ones.

To help put this controversy into context, it may be useful to review the bigger picture. A central goal of machine learning research is to create a classifier that can, by being trained on a finite input context, extract features that generalize to other contexts (and, ideally, beyond any fixed context). This goal is challenging because input data are full of noise and variance, and features learned from training sets may reflect either real patterns or spurious correlations. DNNs attempt to solve this problem by using the activation functions in their hidden layers (and regularization methods) to iteratively transform input signals in a way that accentuates more generalizable patterns, while minimizing less generalizable ones. Adversarial attacks show that in the course of these transformations, DNNs often extract non-robust features. Researchers' attitudes towards non-robust features have tended to extremes, with critics arguing that they must be aggregations of random noise, and defenders holding that they constitute real, reliably predictive patterns that are part of the "inherent geometry of the data".

These positions are not exhaustive, for there remains a class of features that are neither inherent signal nor random noise: processing artefacts. Artefacts are systematic, reproducible patterns in transformed signals that are created by interactions between our instruments and the world. Examples from other domains include perceptual illusions, lens flares, and Doppler effects. We have some familiarity with such artefacts in our own sensory perception, and so might defer to our own judgment when we can ourselves scrutinize the features used by a classifier. However, non-robust features make us nervous in part because we cannot certify their provenance using our own perceptual and cognitive faculties. Until we understand their origins, artefacts are easily misinterpreted; for example, the Doppler effect can cause naive observers to conclude that a train's horn changes frequency as it passes. Though they contain predictive information, we thus may not want to devote a classifier's limited resources to tracking them, since they can lead to incorrect inferences. Once their origins are understood, artefacts can be eliminated with further processing—or even used as a reliable source of evidence, as Doppler shifts allow us to estimate the relative speed of an approaching signal source in weather forecasting.

The deeper riddle of induction

In attempting to determine whether non-robust features are suitable targets for scientific investigation, machine learning researchers are confronting foundational questions in philosophy of science akin to those raised by the philosopher Nelson Goodman[24] in what he called his "new riddle of induction" (expanding on the 'old' riddle posed by David Hume). Goodman challenged us to explain why scientific hypotheses like "emeralds are green" are suitable subjects for empirical investigation, whereas hypotheses like "emeralds are grue"—where grue is defined as "green before time t or blue after time t"—are not. Goodman explored several different ways to cash out the intuitive asymmetry between green and grue: perhaps the problem was that the definition of 'grue' makes reference to limited spatial or temporal coordinates, or that it is defined in terms of other, more basic features. All of these attempts failed, however, because

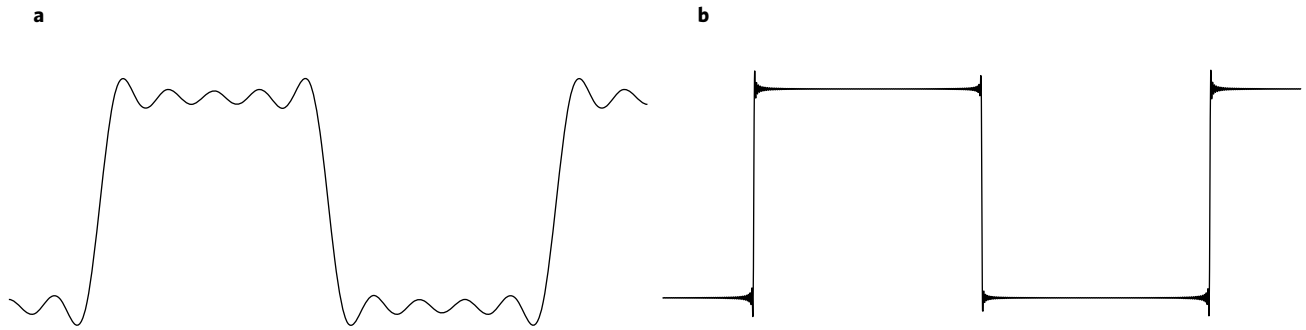

Fig. 2 | A periodic signal function approximated by a Fourier series. A periodic step function can be approximated by a summed series of weighted sinusoids. **a**, The summation with five sinusoids 'overshoots' the peak of the signal before settling into a stable echo. **b**, Although adding more elements (125 total sinusoids) to the summation can 'squash' the overshoot closer to the point of the jump discontinuity, the overshoot is never fully eliminated by adding any finite number of sinusoids.

