nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-020-00257-z

Shortcut learning in deep neural networks

Robert Geirhos[1,2,4] ✉, Jörn-Henrik Jacobsen[3,4], Claudio Michaelis[1,2,4], Richard Zemel[3,5], Wieland Brendel[1,5], Matthias Bethge[1,5] and Felix A. Wichmann[1,5]

Deep learning has triggered the current rise of artificial intelligence and is the workhorse of today's machine intelligence. Numerous success stories have rapidly spread all over science, industry and society, but its limitations have only recently come into focus. In this Perspective we seek to distil how many of deep learning's failures can be seen as different symptoms of the same underlying problem: shortcut learning. Shortcuts are decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios. Related issues are known in comparative psychology, education and linguistics, suggesting that shortcut learning may be a common characteristic of learning systems, biological and artificial alike. Based on these observations, we develop a set of recommendations for model interpretation and benchmarking, highlighting recent advances in machine learning to improve robustness and transferability from the lab to real-world applications.

The success of deep learning has perhaps overshadowed the need to thoroughly understand the behaviour of deep neural networks (DNNs). In an ever-increasing pace, DNNs were reported as having achieved human-level object classification performance[1], beating world-class human Go, poker and Starcraft players[2,3], detecting cancer from X-ray scans[4], translating text across languages[5], helping combat climate change[6], and accelerating the pace of scientific progress itself[7]. Because of these successes, deep learning has gained a strong influence on our lives and society. At the same time, however, researchers are unsatisfied about the lack of a thorough understanding of the underlying principles and limitations. Tackling this lack of understanding has become an urgent necessity due to the growing societal impact of machine learning applications. If we are to trust algorithms with our lives by being driven in an autonomous vehicle, if our job applications are to be evaluated by neural networks, if our cancer screening results are to be assessed with the help of deep learning, then we indeed need to understand thoroughly: when does deep learning work? When does it fail, and why?

We are currently observing a large number of failure cases, some of which are visualized in Fig. 1. DNNs achieve super-human performance recognizing objects, but even small invisible changes[8] or a different background context[9,10] can completely derail predictions. DNNs can generate a plausible caption for an image, but—worryingly—they can do so without really looking at that image[11]. DNNs can accurately recognize faces, but they show high error rates for faces from marginalized groups[12]. DNNs can predict hiring decisions on the basis of résumés, but the algorithm's decisions are biased towards selecting men[13].

How can this discrepancy between super-human performance on one hand and astonishing failures on the other hand be reconciled? One central observation is that many failure cases are not independent phenomena, but are instead connected in the sense that DNNs follow unintended 'shortcut' strategies. While superficially successful, these strategies typically fail under slightly different circumstances. For instance, a DNN may appear to caption images perfectly well, but describes a typical grass landscape as a 'herd of grazing sheep', revealing 'grass' as an unintended (shortcut) predictor for 'sheep'[14]. Likewise, a language model may appear to have learned to reason, but drops to chance performance when superficial correlations are removed from the dataset[15,16]. Worse yet, a machine classifier successfully detected pneumonia from X-ray scans of a number of hospitals, but its performance was surprisingly low for scans from novel hospitals: the model had unexpectedly learned to identify particular hospital systems with near-perfect accuracy (for example, by detecting a hospital-specific metal token on the scan; see Fig. 1). Together with the hospital's pneumonia prevalence rate it was able to achieve a reasonably good prediction—without learning much about pneumonia at all[17].

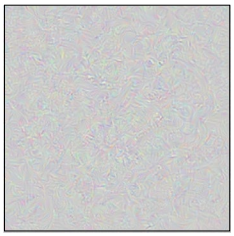
At a principal level, shortcut learning is not a novel phenomenon. The field of machine learning has long aspired to develop a formal understanding of shortcut learning, which has led to an increasing amount of work under different terms such as learning under covariate shift[18], anti-causal learning[19], dataset bias[20], the tank legend[21] and the Clever Hans effect[22]. This Perspective aims to present a unifying view of the various phenomena that can be collectively termed shortcuts, to describe common themes underlying them, and lay out some approaches that are being taken to address them both in theory and in practice.

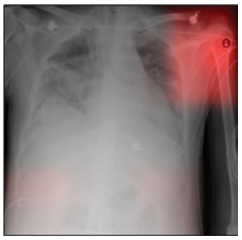
Shortcut learning in biological neural networks

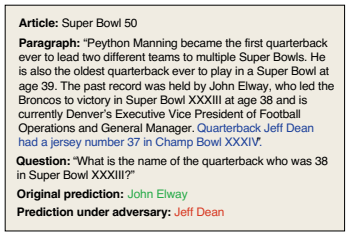
Shortcut learning is not only a problem in machine learning: from the way students learn, to the unintended strategies rats use in behavioural experiments—variants of shortcut learning are also common for 'biological neural networks'. We here point out two examples of unintended learning strategies by biological systems in the hope that this may provide an interesting frame of reference for thinking about shortcut learning within and beyond artificial systems.

Shortcut learning in comparative psychology (learning unintended cues). Rats learned to navigate a complex maze apparently based on subtle colour differences—very surprising given that the rat retina supports at best somewhat crude colour vision. Investigations into this curious finding revealed that the rats had tricked the researchers: they did not use their visual system at all in

[1]University of Tübingen, Tübingen, Germany. [2]International Max Planck Research School for Intelligent Systems, Tübingen, Germany. [3]Vector Institute, University of Toronto, Toronto, Ontario, Canada. [4]These authors contributed equally: Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis. [5]These authors jointly supervised this work: Richard Zemel, Wieland Brendel, Matthias Bethge, Felix A. Wichmann. ✉e-mail: robert.geirhos@wichmannlab.org







	Shane 2018		Zech 2018	Jia 2017
Task for DNN	Caption image	Recognize object	Recognize pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognize primary object	Uses features unrecognizable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Fig. 1 | Examples of shortcut learning. Deep neural networks often solve problems by taking shortcuts instead of learning the intended solution, leading to a lack of generalization and unintuitive failures. This pattern can be observed in many real-world applications. Figure adapted with permission from ref. [14], AI Weirdness (left); ref. [17], PLOS (third from left).

the experiment but simply discriminated the colours by the odour of the colour paint. Once smell was controlled for, the remarkable colour discrimination ability disappeared (N. Rawlins, private communication).

