Shortcut Learning in Deep Neural Networks

Robert Geirhos[1,2,*,§], Jörn-Henrik Jacobsen[3,*], Claudio Michaelis[1,2,*],
Richard Zemel[†,3], Wieland Brendel[†,1], Matthias Bethge[†,1] & Felix A. Wichmann[†,1]

[1]*University of Tübingen, Germany*
[2]*International Max Planck Research School for Intelligent Systems, Germany*
[3]*University of Toronto, Vector Institute, Canada*
[*]*Joint first /* [†] *joint senior authors*
[§]*To whom correspondence should be addressed:* `robert.geirhos@wichmannlab.org`
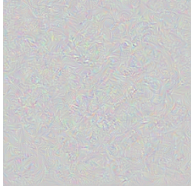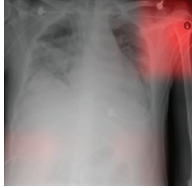
Abstract

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 problem 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.

1   Introduction

If science was a journey, then its destination would be the discovery of simple explanations to complex phenomena. There was a time when the existence of tides, the planet's orbit around the sun, and the observation that "things fall down" were all largely considered to be independent phenomena—until 1687, when Isaac Newton formulated his law of gravitation that provided an elegantly simple explanation to all of these (and many more). Physics has made tremendous progress over the last few centuries, but the thriving field of deep learning is still very much at the beginning of its journey—often lacking a detailed understanding of the underlying principles.

For some time, the tremendous 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

	Caption image	Recognise object	Recognise pneumonia	Answer question
Task for DNN	Caption image	Recognise object	Recognise 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 recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Figure 1. Deep neural networks often solve problems by taking shortcuts instead of learning the intended solution, leading to a lack of generalisation and unintuitive failures. This pattern can be observed in many real-world applications.

the same time, however, researchers are unsatisfied about the lack of a deeper understanding of the underlying principles and limitations. Different from the past, tackling this lack of understanding is not a purely scientific endeavour anymore but 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?

In terms of understanding the limitations of deep learning, we are currently observing a large number of failure cases, some of which are visualised in Figure 1. DNNs achieve super-human performance recognising 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 ever looking at that image [11]. DNNs can accurately recognise faces, but they show high error rates for faces from minority 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 classify cows perfectly well—but fails when tested on pictures where cows appear outside the typical grass landscape, revealing "grass" as an unintended (shortcut) predictor for "cow" [9]. Likewise, a language model may appear to have learned to reason—but drops to chance performance when superficial correlations are removed from the dataset [14]. 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 (e.g. by detecting a hospital-specific metal token on the scan, see Figure 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 [15].

At a principal level, shortcut learning is not a novel phenomenon. The field of machine learning with its strong mathematical underpinnings has long aspired to develop a formal

understanding of shortcut learning which has led to a variety of mathematical concepts and an increasing amount of work under different terms such as *learning under covariate shift* [16], *anti-causal learning* [17], *dataset bias* [18], the *tank legend* [19] and the *Clever Hans effect* [20]. 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 the approaches that are being taken to address them both in theory and in practice.

The structure of this perspective is as follows. Starting from an intuitive level, we introduce shortcut learning across biological neural networks (Section 2) and then approach a more systematic level by introducing a taxonomy (Section 3) and by investigating the origins of shortcuts (Section 4). In Section 5, we highlight how these characteristics affect different areas of deep learning (Computer Vision, Natural Language Processing, Agent-based Learning, Fairness). The remainder of this perspective identifies actionable strategies towards diagnosing and understanding shortcut learning (Section 6) as well as current research directions attempting to overcome shortcut learning (Section 7). Overall, our selection of examples is biased towards Computer Vision since this is one of the areas where deep learning has had its biggest successes, and an area where examples are particularly easy to visualise. We hope that this perspective facilitates the awareness for shortcut learning and motivates new research to tackle this fundamental challenge we currently face in machine learning.

2 Shortcut learning in biological neural networks

Shortcut learning typically reveals itself by a strong discrepancy between intended and actual learning strategy, causing an unexpected failure. Interestingly, machine learning is not alone with this issue: 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 natural systems in the hope that this may provide an interesting frame of reference for thinking about shortcut learning within and beyond artificial systems.