the supposedly reliable features like green could also be defined in these ways (for example, we could define green as "grue before time *t*, but bleen afterwards"). Goodman concluded pessimistically that the only reason to favour green over grue is its entrenchment in our classificatory and justificatory practices, which might reflect an arbitrary historical preference.

Here, Quine[25] famously suggested that the preference for certain features in scientific practice is not an accident: some jump out at us as natural candidates for investigation because evolution has shaped our perceptual and cognitive faculties to respond to them, given that tracking those features reliably allowed our ancestors survive and reproduce. In short, evolutionary biology provides us with some justification to trust that our default preference for certain 'natural' features will not lead us astray. Other influential philosophers of science followed Quine here; Putnam, Millikan and Boyd[26–28] all emphasized natural features in their philosophy of science, and some of their most influential work explores how we are justified in conducting scientific investigations using these features before we know whether it will pay off.

Whether or not Quine offers a satisfactory solution to Goodman's riddle, we might now wonder whether research on adversarial examples has revealed an alternative fork in this road. The relevance of machine learning to these foundational questions about scientific reasoning has been recognized before[29–31], but the discovery of adversarial examples invites us to reconsider them afresh. Since adversarial attacks are defined as those that change the verdicts of machine learning systems but not those of humans, non-robust features are non-natural in the Quinean sense. However, Quine's arguments did not establish that non-natural features might not also be good subjects for science; and in fact Quine suggested that our reliance on our naive sense of salience was only a waystation in the development of mature science. Mature sciences, Quine suggested, would eventually "slough off the muddy old notion of kind or similarity piecemeal, a vestige here and a vestige there"[25] until all that mattered was which features enabled the most highly confirmed and empirically fecund scientific investigations.

DNNs offer us a different starting point for this scientific expedition—not a narrow path constrained by the tenuous course of hominid evolution and perceptual failings, but rather a wider exploration of feature space enabled by artificially engineered DNNs. If scientific investigation would become more productive by tracking non-natural features—allowing more prediction, control and other scientific goods—then even Quine would be likely to embrace this alternative route to scientific progress. However, we will eventually have to consider a roadblock that has been laid elsewhere in philosophy of science regarding the nature of explanation. Assuming that humans are never able to intuitively grasp these non-robust features—through the use of specialized training or augmented-reality headsets, for example—then it is unlikely that explanations phrased in terms of them should ever produce in us that satisfying feeling of understanding that many regard as the endpoint of successful scientific investigation[32–35]. Perhaps this concern could be at least partially allayed by developing a taxonomy of non-robust features and exploring the properties of each taxon. Doing so might help us decide which non-robust features to discard, and which to retain for inferential and explanatory work.

Beyond signal and noise

Returning to debate over Ilyas et al.'s results, suppose for the sake of argument that there are scientific disciplines in which progress may depend in some crucial way on detecting or modelling predictively useful but human-inscrutable features. To ground the discussion in a speculative but plausible example, let us return to protein folding. For many years in the philosophy of science, protein folding was regarded as paradigm evidence for 'emergent' properties[36]—properties that only appear at higher levels of investigation, and which humans cannot reduce to patterns in lower-level structures. The worry here is that the interactions among amino acids in a protein chain are so complex that humans would never be able to explain biochemical folding principles in terms of lower-level physics[37]. Instead, scientists have relied on a series of analytical 'energy landscape' or 'force field' models that can predict the stability of final fold configurations with some degree of success. These principles are intuitive and elegant once understood, but their elements cannot be reduced to the components of a polypeptide chain in any straightforward manner, and there seem to be stark upper limits on their prediction accuracy. By contrast, AlphaFold[38] on its first entry in the CASP protein-folding competition was able to beat state-of-the-art analytical models on 40 out of 43 of the test proteins, and achieve an unprecedented 15% jump in accuracy across the full test set.