Animals are no strangers to finding simple, unintended solutions that fail unexpectedly: they are prone to learning unintended cues, as shortcut learning is called in comparative psychology and the behavioural neurosciences. When discovering cases of unintended cue learning, there usually was a crucial difference between performance in the experimental paradigm (for example, rewarding rats to identify different colours) and the investigated mental ability (for example, visual colour discrimination). In analogy to machine learning, we have a striking discrepancy between intended and actual learning outcome.

Shortcut learning in education (surface learning). Alice loves history—but at this very moment, she is cursing the subject: after spending weeks immersing herself in the world of Hannibal, she is faced with a number of exam questions that are (in her opinion) to equal parts dull and difficult. "How many elephants did Hannibal employ in his army—19, 34 or 40?" Alice notices that Bob, sitting in front of her, seems to be doing very well. Bob of all people, who had just boasted how he hadn't learned anything except for dates and numbers last night.

In educational research, Bob's reproductive learning strategy would be considered surface learning, an approach that relies on narrow testing conditions where simple discriminative generalization strategies can be highly successful. This fulfils the characteristics of shortcut learning by giving the appearance of good performance, but failing immediately under more general test settings. Worryingly, surface learning helps rather than hurts test performance on typical multiple-choice exams[23]: Bob is likely to receive a better grade than Alice in spite of her focus on understanding. Thus, in analogy to machine learning, we again have a striking discrepancy between intended and actual learning outcome.

Shortcuts defined—a taxonomy of decision rules

With examples of biological shortcut learning in mind, what does shortcut learning in artificial neural networks look like? Fig. 2 shows a simple classification problem that a neural network is trained on. When testing the model on similar data (blue) the network does very well—or so it may seem. Very much like the smart rats that tricked the experimenter, the network uses a shortcut to solve the classification problem by relying on the location of stars and moons instead of their shapes. When location is controlled for, network performance deteriorates to random guessing (red).

Any neural network (or machine learning algorithm) implements a decision rule that defines a relationship between input and output—in this example assigning a category to every input image. Shortcuts are one particular group of decision rules. In order to distinguish them from other decision rules, we here introduce a taxonomy of decision rules (visualized in Fig. 3).

1. All possible decision rules, including non-solutions. Imagine a model that tries to solve the problem of separating stars and moons by predicting 'star' every time it detects a white pixel in the image. This model uses an uninformative feature and does not reach good performance on the data it was trained on, since it implements a poor decision rule (both moon and star images contain white pixels). Typically, interesting problems have numerous non-solutions.
2. Training solutions, including overfitting solutions. In machine learning, it is common practice to split the available data randomly into a training and a test set. The training set is used to guide the model in its selection of a (hopefully useful) decision rule, and the test set is used to check whether the model achieves good performance on similar data it has not seen before. Mathematically, the notion of similarity between training and test set is the assumption that the samples in both sets are drawn from the same distribution (called independent and identically distributed, or i.i.d.). If, however, a decision rule only predicts correctly on the training images but not on the i.i.d. test images, the learning machine uses overfitting features.
3. I.i.d. test solutions, including shortcuts. Decision rules that solve both the training and i.i.d. test set typically score high on standard benchmarks. However, even the simple toy example can be solved through at least three different decision rules: (1) by shape, (2) by counting the number of white pixels (moons are smaller than stars) or (3) by location. It is impossible to distinguish between these by their i.i.d. performance alone. Therefore, one needs to test models on datasets that are systematically different from the i.i.d. training and test data (also called out-of-distribution or o.o.d. data). For example, an o.o.d. test set with randomized object size will instantly invalidate a rule that counts white pixels. Which decision rule is the intended solution is clearly in the eye of the beholder, but humans often

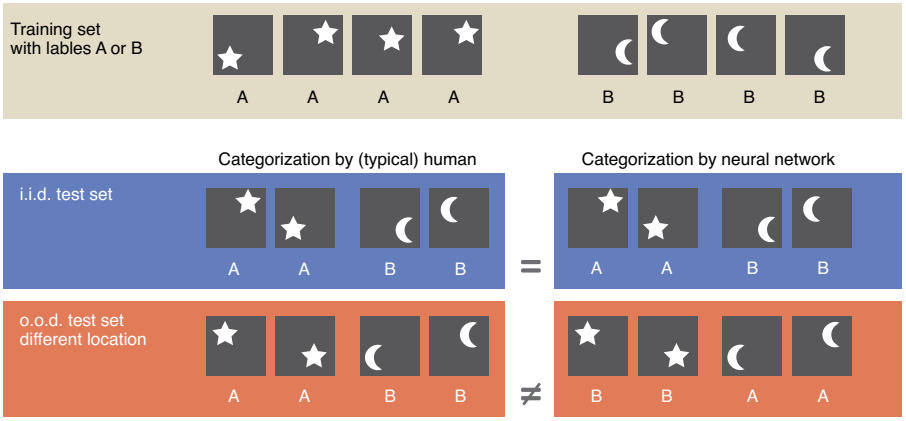


Fig. 2 | Toy example of shortcut learning in neural networks. When trained on a simple dataset of stars and moons, a standard fully connected neural network learns a shortcut strategy: classifying based on the location (stars in the top right or bottom left; moons in the top left or bottom right) rather than the shape of the objects.