2.1 Shortcut learning in Comparative Psychology: unintended cue learning

Rats learned to navigate a complex maze apparently based on subtle colour differences— very surprising given that the rat retina has only rudimentary machinery to support at best somewhat crude colour vision. Intensive investigation into this curious finding revealed that the rats had tricked the researchers: They did not use their visual system at all in the experiment and instead simply discriminated the colours by the odour of the colour paint used on the walls of the maze. Once smell was controlled for, the remarkable colour discrimination ability disappeared ...[1]

Animals are no strangers to finding simple, unintended solutions that fail unexpectedly: They are prone to *unintended cue learning*, as shortcut learning is called in Comparative Psychology and the Behavioural Neurosciences. When discovering cases of unintended cue learning, one typically has to acknowledge that there was a crucial difference between

---

[1]Nicholas Rawlins, personal communication with F.A.W. some time in the early 1990s, confirmed via email on 12.11.2019.

performance in a given experimental paradigm (e.g. rewarding rats to identify different colours) and the investigated mental ability one is actually interested in (e.g. visual colour discrimination). In analogy to machine learning, we have a striking discrepancy between intended and actual learning outcome.

2.2   Shortcut learning in Education: surface learning

Alice loves history. Always has, probably always will. At this very moment, however, she is cursing the subject: After spending weeks immersing herself in the world of Hannibal and his exploits in the Roman Empire, she is now 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 had learned the whole book chapter by rote 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 generalisation 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 [21]: Bob is likely to receive a good grade, and judging from grades alone Bob would appear to be a much better student 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.

3   Shortcuts defined: a taxonomy of decision rules

With examples of biological shortcut learning in mind (examples which we will return to in Section 6), what does shortcut learning in artificial neural networks look like? Figure 2 shows a simple classification problem that a neural network is trained on (distinguishing a star from a moon).[2] 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. When location is controlled for, network performance deteriorates to random guessing (red). In this case (as is typical for object recognition), classification based on object shape would have been the intended solution, even though the difference between intended and shortcut solution is not something a neural network can possibly infer from the training data.

On a general level, any neural network (or machine learning algorithm) implements a decision rule which defines a relationship between input and output—in this example assigning a category to every input image. Shortcuts, the focus of this article, are one particular group of decision rules. In order to distinguish them from other decision rules, we here introduce a taxonomy of decision rules (visualised in Figure 3), starting from a very general rule and subsequently adding more constraints until we approach the intended solution.

---

[2]Code is available from `https://github.com/rgeirhos/shortcut-perspective`.
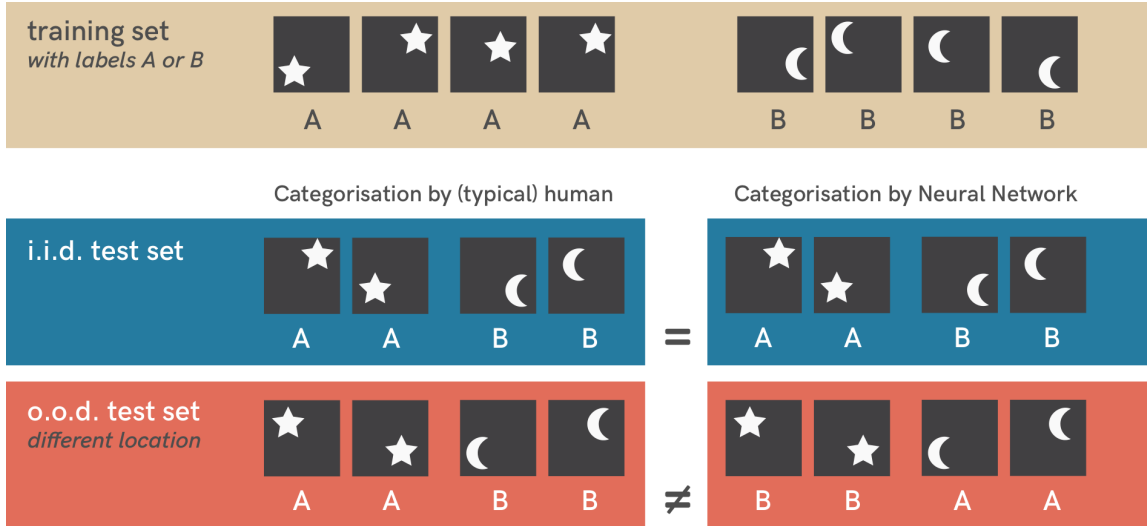
Figure 2. Toy example of shortcut learning in neural networks. When trained on a simple dataset of stars and moons (top row), a standard neural network (three layers, fully connected) can easily categorise novel similar exemplars (mathematically termed i.i.d. test set, defined later in Section 3). However, testing it on a slightly different dataset (o.o.d. test set, bottom row) reveals a shortcut strategy: The network has learned to associate object location with a category. During training, stars were always shown in the top right or bottom left of an image; moons in the top left or bottom right. This pattern is still present in samples from the i.i.d. test set (middle row) but not in o.o.d. test images (bottom row), exposing the shortcut.