Subsequent work[39] has suggested that the ability of DNNs to so successfully predict final fold configurations may depend on the identification of 'interaction fingerprints', which are distributed across the full polypeptide chain. We might speculate that these interaction fingerprints are like the non-robust features that cause image-classifying networks to be susceptible to adversarial attacks, in that they are complex, spatially distributed, predictively useful, and not amenable to human understanding. Suppose this is all the case, for the sake of argument; whether protein science should rely on such fingerprints depends on whether they are artefacts, and if so whether we can understand their origins.

We have already mentioned how understanding the origins of the Doppler effect allows us to turn a source of confusion into a

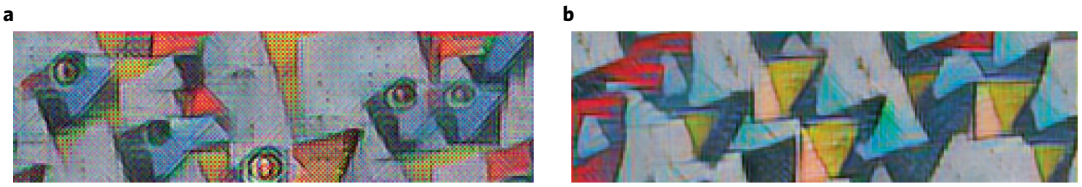

Fig. 3 | Checkerboard artefacts produced by image deconvolution in GANs. a, A generated example containing grainy, unrealistic checkerboard artefacts. **b**, Similar output smoothed by corrective measures. Specifically, Odena et al.[41] attempt to eliminate the artefacts using an alternative deconvolution method—'resize-convolution'—that reshapes images to fit kernels. Figure reproduced with permission from ref. [41], Distill.

source of reliable prediction. To review another example from a discipline more closely related to machine learning, consider the Gibbs phenomenon (Fig. 2). The Gibbs phenomenon is caused by an 'overshoot' in the Fourier series of an input signal when the target function approaches a jump discontinuity. To explain, a Fourier series for some differentiable function is a decomposition of that signal into a weighted summation of sinusoid waves with different amplitudes and frequencies; approximating a function with a Fourier series can help simplify the solution to a variety of mathematical problems. As the number of sinusoids in the summation increases, the Fourier series better approximates a variety of differentiable functions. However, jump discontinuities present an enduring challenge, because adding more sinusoids does not eliminate the overshoot; it only 'squashes' the overshoot closer and closer to the jump discontinuity. Like Ilyas et al.'s non-robust features, this overshoot is useful in the sense that it predicts the location of the jump discontinuity in the target signal; but it can also be misleading about the value of the signal for the duration of the overshoot. Whether or not we should deploy the overshoot in our data analysis depends upon our purposes and how we interpret it.

It is possible—perhaps even likely—that the late-stage, transformed signals at the end of a DNN's processing hierarchy could similarly contain predictive artefacts. We already know that distinctive artefacts can be found in the products of generative adversarial networks (GANs)[40], a DNN architecture that can produce the photorealistic 'deepfakes' that have captivated the popular press. The images produced by the initial version of this technology contained unrealistic-looking 'checkerboard artefacts'[41] that expose the generated exemplars as fakes (Fig. 3). As with Doppler effects, researchers developed a theory of the origins of checkerboard artefacts: an interaction between a GAN's hyperparameter choices and statistical properties of the input data. Once the source of the artefacts was identified, we could predict their appearance and deploy countermeasures to lessen or eliminate them.