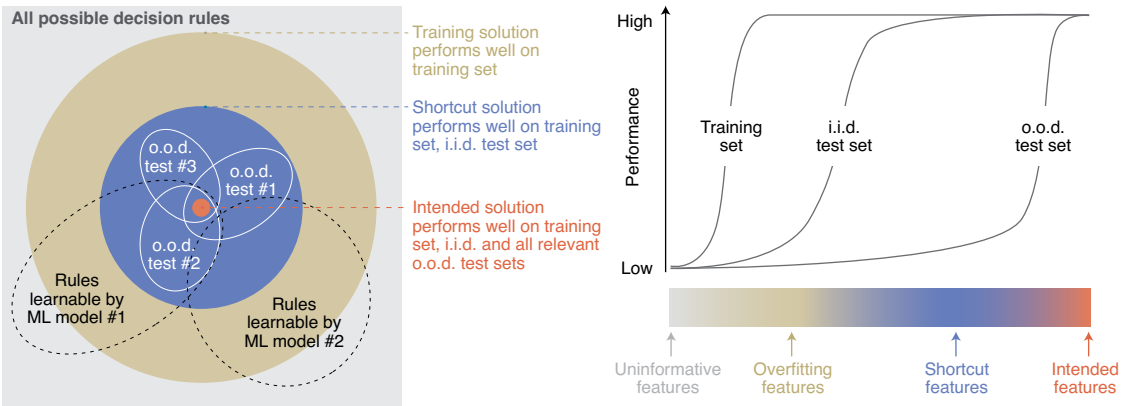


Fig. 3 | Taxonomy of decision rules. Among the set of all possible rules, only some solve the training data. Among the solutions that solve the training data, only some generalize to an i.i.d. test set. Among those solutions, shortcuts fail to generalize to different data (o.o.d. test sets), but the intended solution does generalize.

have clear expectations (here: classification by shape). A standard fully connected neural network trained on this dataset, however, learns a location-based rule. In this case, the network has used a shortcut feature: a feature that helps to perform well on i.i.d. test data but fails in o.o.d. generalization tests. Section A in the Supplementary Information discusses how different areas across deep learning (computer vision, natural language processing, reinforcement learning, and fairness) are affected by shortcut learning.

4. Intended solution. Decision rules that use intended features work well not only on an i.i.d. test set but also on o.o.d. tests where shortcut solutions fail. In the toy example, a decision rule based on object shape (the intended feature) would generalize to objects at a different location. Humans typically have a strong intuition for what the intended solution should be capable of. Yet, for complex problems, intended solutions are mostly impossible to formalize, so machine learning is needed to estimate these solutions from examples. Therefore, the choice of examples, among other aspects, influence how closely the intended solution can be approximated.

Where shortcuts come from

Following this taxonomy, shortcuts are decision rules that perform well on i.i.d. test data but fail on o.o.d. tests, revealing a mismatch between intended and learned solution. It is clear that shortcut learning is to be avoided, but where do shortcuts come from, and what are the defining real-world characteristics of shortcuts that one needs to look out for when assessing a model or task through the lens of shortcut learning? There are two different aspects that one needs to take into account. First, shortcut opportunities (or shortcut features) in the data: possibilities for solving a problem differently than intended. Second, the decision rule: how different features are combined. Together, these aspects determine how a model generalizes.

Dataset shortcut opportunities. What makes a cow a cow? To DNNs, a familiar background can be as important for recognition as the object itself, and sometimes even more important: a cow at an unexpected location (such as a beach rather than grassland) is not classified correctly[9]. Conversely, a lush hilly landscape without any animal at all might be labelled as a 'herd of grazing sheep' by a DNN[14].

This example highlights how a systematic relationship between object and background or context can easily create a shortcut opportunity. And indeed many models base their predictions on context[9,10,24–28]. These so-called dataset biases have long been known to be problematic for machine learning algorithms[20]. Humans, too, are influenced by contextual biases (as evident from faster reaction times when objects appear in the expected context), but their predictions are much less affected when context is missing[29–32].

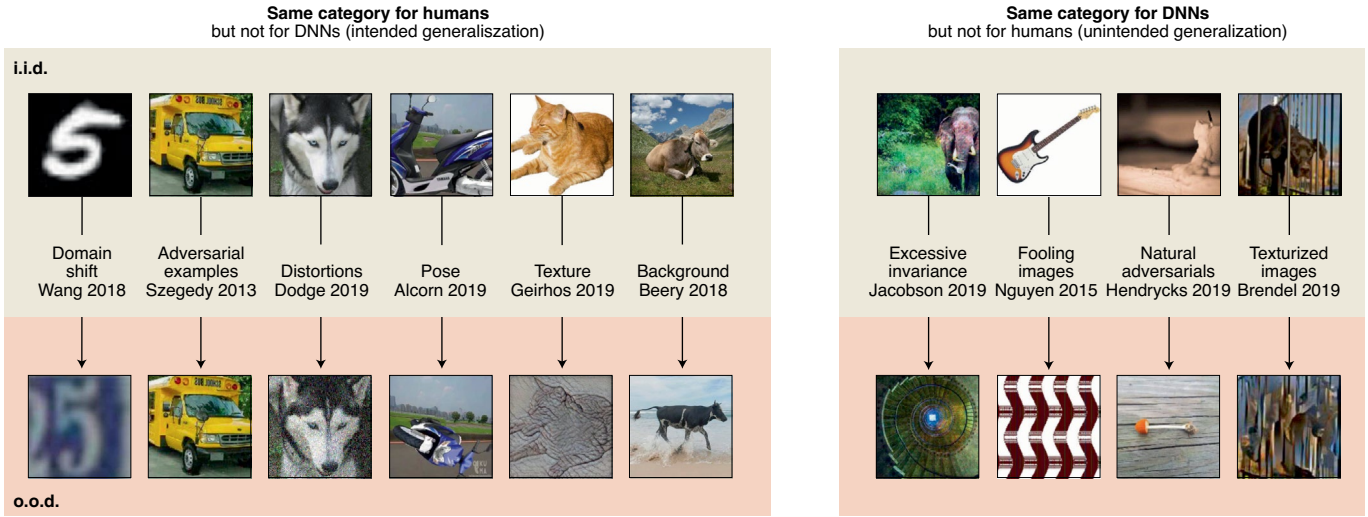


Fig. 4 | Humans and DNNs both generalize, but they generalize very differently. Left: image pairs that belong to the same category for humans, but not for DNNs. Right: image pairs assigned to the same category by a variety of DNNs, but not by humans. Figure adapted with permission from ref. [47], Elsevier (5 images); ref. [8], ICLR (trucks); ref. [38], ICLR (bottom cat); ref. [9], Springer (bottom cow); ref. [45], IEEE (curved pattern).