(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* (the grey area in Figure 3) 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 an abundant amount of 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 commonly referred to in machine learning is the assumption that the samples in both sets are drawn from the same distribution. This is the case if both the data generation mechanism and the sampling mechanism are identical. In practice this is achieved by randomising the split between training and test set. The test set is then called independent and identically distributed (i.i.d.) with regard to the training set. In order to achieve high average performance on the test set, a model needs to learn a function that is approximately correct within a subset of the input domain which covers most of the probability of the distribution. If a function is learned that yields the correct output on the training images but not on the i.i.d. test images, the learning machine uses *overfitting features* (the blue area in Figure 3).
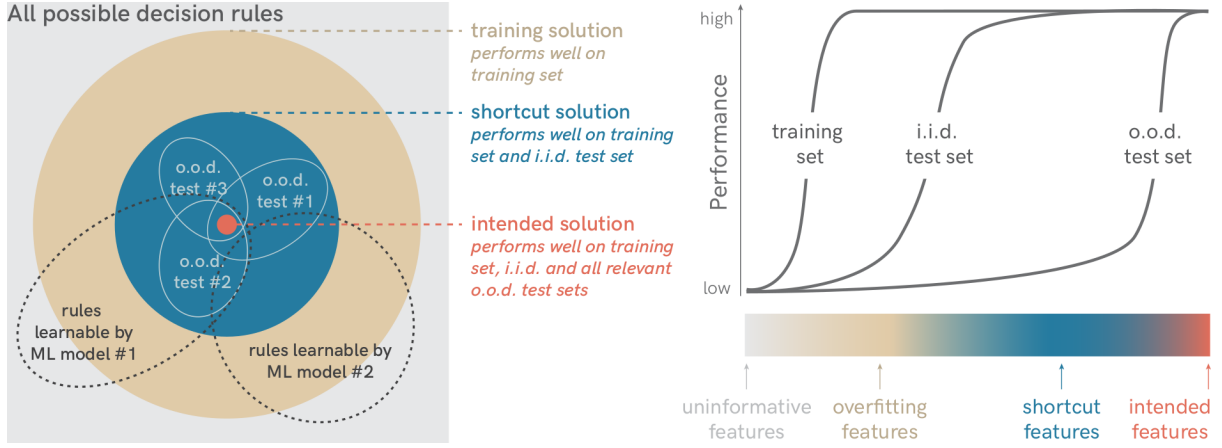
Figure 3. Taxonomy of decision rules. Among the set of all possible rules, only some solve the training data. Among the solutions that solve the training data, only some generalise to an i.i.d. test set. Among those solutions, shortcuts fail to generalise to different data (o.o.d. test sets), but the intended solution does generalise.

(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 benchmark leaderboards. However, even the simple toy example can be solved through at least three different decision rules: (a) by shape, (b) by counting the number of white pixels (moons are smaller than stars) or (c) by location (which was correlated with object category in the training and i.i.d. test sets). As long as tests are performed only on i.i.d. data, it is impossible to distinguish between these. However, one can instead 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 randomised 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 have clear expectations. In our toy example, humans typically classify by shape. A standard fully connected neural network[3] trained on this dataset, however, learns a location-based rule (see Figure 2). In this case, the network has used a *shortcut feature* (the blue area in Figure 3): a feature that helps to perform well on i.i.d. test data but fails in o.o.d. generalisation tests.