However, the possibility that the non-robust features studied by Ilyas et al. are similarly undesirable artefacts needs to be balanced against the possibility that they are inherent patterns in the data available only to 'alien' perception or cognition. It is plausible that DNNs outperform humans in at least some domains because they can track inherent data patterns that humans cannot; progress has been elusive in domains like Go, particle physics and protein folding precisely because these domains are characterized by complex, non-local patterns that resist human understanding. To consider a more mundane example, many animals (and some humans) are tetrachromats, possessing four different colour-detecting cells in their retinas. This allows them to detect stark differences in perceived colours that would be invisible and inscrutable to a trichromat (such as your average human). These additional colour contrasts might be important for understanding the allure of a bird's mating display, which can hinge on presenting an area of plumage that appears highly salient to tetrachromats, but bland to trichromats. In this sense, we may call the additional colours perceived by tetrachromats real patterns in this plumage, even if they are not graspable by trichromats.

The second and related reason we cannot so quickly dismiss Ilyas et al.'s non-robust but useful features as artefacts is that the concept of 'artefact' is surprisingly difficult to define. In characterizing artefacts above, we focused on whether signal patterns are created by an interaction between processing methods and the world; but if artefacts are defined as errors—as patterns introduced by processing that are undesirable—then whether some feature counts as an artefact may be domain- and even purpose-specific. For example, we might imagine scientific applications of the Gibbs phenomenon in which the overshoots are desirable, because they accentuate useful information about the location of a jump discontinuity. If we were using the Fourier transform on electrocardiography data to obtain heartbeat frequency, for example, then the Gibbs phenomenon may help us emphasize aspects of the signal needed to distinguish systolic and diastolic components of heart rhythms. On the other hand, if we were using the transform to gauge the voltage of the heart's output signal to calibrate a pacemaker, the value of the Gibbs overshoot could lead to serious mistakes. Only an understanding of the overshoot's origins could allow us to tailor its use to our purposes in a responsible manner.

Towards a deeper understanding of non-robust features

I suggest three interlocking strategies to advance the current debate over adversarial examples and human-inscrutable science. First and most generally, we need a taxonomy of the non-robust features detected by DNNs. This work has already begun in response to Ilyas et al.'s original finding; Goh[42] proposed at least two different kinds of non-robust features, 'ensembles' and 'containments'. Goh defined ensembles as collections of non-robust and non-useful features which, if sufficiently uncorrelated, could be combined into a single useful and robust feature. We might want to retain such ensembles in mature science. Containments, on the other hand, are interpolations of a useful, robust feature and a useless, non-robust feature—something which seems undesirably gerrymandered in the same way as Goodman's grue. Because containments could always be replaced with a more reliable feature that is at least as predictively useful, they may be regarded as unwanted artefacts.

As we develop a taxonomy of such features, we should also try to uncover conditions that foster the manifestation of each sub-type—that is, to discover which environments, architectures and/or hyperparameter choices tend to produce which types of features. Many choices go into the construction of these networks: number of nodes in each layer, number of layers, types of activation functions, regularization methods, and so on. As we discovered with checkerboard artefacts—they are a product of the stride length chosen for the deconvolution operation—we are likely to find that certain non-robust features are produced only by certain architectures or hyperparameter choices. This would be a step towards anticipating and mitigating the appearance of these features in our outputs, should we decide that they are undesirable. This step would in turn help address a key concern about the opacity of DNNs used in science: a lack of knowledge regarding empirical linkages between the representations learned by networks and the phenomena under investigation[35].