In addition to shortcut opportunities that are fairly easy to recognize, deep learning has led to the discovery of much more subtle shortcut features, including high frequency patterns that are almost invisible to the human eye[33,34]. Systematic biases are still present even in 'big data' with large volume and variety, and consequently even large real-world datasets usually contain numerous shortcut opportunities[24,35].

Decision rule (shortcuts from discriminative learning). What makes a cat a cat? To standard DNNs, the example image in the bottom row of Fig. 4 (cat with elephant texture) clearly shows an elephant, not a cat. Object textures and other local structures in images are highly useful for object classification in standard datasets[36], and DNNs strongly rely on texture cues for object classification, largely ignoring global object shape[37,38].

Discriminative learning differs from generative modelling by picking any feature that is sufficient to reliably discriminate on a given dataset but the learning machine has no notion of how realistic examples typically look and how the features used for discrimination are combined with other features that define an object. In our example, using textures for object classification becomes problematic if other intended attributes (like shape) are ignored entirely. This exemplifies the importance of feature combination: the definition of an object relies on a (potentially highly non-linear) combination of information from different sources or attributes that influence a decision rule (in cognitive science, this process is called cue combination). A shape-agnostic decision rule that merely relies on texture properties clearly fails to capture the task of object recognition as it is understood for human vision. Of course, being aligned with the human decision rule does not always conform to our intention. In medical or safety-critical applications, for instance, we may instead seek an improvement over human performance.

Within standard discriminative feature learning, some decision rules even depend on a single predictive pixel[39–41] while all other evidence is ignored. In models of animal learning, the blocking effect is a related phenomenon. Once a predictive cue/feature (say, a light flash) has been associated with an outcome (for example, food), animals sometimes fail to associate a new, equally predictive cue with the same outcome[42–44]. In principle, ignoring some evidence can be beneficial. In object recognition, for example, we want the decision rule to be invariant to an object shift. However, undesirable invariance (sometimes called excessive invariance) is harmful.

Generalization reveals shortcuts. What makes a guitar a guitar? When tested on a pattern never seen before, the brown curves at the bottom of Fig. 4 ('fooling images'), standard DNNs predict 'guitar' with high certainty[45]. Exposed by the generalization test, it seems that DNNs learned to detect certain patterns (curved guitar body, strings?) instead of guitars: a successful strategy on training and i.i.d. test data that leads to unintended generalization on o.o.d. data.

This exemplifies the inherent link between shortcut learning and generalization. Often, shortcut learning is discovered through cases of unintended generalization, revealing a mismatch between human-intended and model-learned solution. Interestingly, DNNs do not suffer from a general lack of o.o.d. generalization (Fig. 4)[36,41,45,46]. The set of images that DNNs classify as 'guitar' with high certainty is incredibly big. To humans, only some of these look like guitars, others like patterns (interpretable or abstract) and many more resemble white noise or even look like airplanes, cats or food[8,41,45]. The right side of Fig. 4, for example, highlights a variety of image pairs that have hardly anything in common for humans but belong to the same category for DNNs. Conversely, to the human eye an image's category is not altered by innocuous distribution shifts like rotating objects or adding a bit of noise, but if these changes interact with the shortcut features that DNNs are sensitive to, they completely derail neural network predictions[8,9,38,47–50]. This highlights that generalization failures are neither a failure to learn nor a failure to generalize, but instead a failure to generalize in the intended direction.

Diagnosing and understanding shortcut learning

Many individual elements of shortcut learning have been identified long ago by parts of the machine learning community and some have already seen substantial progress, but currently a variety of approaches are explored without a commonly accepted strategy. We here outline three actionable steps towards diagnosing and understanding shortcut learning, and refer the interested reader to Section B of the Supplementary Information for a discussion of techniques that may help us to mitigate shortcut learning.

Interpreting results carefully. Shortcut learning is most deceptive when gone unnoticed. The following two recommendations may help in this regard.

Distinguishing datasets and underlying abilities. There is often a discrepancy between the simplicity with which a dataset can be

solved and the complexity evoked by the high-level description of the underlying ability. For example, the ImageNet dataset[51] was intended to measure the 'object recognition' ability, but DNNs seem to rely mostly on 'counting texture patches'[36]. Likewise, instead of performing 'natural language inference', some language models simply detect correlated key words[52]. As a consequence, it is important to regularly verify that a dataset is (still) a good proxy for the ability we are truly interested in[52,53].

Morgan's canon for machine learning. Recall the cautionary tale of rats sniffing rather than seeing colour, described in the beginning. There is often a tacit assumption that human-like performance implies human-like strategy[54,55]. This same strategy assumption is paralleled by deep learning: surely, at Marr's implementational level[56], DNNs are different from brains—but if DNNs successfully recognize objects, it seems natural to assume that they are using object shape like humans do[37,38].

Comparative psychology with its long history of comparing mental abilities across species has coined the term anthropomorphism, "the tendency of humans to attribute human-like psychological characteristics to nonhumans on the basis of insufficient empirical evidence"[57], for this fallacy. As a reaction, psychologist Lloyd Morgan developed a conservative guideline for interpreting non-human behaviour known as Morgan's canon: "In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower on the scale of psychological evolution and development"[58]. Picking up on a simple correlation, for example, would be considered a process that stands low on this psychological scale, whereas 'understanding a scene' would be considered much higher. Consequently, we need to interpret machine learning carefully by using what we call 'Morgan's canon for machine learning': never attribute to high-level abilities that which can be adequately explained by shortcut learning.

Towards o.o.d. generalization tests for detecting shortcuts. Testing o.o.d. generalization is the single most important recommendation that we can make: o.o.d. generalization tests need to become the rule rather than the exception.

Making o.o.d. generalization tests a standard practice. In current benchmarks, model performance is usually assessed on an i.i.d. test set. Unfortunately, in real-world settings, the i.i.d. assumption is rarely justified; in fact, this assumption has been called "the big lie in machine learning"[59]. While any metric is typically an approximation of what we truly intend to measure, the i.i.d. performance metric may not be a good approximation as it can often be misleading, giving a false sense of security. We previously described how Bob gets a good grade on a multiple-choice exam through rote learning. Bob's reproductive approach gives the superficial appearance of excellent performance, but it would not generalize to a more challenging test. Worse yet, as long as Bob continues to receive good grades through so-called surface learning, he is unlikely to change his learning strategy.