(4) intended solution

Decision rules that use the *intended features* (the red area in Figure 3) work well not only on an i.i.d. test set but also perform as intended on o.o.d. tests, where shortcut solutions fail. In the toy example, a decision rule based on object shape (the intended feature) would generalise to objects at a different location or with a different size. 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 formalise, 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.

---

[3]A convolutional (rather than fully connected) network would be prevented from taking this shortcut by design.

4 Shortcuts: where do they 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 (Section 4.1). Second, feature combination: how different features are combined to form a decision rule (Section 4.2). Together, these aspects determine how a model generalises (Section 4.3).

4.1 Dataset: shortcut opportunities
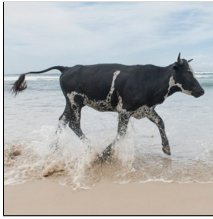
What makes a cow a cow? To DNNs, a familiar background can be as important for recognition as the object itself, and sometimes even more important: A cow at an unexpected location (such as a beach rather than grassland) is not classified correctly [9]. Conversely, a lush hilly landscape without any animal at all might be labelled as a "herd of grazing sheep" by a DNN [22].

This example highlights how a systematic relationship between object and background or context can easily create a shortcut opportunity. If cows happen to be on grassland for most of the training data, detecting grass instead of cows becomes a successful strategy for solving a classification problem in an unintended way; and indeed many models base their predictions on context [23, 24, 25, 26, 9, 27, 10]. Many shortcut opportunities are a consequence of natural relationships, since grazing cows are typically surrounded by grassland rather than water. These so-called *dataset biases* have long been known to be problematic for machine learning algorithms [18]. 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 [28, 29, 30, 31]. In addition to shortcut opportunities that are fairly easy to recognise, 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 [32, 33]. Whether easy to recognise or hard to detect, it is becoming more and more evident that shortcut opportunities are by no means disappearing when the size of a dataset is simply scaled up by some orders of magnitude (in the hope that this is sufficient to sample the diverse world that we live in [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. Overall, it is quite clear that data alone rarely constrains a model sufficiently, and that data cannot replace making assumptions [35]. The totality of all assumptions that a model incorporates (such as, e.g., the choice of architecture) is called the *inductive bias* of a model and will be discussed in more detail in Section 6.3.

4.2 Decision rule: shortcuts from discriminative learning

What makes a cat a cat? To standard DNNs, the example image on the left 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].

In many cases, relying on object textures can be sufficient to solve an object categorisation task. Obviously, however, texture is only one of many attributes that define an object. Discriminative learning differs from generative modeling 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 like 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.[4] In the example of the cat with elephant texture above, 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. While the model uses an important attribute (texture) it tends to equate it with the definition of the object missing out other important attributes such as shape. 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.
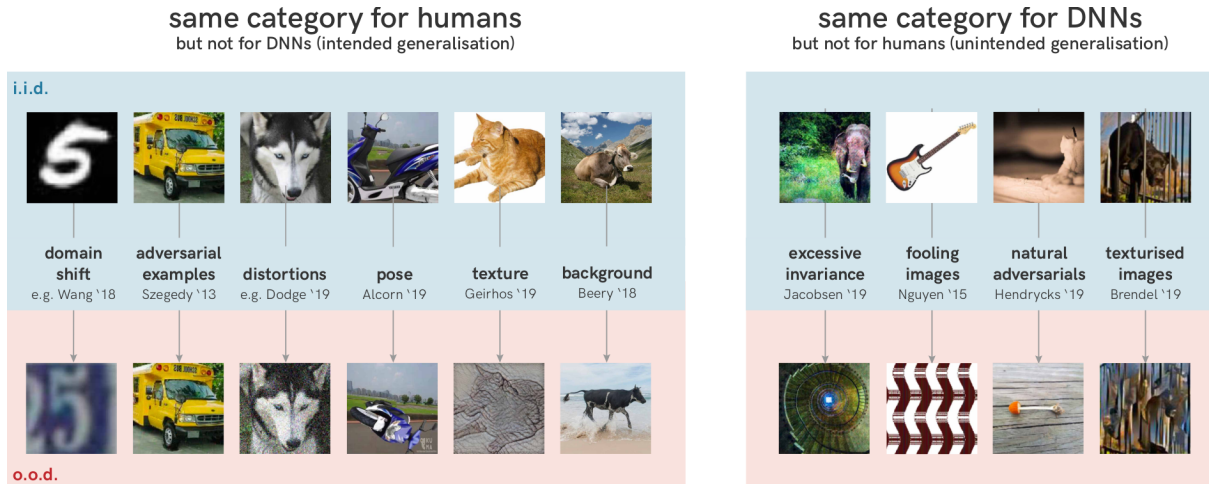
Inferring human-interpretable object attributes like shape or texture from an image requires specific nonlinear computations. In typical end-to-end discriminative learning, this again may be prone to shortcut learning. Standard DNNs do not impose any human-interpretability requirements on intermediate image representations and thus might be severely biased to the extraction of overly simplistic features which only generalise under the specific design of the particular dataset used but easily fail otherwise. Discriminative feature learning goes as far that some decision rules only depend on a single predictive pixel [39, 40, 41] while all other evidence is ignored.[5] 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 and modern machine learning models can be invariant to almost all features that humans would rely on when classifying an image [41].