Discovering the set of conditions that foster different types of features would also help with a second strategy for detecting artefacts in DNNs—the method of triangulation. Many sciences already make do with methods of investigation that are not fully trustworthy and cannot be calibrated against gold standard data or independent accounts of ground truth. Sociology, for example, utilizes a variety of different survey and investigative methods, none of which can be regarded as fully trustworthy. A standard method in sociology is triangulation—researchers deploy many qualitatively different methods to ask the same question, and regard an answer to reflect a real pattern if it arises independently from multiple independent methods[43,44]. In other words, we may want to apply many different machine learning methods with different hyperparameters to the same data; if the same type of feature reliably appears in the same way on many different methods, it may be less likely to be an artefact (similar to some existing uses of ensemble learning[45]). A complication of this approach is that without knowing which hyperparameter choices produce which type of feature, we will not know what aspects of models to vary.

Finally, a more multi-dimensional approach to explanatory power in sciences using DNNs for data analysis may soften the blow of 'unintelligible' progress and help us calibrate reliance on non-robust features in particular applications. More generally, philosophers of science have distinguished a variety of different dimensions of explanatory power, only one of which is 'cognitive salience' to humans. Many others—such as non-sensitivity to background conditions, precision, factual accuracy, and degree of integration with background theory[46]—may be satisfied by useful-but-inscrutable features. These other dimensions can be traded off against losses in cognitive salience, providing us with a principled way to decide when non-robust features or even artefacts should be deployed in particular scientific applications.

Conclusion

Researchers should develop a systematic taxonomy of the kinds of features learned by DNNs and tools to distinguish them from one another and gauge their suitability for various scientific projects. The first cut in this taxonomy would divide those features that are reliably predictive from those that are not; this distinction has long been a central focus of research in machine learning and is explored by standard methods like cross-validation. The next cut would distinguish predictive features that are scrutable to humans (robust) from those that humans find inscrutable (non-robust); this is the cut that Ilyas et al., and Zhou and Firestone have begun to explore. Finally, the third cut divides the predictive-but-inscrutable features into artefacts and inherent data patterns detectable only by non-human processing, with the former targeted for more suspicion until a theory of their origins and techniques for mitigation can be deployed; Goh's *Distill* response has made some initial steps here. More research on the last two cuts is urgently needed to understand the full implications of DNNs' susceptibility to adversarial attacks.

Received: 20 March 2020; Accepted: 28 October 2020;
Published online: 23 November 2020