Educational research suggests to change the type of examination: surface approaches (successful on multiple-choice exams) typically fail on essay questions[23], and so-called deep or transformational learning strategies[60,61] are encouraged; strategies that enable transferring the learned content to novel problems[62]. We can easily see the connection to machine learning—transferring knowledge to novel problems corresponds to testing generalization beyond the narrowly learned setting[63–65]. If model performance is assessed only on i.i.d. test data, we cannot tell whether the model is actually acquiring the ability we think it is, since exploiting shortcuts often leads to deceptively good results on standard metrics[66]. Fortunately, o.o.d. generalization tests are beginning to gain traction across deep learning[54,67–71] and will hopefully become a standard method for benchmarking models in the future (a few current encouraging examples are listed in Box 1).

Designing good o.o.d. tests. We believe that good o.o.d. tests should fulfil at least the following three conditions: First, per definition, there needs to be a clear distribution shift, a shift that may or may not be distinguishable by humans. Second, it should have a well-defined intended solution. Training on natural images while testing on white noise would technically constitute an o.o.d. test but lacks a solution. Third, a good o.o.d. test is a test where the majority of current models struggle. The space of conceivable o.o.d. tests includes

Box 1 | Examples of interesting o.o.d. benchmarks

We here list a few selected, encouraging examples of o.o.d. benchmarks.

Adversarial attacks can be seen as testing on model-specific worst-case o.o.d. data, which makes them an interesting diagnostic tool. If a successful adversarial attack[8] can change model predictions without changing semantic content, this is an indication that something akin to shortcut learning may be occurring[34,84].

Argument Reasoning Comprehension Task (ARCT) with removed shortcuts is a language argument comprehension dataset that follows the idea of removing known shortcut opportunities from the data itself in order to create harder test cases[15].

Cue conflict stimuli like images with conflicting texture and shape information pitch features/cues against each other, such as an intended against an unintended cue[38]. This approach can easily be compared to human responses, even on a detailed image-by-image level[55].

ImageNet-A is a collection of natural images that several state-of-the-art models consistently classify wrongly. It thus benchmarks models on worst-case natural images[46].

ImageNet-C applies 15 different image corruptions to standard test images, an approach we find appealing for its variety and usability[71].

ObjectNet introduces the idea of scientific controls into o.o.d. benchmarking, allowing to disentangle the influence of background, rotation and viewpoint[85].

PACS and other domain generalization datasets require extrapolation beyond i.i.d. data per design by testing on a domain different from training data (for example, cartoon images)[86]. 3D renderers[87,88] may be a promising avenue for additionally controlling factors of variation.

Shift-MNIST/ biased CelebA/unfair dSprites are controlled toy datasets that introduce correlations in the training data (for example, class-predictive pixels or image quality) and record the accuracy drop on clean test data as a way of finding out how prone a given architecture and loss function are to picking up on shortcuts[39–41,89].

Testing surprisingly strong baselines. Complementary to o.o.d. benchmarks, one can test whether a baseline model exceeds expectations despite not using intended features. Examples include using nearest neighbours[90,91], object recognition with local features only[36], reasoning based on single cue words[15,92] or answering questions about a movie without ever showing the movie to a model[93].

While benchmarks are a great way to assess and compare performance, it is equally important to keep in mind that benchmarks tend to follow Goodhart's law over time: "When a measure becomes a target, it ceases to be a good measure".

numerous uninteresting tests. Thus, we want to focus on challenging test cases. As models evolve, generalization benchmarks need to evolve as well, which is exemplified by the Winograd Schema Challenge[72]. Initially designed to overcome shortcut opportunities caused by the open-ended nature of the Turing test, this common-sense reasoning benchmark was scrutinized after modern language models started to perform suspiciously well—and it indeed contained more shortcut opportunities than originally envisioned[73], highlighting the need to evolve tests alongside models.

Why shortcuts are learned. Understanding where shortcuts come from and why they are learned will be key towards mitigating them.

The principle of least effort. Why are machines detecting grass instead of cows[9] or a metal token instead of pneumonia[17]? Exploiting those shortcuts seems easier for DNNs than learning the intended solution. But what determines whether a solution is easy to learn? In linguistics, a related phenomenon is called the 'principle of least effort'[74], the observation that language speakers generally try to minimize the amount of effort involved in communication, while remaining understandable (a central goal of communication). For example, the use of 'plane' is becoming more common than 'airplane', and in pronouncing 'cupboard', 'p' and 'b' are merged into a single sound[75,76]. Interestingly, whether a language change makes it easier for the speaker does not always simply depend on objective measures like word length. On the contrary, this process is shaped by a variety of different factors, including the anatomy (architecture) of the human speech organs and previous language experience (training data).

Understanding the influence of inductive biases. In a similar vein, whether a solution is easy to learn for machines does not simply depend on the data but on all of the four components of a machine learning algorithm: architecture, training data, loss function, and optimization. These components—the inductive bias of a model—influence which solutions are easier to learn than others, and thus ultimately determine whether a shortcut is learned instead of the intended solution[77]. Box 2 provides an overview of the connections between shortcut learning and inductive biases. A few hypotheses have been proposed to explain why models tend to learn simple solutions—often, these are shortcuts—first[78–80]. For instance, DNNs are biased towards learning features linearly decodable from a randomly initialized model[81,82].

Box 2 | Shortcut learning and inductive biases

The four components listed below determine the inductive bias of a model and dataset: the set of assumptions that influence which solutions are learnable, and how readily they can be learned. Although in theory DNNs can approximate any function (given potentially infinite capacity)[94], their inductive bias plays an important role for the types of patterns that they prefer to learn given finite capacity and data.

Structure: architecture. Convolutions make it harder for a model to use location—a prior[95] that is so powerful for natural images that even untrained networks can be used for tasks like image inpainting and denoising[96]. In natural language processing, transformer architectures[97] use attention layers to understand the context by modelling relationships between words. In most cases, however, it is hard to understand the implicit priors in a DNN and even standard elements like ReLU activations can lead to unexpected effects like unwarranted confidence[98].