[4]In Cognitive Science, this process is called *cue combination*.

[5]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 (e.g. food), animals sometimes fail to associate a new, equally predictive cues with the same outcome [42, 43, 44].

same category for humans
but not for DNNs (intended generalisation)

same category for DNNs
but not for humans (unintended generalisation)

i.i.d.

domain shift	adversarial examples	distortions	pose	texture	background
e.g. Wang '18	Szegedy '13	e.g. Dodge '19	Alcorn '19	Geirhos '19	Beery '18

excessive invariance	fooling images	natural adversarials	texturised images
Jacobsen '19	Nguyen '15	Hendrycks '19	Brendel '19

o.o.d.

Figure 4. Both human and machine vision generalise, but they generalise very differently. Left: image pairs that belong to the same category for humans, but not for DNNs. Right: images pairs assigned to the same category by a variety of DNNs, but not by humans.

4.3 Generalisation: how shortcuts can be revealed
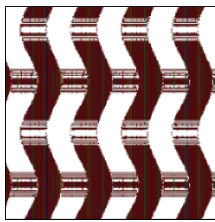
What makes a guitar a guitar? When tested on this pattern never seen before, standard DNNs predict "guitar" with high certainty [45]. Exposed by the generalisation test, it seems that DNNs learned to detect certain patterns (curved guitar body? strings?) instead of guitars: a successful strategy on training and i.i.d. test data that leads to unintended generalisation on o.o.d. data.

This exemplifies the inherent link between shortcut learning and generalisation. By itself, generalisation is not a part of shortcut learning—but more often than not, shortcut learning is discovered through cases of unintended generalisation, revealing a mismatch between human-intended and model-learned solution. Interestingly, DNNs do not suffer from a general lack of o.o.d. generalisation (Figure 4) [45, 36, 46, 41]. DNNs recognise guitars even if only some abstract pattern is left—however, this remarkable generalisation performance is undesired, at least in this case. In fact, 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, 45, 41]. Figure 4 on the right, 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, 47, 9, 48, 49, 50, 38]. This highlights that generalisation failures are neither a failure to learn nor a failure to generalise at all, but instead a failure to generalise in the intended direction—generalisation and robustness can be considered the flip side of shortcut learning. Using a certain set of features creates insensitivity towards other features. Only if the selected features are still present after a distribution shift, a model generalises o.o.d.

9

5 Shortcut learning across deep learning

Taken together, we have seen how shortcuts are based on dataset shortcut opportunities and discriminative feature learing that result in a failure to generalise as intended. We will now turn to specific application areas, and discover how this general pattern appears across Computer Vision, Natural Language Processing, Agent-based (Reinforcement) Learning and Fairness / algorithmic decision-making. While shortcut learning is certainly not limited to these areas, they might be the most prominent ones where the problem has been observed.