References

1. Silver, D. et al. Mastering the game of go without human knowledge. *Nature* **550**, 354–359 (2017).
2. Shallue, C. J. & Vanderburg, A. Identifying exoplanets with deep learning: a five-planet resonant chain around Kepler-80 and an eighth planet around Kepler-90. *Astron. J.* **155**, 94 (2018).
3. Albertsson, K. et al. Machine learning in high energy physics community white paper. *J. Phys. Conf. Ser.* **1085**, 022008 (2018).
4. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
5. Fukushima, K. Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron. *IEICE Techn. Rep. A* **62**, 658–665 (1979).
6. Hubel, D. H. & Wiesel, T. N. Cortical and callosal connections concerned with the vertical meridian of visual fields in the cat. *J. Neurophysiol.* **30**, 1561–1573 (1967).
7. Rajalingham, R. 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).
8. Guest, O. & Love, B. Levels of representation in a deep learning model of categorization. Preprint at https://doi.org/10.1101/626374 (2019).
9. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
10. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
11. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
12. Buckner, C. Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* **195**, 5339–5372 (2018).
13. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint https://arxiv.org/abs/1412.6572 (2014).
14. Eykholt, K. et al. Robust physical-world attacks on deep learning visual classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* 1625–1634 (IEEE, 2018).
15. Sharif, M., Bhagavatula, S., Bauer, L. & Reiter, M. K. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In *Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security* 1528–1540 (ACM, 2016).
16. Yuan, X., He, P., Zhu, Q. & Li, X. Adversarial examples: attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 2805–2824 (2019).
17. Szegedy, C. et al. Intriguing properties of neural networks. Preprint at https://arxiv.org/abs/1312.6199 (2013).
18. Xu, W., Evans, D. & Qi, Y. Feature squeezing: detecting adversarial examples in deep neural networks. Preprint at https://arxiv.org/abs/1704.01155 (2017).
19. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: high confidence. In *IEEE Conf. Computer Vision and Pattern Recognition* 427–436 (IEEE, 2015).
20. Elsayed, G. F. et al. Adversarial examples that fool both computer vision and time-limited humans. In *Proc. 32nd Int. Conf. Neural Information Processing Systems* 3914–3924 (NeurIPS, 2018).
21. Zhou, Z. & Firestone, C. Humans can decipher adversarial images. *Nat. Commun.* **10**, 1334 (2019).
22. Ilyas, A. et al. Adversarial examples are not bugs, they are features. Preprint at https://arxiv.org/abs/1905.02175 (2019).
23. Wallace, E. A Discussion of 'adversarial examples are not bugs, they are features': learning from incorrectly labeled data. *Distill* **4**, e00019.6 (2019).
24. Goodman, N. *Fact, Fiction, and Forecast* (Harvard Univ. Press, 1983).
25. Quine, W. V. in *Essays in Honor of Carl G. Hempel* 5–23 (Springer, 1969).
26. Boyd, R. Kinds, complexity and multiple realization. *Philos. Stud.* **95**, 67–98 (1999).
27. Millikan, R. G. Historical kinds and the "special sciences". *Philos. Stud.* **95**, 45–65 (1999).
28. Putnam, H. in *Vetus Testamentum* Vol. 7 (ed. Gunderson, K.) 131–193 (Univ. Minnesota Press, 1975).
29. Harman, G. & Kulkarni, S. *Reliable Reasoning: Induction and Statistical Learning Theory* (MIT Press, 2012).
30. Suppes, P. in *Grue! The New Riddle of Induction* (ed. Stalker, D.) 263–272 (Open Court, 1994).
31. Thagard, P. Philosophy and machine learning. *Can. J. Philos.* **20**, 261–276 (1990).
32. Arango-Muñoz, S. The nature of epistemic feelings. *Philos. Psychol.* **27**, 193–211 (2014).
33. Khalifa, K. The role of explanation in understanding. *Br. J. Philos. Sci.* **64**, 161–187 (2013).
34. Potochnik, A. Explanation and understanding. *Eur. J. Philos.* **1**, 29–38 (2011).
35. Sullivan, E. Understanding from machine learning models. *Br. J. Philos. Sci.* https://doi.org/10.1093/bjps/axz035 (2019).
36. Humphreys, P. *Emergence: A Philosophical Account* (Oxford Univ. Press, 2016).
37. Theurer, K. L. Complexity-based theories of emergence: criticisms and constraints. *Int. Stud. Philos. Sci.* **28**, 277–301 (2014).
38. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
39. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
40. Goodfellow, I. NIPS 2016 tutorial: generative adversarial networks. Preprint at https://arxiv.org/abs/1701.00160 (2016).
41. Odena, A., Dumoulin, V. & Olah, C. Deconvolution and checkerboard artifacts. *Distill* **1**, e3 (2016).
42. Goh, G. A Discussion of 'adversarial examples are not bugs, they are features': two examples of useful, non-robust features. *Distill* **4**, e00019.3 (2019).

43. Denzin, N. K. *The Research Act: A Theoretical Introduction to Sociological Methods* (Routledge, 2017).
44. Heesen, R., Bright, L. K. & Zucker, A. Vindicating methodological triangulation. *Synthese* **196**, 3067–3081 (2019).
45. Allman, D., Reiter, A. & Bell, M. A. L. Photoacoustic source detection and reflection artifact removal enabled by deep learning. *IEEE Trans. Med. Imaging* **37**, 1464–1477 (2018).
46. Ylikoski, P. & Kuorikoski, J. Dissecting explanatory power. *Philos. Stud.* **148**, 201–219 (2010).

Acknowledgements

This work has been supported by National Science Foundation grant 2020585.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to C.B.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.