Experience: training data. Shortcut opportunities are present in most data and rarely disappear by adding more data[33,34,38,52,99]. Modifying the training data to block specific shortcuts has been demonstrated to work for reducing adversarial vulnerability[100] and texture bias[38].

Goal: loss function. The most commonly used loss function for classification, cross-entropy, encourages DNNs to stop learning once a simple predictor is found; a modification can force neural networks to use all available information[41]. Regularization terms that use additional information about the training data have been used to disentangle intended features from shortcut features[39,101].

Learning: optimization. Stochastic gradient descent and its variants bias DNNs towards learning simple functions[102–105]. The learning rate influences which patterns networks focus on: large learning rates lead to learning simple patterns that are shared across examples, while small learning rates facilitate complex pattern learning and memorization[78,106]. The complex interactions between training method and architecture are poorly understood so far; strong claims can only be made for simple cases[107].

Conclusion

"The road reaches every place, the short cut only one"
— James Richardson[83]

Shortcut learning is one of the key roadblocks towards fair, robust, deployable and trustworthy machine learning. While overcoming shortcut learning in its entirety may potentially be impossible, any progress towards mitigating it will lead to a better alignment between learned and intended solutions. This holds the promise that machines behave much more reliably in our complex and ever-changing world, even in situations far away from their training experience. Furthermore, machine decisions would become more transparent, enabling us to detect and remove biases more easily. Currently, the research on shortcut learning is still fragmented into various communities. With this Perspective, we hope to fuel discussions across these different communities and to initiate a movement that pushes for a new standard paradigm of o.o.d. generalization that is able to replace the current i.i.d. tests. To increase understanding and mitigate instances of shortcut learning, we offer the following recommendations:

1. **Connecting the dots: shortcut learning is ubiquitous.** Shortcut learning appears to be a ubiquitous characteristic of learning systems, biological and artificial alike. Many of deep learning's problems are connected through shortcut learning—models exploit dataset shortcut opportunities, select only a few predictive features instead of taking all evidence into account, and consequently suffer from unexpected generalization failures. 'Connecting the dots' between affected areas is likely to facilitate progress.
2. **Interpreting results carefully.** Discovering a shortcut often reveals the existence of an easy solution to a seemingly complex dataset. We will need to exercise great care before attributing high-level abilities like 'object recognition' or 'language understanding' to machines, since there is often a much simpler explanation.
3. **Testing o.o.d. generalization.** Assessing model performance on i.i.d. test data (as the majority of current benchmarks do) is insufficient to distinguish between intended and unintended (shortcut) solutions. Consequently, o.o.d. generalization tests will need to become the rule rather than the exception.
4. **Understanding what makes a solution easy to learn.** DNNs always learn the easiest possible solution to a problem, but understanding which solutions are easy (and thus likely to be learned) requires disentangling the influence of structure (architecture), experience (training data), goal (loss function) and learning (optimization), as well as a thorough understanding of the interactions between these factors.

5. **Asking whether a task should be solved in the first place.** DNNs will often find (shortcut) solutions no matter the task. For instance, they might use shortcuts to assess credit-scores from sensitive demographics. Shortcuts can make questionable or harmful tasks appear perfectly solvable. However, the ability of DNNs to tackle a task with high performance can never justify the task's existence or underlying assumptions. Thus, before assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by artificial intelligence?

Code availability

Code to reproduce the toy experiment (Fig. 2) is available at: https://github.com/rgeirhos/shortcut-perspective.

Received: 25 June 2020; Accepted: 9 October 2020;
Published online: 10 November 2020