Computer Vision To humans, for example, a photograph of a car still shows the same car even when the image is slightly transformed. To DNNs, in contrast, innocuous transformations can completely change predictions. This has been reported in various cases such as shifting the image by a few pixels [47], rotating the object [49], adding a bit of random noise or blur [51, 50, 52, 53] or (as discussed earlier) by changing background [9] or texture while keeping the shape intact [38] (see Figure 4 for examples). Some key problems in Computer Vision are linked to shortcut learning. For example, transferring model performance across datasets (*domain transfer*) is challenging because models often use domain-specific shortcut features, and shortcuts limit the usefulness of unsupervised representations [54]. Furthermore, *adversarial examples* are particularly tiny changes to an input image that completely derail model predictions [8] (an example is shown in Figure 4). Invisible to the human eye, those changes modify highly predictive patterns that DNNs use to classify objects [33]. In this sense, adversarial examples—one of the most severe failure cases of neural networks—can at least partly be interpreted as a consequence of shortcut learning.

Natural Language Processing The widely used language model BERT has been found to rely on superficial cue words. For instance, it learned that within a dataset of natural language arguments, detecting the presence of "not" was sufficient to perform above chance in finding the correct line of argumentation. This strategy turned out to be very useful for drawing a conclusion without understanding the content of a sentence [14]. Natural Language Processing suffers from very similar problems as Computer Vision and other fields. Shortcut learning starts from various dataset biases such as annotation artefacts [55, 56, 57, 58]. Feature combination crucially depends on shortcut features like word length [59, 60, 14, 61], and consequently leads to a severe lack of robustness such as an inability to generalise to more challenging test conditions [62, 63, 64, 65]. Attempts like incorporating a certain degree of unsupervised training as employed in prominent language models like BERT [5] and GPT-2 [66] did not resolve the problem of shortcut learning [14].

Agent-based (Reinforcement) Learning Instead of learning how to play Tetris, an algorithm simply learned to pause the game to evade losing [67]. Systems of Agent-based Learning are usually trained using Reinforcement Learning and related approaches such as evolutionary algorithms. In both cases, designing a good reward function is crucial, since a reward function measures how close a system is to solving the problem. However, they all too often contain unexpected shortcuts that allow for so-called *reward hacking* [68]. The existence of loopholes exploited by machines that follow the letter (and not the spirit) of the reward function highlight how difficult it is to design a shortcut-free reward function [69]. Reinforcement Learning is also a widely used method in Robotics, where there is a commonly observed *generalisation* or *reality gap* between simulated training

environment and real-world use case. This can be thought of as a consequence of narrow shortcut learning by adapting to specific details of the simulation. Introducing additional variation in colour, size, texture, lighting, etc. helps a lot in closing this gap [70, 71].

Fairness & algorithmic decision-making Tasked to predict strong candidates on the basis of their résumés, a hiring tool developed by Amazon was found to be biased towards preferring men. The model, trained on previous human decisions, found gender to be such a strong predictor that even removing applicant names would not help: The model always found a way around, for instance by inferring gender from all-woman college names [13]. This exemplifies how some—but not all—problems of (un)fair algorithmic decision-making are linked to shortcut learning: Once a predictive feature is found by a model, even if it is just an artifact of the dataset, the model's decision rule may depend entirely on the shortcut feature. When human biases are not only replicated, but worsened by a machine, this is referred to as *bias amplification* [72]. Other shortcut strategies include focusing on the majority group in a dataset while accepting high error rates for underrepresented groups [12, 73], which can amplify existing societal disparities and even create new ones over time [74]. In the dynamical setting a related problem is called *disparity amplification* [74], where sequential feedback loops may amplify a model's reliance on a majority group. It should be emphasised, however, that fairness is an active research area of machine learning closely related to invariance learning that might be useful to quantify and overcome biases of both machine and human decision making.

6 Diagnosing and understanding shortcut learning

Shortcut learning currently occurs across deep learning, causing machines to fail unexpectedly. 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 analysing shortcut learning.

6.1 Interpreting results carefully

Distinguishing datasets and underlying abilities Shortcut learning is most deceptive when gone unnoticed. The most popular benchmarks in machine learning still rely on i.i.d. testing which drags attention away from the need to verify how closely this test performance measures the *underlying ability* one is actually interested in. For example, the ImageNet dataset [75] was intended to measure the ability "object recognition", but DNNs seem to rely mostly on "counting texture patches" [36]. Likewise, instead of performing "natural language inference", some language models perform well on datasets by simply detecting correlated key words [56]. Whenever there is a discrepancy between the simplicity with which a dataset (e.g. ImageNet, SQuAD) can be solved and the complexity evoked by the high-level description of the underlying ability (e.g. object recognition, scene understanding, argument comprehension), it is important to bear in mind that a dataset is useful only for as long as it is a good proxy for the ability one is actually interested in [56, 76]. We would hardly be intrigued by reproducing human-defined labels on datasets per se (a lookup table would do just as well in this case)—it is the underlying generalisation ability that we truly intend to measure, and ultimately improve upon.