References

1. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proc. IEEE Int. Conf. Computer Vision* 1026–1034 (ACM, 2015).
2. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
3. Moravčík, M. et al. Deepstack: expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**, 508–513 (2017).
4. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
5. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. Annual Conf. North American Chapter of the Association for Computational Linguistics* (ACL, 2019).
6. Rolnick, D. et al. Tackling climate change with machine learning. Preprint at https://arxiv.org/abs/1906.05433 (2019).
7. Reichstein, M. et al. Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).
8. Szegedy, C. et al. Intriguing properties of neural networks. In *Proc. Int. Conf. Learning Representations* (ICLR, 2014).
9. Beery, S., Van Horn, G. & Perona, P. Recognition in terra incognita. In *European Conf. Computer Vision* 456–473 (Springer, 2018).
10. Rosenfeld, A., Zemel, R. & Tsotsos, J. K. The elephant in the room. Preprint at https://arxiv.org/abs/1808.03305 (2018).
11. Heuer, H., Monz, C. & Smeulders, A. W. Generating captions without looking beyond objects. Preprint at https://arxiv.org/abs/1610.03708 (2016).
12. Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. ACM Fairness Accountability and Transparency* 77–91 (PMLR, 2018).
13. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* https://reut.rs/2Od9fPr (2018).
14. Shane, J. Do neural nets dream of electric sheep? *AI Wierdness* https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep (2018).
15. Niven, T. & Kao, H.-Y. Probing neural network comprehension of natural language arguments. In *Proc. 57th Annual Meeting of the Association of Computational Linguistics* 4658–4664 (2019).
16. Jia, R. & Liang, P. Adversarial examples for evaluating reading comprehension systems. Preprint at https://arxiv.org/1707.07328 (2017).
17. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
18. Bickel, S., Bru¨ckner, M. & Scheffer, T. Discriminative learning under covariate shift. *J. Mach. Learn. Res.* **10**, 2137–2155 (2009).
19. Schölkopf, B. et al. On causal and anticausal learning. In *Proc. Int. Conf. Machine Learning* 1255–1262 (ICML, 2012).
20. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2011).
21. Branwen, G. The neural net tank urban legend. *Gwern.net* https://www.gwern.net/Tanks (2011).
22. Pfungst, O. *Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology* (Holt, Rinehart and Winston, 1911).
23. Scouller, K. The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. *Higher Educ.* **35**, 453–472 (1998).
24. Wichmann, F. A., Drewes, J., Rosas, P. & Gegenfurtner, K. R. Animal detection in natural scenes: critical features revisited. *J. Vis.* **10**, 6 (2010).
25. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
26. Zhu, Z., Xie, L. & Yuille, A. L. Object recognition with and without objects. In *Proc. 26th Int. Joint Conf. Artificial Intelligence* 3609–3615 (IJCAI, 2017).
27. Wang, J. et al. Visual concepts and compositional voting. *Ann. Math. Sci. Appl.* **3**, 151–188 (2018).
28. Dawson, M., Zisserman, A. & Nellåker, C. From same photo: cheating on visual kinship challenges. In *Asian Conf. Computer Vision* 654–668 (Springer, 2018).
29. Biederman, I. *On the Semantics of a Glance at a Scene* (Erlbaum, 1981).
30. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**, 143–177 (1982).
31. Oliva, A. & Torralba, A. The role of context in object recognition. *Trends Cogn. Sci.* **11**, 520–527 (2007).
32. Castelhano, M. S. & Heaven, C. Scene context influences without scene gist: eye movements guided by spatial associations in visual search. *Psychon. Bull Rev.* **18**, 890–896 (2011).
33. Jo, J. & Bengio, Y. Measuring the tendency of CNNs to learn surface statistical regularities. Preprint at https://arxiv.org/abs/1711.11561 (2017).
34. Ilyas, A. et al. Adversarial examples are not bugs, they are features. In *Proc. Advances NeurIPS* 125–136 (NeurIPS, 2019).
35. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE T. Evolut. Comput.* **1**, 67–82 (1997).
36. Brendel, W. & Bethge, M. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
37. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comp. Biol.* **14**, e1006613 (2018).
38. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
39. Heinze-Deml, C. & Meinshausen, N. Conditional variance penalties and domain shift robustness. Preprint at https://arxiv.org/abs/1710.11469 (2017).
40. Malhotra, G. & Bowers, J. What a difference a pixel makes: an empirical examination of features used by CNNs for categorisation. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
41. Jacobsen, J.-H., Behrmann, J., Zemel, R. & Bethge, M. Excessive invariance causes adversarial vulnerability. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
42. Kamin, L. J. Predictability, surprise, attention, and conditioning. In *Symp. Punishment and Averse Behavior* (eds Campbell, B. A. & Church, R. M.) 279–296 (Appleton-Century-Crofts, 1969).
43. Dickinson, A. *Contemporary Animal Learning Theory* Vol. 1 (CUP Archive, 1980).
44. Bouton, M. E. *Learning and Behavior: A Contemporary Synthesis* (Sinauer Associates, 2007).
45. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* 427–436 (IEEE, 2015).
46. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. Natural adversarial examples. Preprint at https://arxiv.org/abs/1907.07174 (2019).
47. Wang, M. & Deng, W. Deep visual domain adaptation: a survey. *Neurocomputing* **312**, 135–153 (2018).
48. Alcorn, M. A. et al. Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2019).
49. Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.* **20**, 1–25 (2019).
50. Dodge, S. & Karam, L. Human and DNN classification performance on images with quality distortions: a comparative study. *ACM T. Appl. Perc.* **16**, 7 (2019).
51. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
52. Gururangan, S. et al. Annotation artifacts in natural language inference data. In *Proc. Annual Conf. North American Chapter of the Association for Computational Linguistics* (ACL, 2018).
53. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. HellaSwag: can a machine really finish your sentence? In *Proc. 57th Annual Meeting Asscociation of Computational Linguistics* 4791–4800 (ACL, 2019).
54. Borowski, J. et al. The notorious difficulty of comparing human and machine perception. In *Proc. NeurIPS Shared Representations in Human and Machine Intelligence Workshop* (NeurIPS, 2019).

55. Geirhos, R., Meding, K. & Wichmann, F. A. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. Preprint at https://arxiv.org/abs/2006.16736 (2020).
56. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W. H. Freeman and Company, 1982).
57. Buckner, C. *The Comparative Psychology of Artificial Intelligences* (PhilSci Archive, 2019); http://philsci-archive.pitt.edu/16034/
58. Morgan, C. L. *Introduction to Comparative Psychology* (Scribner, 1903).
59. Ghahramani, Z. Panel of workshop on advances in approximate Bayesian inference (AABI) 2017. *YouTube* https://www.youtube.com/watch?v=x1UByHT60mQ (2017).
60. Marton, F. & Säaljö, R. On qualitative differences in learning—II Outcome as a function of the learner's conception of the task. *Br. J. Educ. Psychol.* **46**, 115–127 (1976).
61. Biggs, J. Individual differences in study processes and the quality of learning outcomes. *Higher Educ.* **8**, 381–394 (1979).
62. Chin, C. & Brown, D. E. Learning in science: a comparison of deep and surface approaches. *J. Res. Sci. Teach.* **37**, 109–138 (2000).
63. Marcus, G. F. Rethinking eliminative connectionism. *Cogn. Psychol.* **37**, 243–282 (1998).
64. Kilbertus, N., Parascandolo, G. & Schölkopf, B. Generalization in anti-causal learning. Preprint at https://arxiv.org/abs/1812.00524 (2018).
65. Marcus, G. Deep learning: a critical appraisal. Preprint at https://arxiv.org/abs/1801.00631 (2018).
66. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
67. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
68. Chollet, F. The measure of intelligence. Preprint at https://arxiv.org/abs/1911.01547 (2019).
69. Crosby, M., Beyret, B. & Halina, M. The Animal-AI Olympics. *Nat. Mach. Int.* **1**, 257–257 (2019).
70. Juliani, A. et al. Obstacle tower: a generalization challenge in vision, control, and planning. In *Proc. 28th Int. Joint Conf. Artificial Intelligence* (IJCAI, 2019).
71. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
72. Levesque, H., Davis, E. & Morgenstern, L. The Winograd Schema Challenge. In *13th Int. Conf. Principles of Knowledge Representation and Reasoning* (KR, 2012).
73. Trichelair, P., Emami, A., Trischler, A., Suleman, K. & Cheung, J. C. K. How reasonable are common-sense reasoning tasks: a case-study on the Winograd Schema Challenge and SWAG. In *Proc. Conf. Empirical Methods in Natural Language Processing and Int. Joint Conf. Natural Language Processing* 3373–3378 (ACL, 2019).
74. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
75. Ohala, J. J. The phonetics and phonology of aspects of assimilation. *Papers Lab. Phono.* **1**, 258–275 (1990).
76. Vicentini, A. The economy principle in language. Notes and Observations from early modern English grammars. *Mots Palabras Words* **3**, 37–57 (2003).
77. Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a less artificial intelligence. *Neuron* **103**, 967–979 (2019).
78. Arpit, D. et al. A closer look at memorization in deep networks. In *Proc. Int. Conf. Machine Learning* (ICML, 2017).
79. Valle-Perez, G., Camargo, C. Q. & Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *Proc. Int. Conf. Learning Representations* (ICLR, 2018).
80. Shah, H., Tamuly, K., Raghunathan, A., Jain, P. & Netrapalli, P. The pitfalls of simplicity bias in neural networks. Preprint at https://arxiv.org/abs/2006.07710 (2020).
81. Kalimeris, D. et al. SGD on neural networks learns functions of increasing complexity. In *Proc. Advances NeurIPS* 3496–3506 (NeurIPS, 2019).
82. Hermann, K. L. & Lampinen, A. K. What shapes feature representations? exploring datasets, architectures, and training. Preprint at https://arxiv.org/abs/2006.12433 (2020).
83. Richardson, J. *Vectors: Aphorisms & Ten-Second Essays* (Ausable, 2001).
84. Engstrom, L. et al. A discussion of 'adversarial examples are not bugs, they are features'. *Distill* https://distill.pub/2019/advex-bugs-discussion/ (2019).
85. Barbu, A. et al. ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proc. Advances NeurIPS* 9448–9458 (NeurIPS, 2019).
86. Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proc. IEEE Int. Conf. Computer Vision* (IEEE, 2017).
87. Qiu, W. & Yuille, A. UnrealCV: connecting computer vision to unreal engine. In *European Conf. Computer Vision* 909–916 (Springer, 2016).
88. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A. & Koltun, V. CARLA: an open urban driving simulator. In *Conf. Robot Learning* 1–16 (CoRL, 2017).
89. Creager, E. et al. Flexibly fair representation learning by disentanglement. In *Proc. Int. Conf. Machine Learning* (ICML, 2019).
90. Hays, J. & Efros, A. A. Scene completion using millions of photographs. *ACM Trans. Graph.* **26**, 4 (2007).
91. Hays, J. & Efros, A. A. IM2GPS: estimating geographic information from a single image. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2008).
92. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. & Van Durme, B. Hypothesis only baselines in natural language inference. In *Proc. 7th Joint Conf. Lexical and Computational Semantics* 180–191 (ACL, 2018).
93. Jasani, B., Girdhar, R. & Ramanan, D. Are we asking the right questions in MovieQA? In *Proc. IEEE/CVF Int. Conf. Computer Vision Workshop* (IEEE, 2019).
94. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
95. d'Ascoli, S., Sagun, L., Bruna, J. & Biroli, G. Finding the needle in the haystack with convolutions: on the benefits of architectural bias. In *Proc. Advances NeurIPS* (NeurIPS, 2019).
96. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Deep image prior. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* 9446–9454 (IEEE, 2018).
97. Vaswani, A. et al. Attention is all you need. In *Proc. Advances NeurIPS* 5998–6008 (NeurIPS, 2017).
98. Hein, M., Andriushchenko, M. & Bitterwolf, J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* 41–50 (IEEE, 2019).
99. Lehman, J. et al. The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. *Art. Life* **26**, 274–306 (2020).
100. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proc. Int. Conf. Learning Representations* (ICLR, 2018).
101. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at https://arxiv.org/abs/1907.02893 (2019).
102. Wu, L., Zhu, Z. & E, W. Towards understanding generalization of deep learning: perspective of loss landscapes. Preprint at https://arxiv.org/abs/1706.10239 (2017).
103. De Palma, G., Kiani, B. T. & Lloyd, S. Deep neural networks are biased towards simple functions. Preprint at https://arxiv.org/abs/1812.10156 (2018).
104. Valle-Perez, G., Camargo, C. Q. & Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *Proc. Int. Conf. Learning Representations* (ICLR, 2019).
105. Sun, K. & Nielsen, F. Lightlike neuromanifolds, Occam's razor and deep learning. Preprint at https://arxiv.org/abs/1905.11027 (2019).
106. Li, Y., Wei, C. & Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Proc. Advances NeurIPS* 11674–11685 (NeurIPS, 2019).
107. Bartlett, P. L., Long, P. M., Lugosi, G. & Tsigler, A. Benign overfitting in linear regression. *Proc. Natl Acad Sci. USA* https://doi.org/10.1073/pnas.1907378117 (2019).

Acknowledgements

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting R.G. and C.M.; the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for supporting C.M. via grant EC 479/1-1; the Collaborative Research Center (Projektnummer 276693517—SFB 1233: Robust Vision) for supporting M.B. and F.A.W.; the German Federal Ministry of Education and Research through the Tübingen AI Center (FKZ 01IS18039A) for supporting W.B. and M.B.; as well as the Natural Sciences and Engineering Research Council of Canada and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 for supporting R.Z. The authors would like to thank J. Borowski, M. Burg, S. Cadena, A. S. Ecker, L. Eisenberg, R. Fleming, I. Fründ, S. Greiner, F. Grießer, S. Keshvari, R. Kessler, D. Klindt, M. Kümmerer, B. Mitzkus, H. Nienborg, J. Rauber, E. Rusak, S. Schneider, L. Schott, T. Sering, Y. Sharma, M. Tangemann, R. Zimmermann and T. Wallis for helpful discussions.

Author contributions

The project was initiated by R.G. and C.M. and led by R.G. with support from C.M. and J.J.; F.A.W. added the cognitive science and neuroscience connection; M.B. and W.B. reshaped the initial thrust of the perspective and together with R.Z. supervised the machine learning components. The toy experiment was conducted by J.J. with input from R.G. and C.M. Most figures were designed by R.G. and W.B. with input from all other authors. Figure 2 (left) was conceived by M.B. The first draft was written by R.G., J.J. and C.M. with input from F.A.W. All authors contributed to the final version and provided critical revisions from different perspectives.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00257-z.

Correspondence should be addressed to R.G.