Morgan's Canon for machine learning Recall the cautionary tale of rats sniffing rather than seeing colour, described in Section 2.1. Animals often trick experimenters by solving an experimental paradigm (i.e., dataset) in an unintended way without using the underlying ability one is actually interested in. This highlights how incredibly difficult it can be for humans to imagine solving a tough challenge in any other way than the human way: Surely, at Marr's implementational level [77] there may be differences between rat and human colour discrimination. But at the algorithmic level there is often a tacit assumption that human-like performance implies human-like strategy (or algorithm) [78]. This *same strategy assumption* is paralleled by deep learning: Surely, DNN units are different from biological neurons—but if DNNs successfully recognise objects, it seems natural to assume that they are using object shape like humans do [37, 36, 38].

Comparative Psychology with its long history of comparing mental abilities across species has coined a term for the fallacy to confuse human-centered interpretations of an observed behaviour and the actual behaviour at hand (which often has a much simpler explanation): *anthropomorphism*, "the tendency of humans to attribute human-like psychological characteristics to nonhumans on the basis of insufficient empirical evidence" [79, p. 5]. As a reaction to the widespread occurrence of this fallacy, psychologist Lloyd Morgan developed a conservative guideline for interpreting non-human behaviour as early as 1903. It later became 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" [80, p. 59]. 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. It has been argued that Morgan's Canon can and should be applied to interpreting machine learning results [79], and we consider it to be especially relevant in the context of shortcut learning. Accordingly, it might be worth acquiring the habit to confront machine learning models with a "Morgan's Canon for machine learning"[6]: *Never attribute to high-level abilities that which can be adequately explained by shortcut learning.*

Testing (surprisingly) strong baselines In order to find out whether a result may also be explained by shortcut learning, it can be helpful to test whether a baseline model exceeds expectations even though it does not use intended features. Examples include using nearest neighbours for scene completion and estimating geolocation [81, 82], object recognition with local features only [36], reasoning based on single cue words [59, 14] or answering questions about a movie without ever showing the movie to a model [83]. Importantly, this is not meant to imply that DNNs cannot acquire high-level abilities. They certainly do have the potential to solve complex challenges and serve as scientific models for prediction, explanation and exploration [84]—however, we must not confuse performance on a *dataset* with the acquisition of an *underlying ability*.

6.2   Detecting shortcuts: towards o.o.d. generalisation tests

Making o.o.d. generalisation tests a standard practice Currently, measuring model performance by assessing validation performance on an i.i.d. test set is at the very heart of the vast majority of machine learning benchmarks. Unfortunately, in real-world settings

---

[6]Our formulation is adapted from Hanlon's razor, "Never attribute to malice that which can be adequately explained by stupidity".

the i.i.d. assumption is rarely justified; in fact, this assumption has been called "the big lie in machine learning" [85]. While any metric is typically only 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. In Section 2.2 we 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 generalise to a more challenging test. Worse yet, as long as Bob continues to receive good grades through surface learning, he is unlikely to change his learning strategy.

Within the field of Education, what is the best practice to avoid surface learning? It has been argued that changing the type of examination from multiple-choice tests to essay questions discourages surface learning, and indeed surface approaches typically fail on these kinds of exams [21]. Essay questions, on the other hand, encourage so-called *deep* or *transformational* learning strategies [86, 87], like Alice's focus on understanding. This in turn enables transferring the learned content to *novel* problems and consequently achieves a much better overlap between the educational objectives of the teacher and what the students actually learn [88]. We can easily see the connection to machine learning—transferring knowledge to novel problems corresponds to testing generalisation beyond the narrowly learned setting [89, 90, 91]. If model performance is assessed only on i.i.d. test data, then we are unable to discover whether the model is actually acquiring the ability we think it is, since exploiting shortcuts often leads to deceptively good results on standard metrics [92]. We, among many others [93, 78, 94, 95, 96], have explored a variety of o.o.d. tests and we hope it will be possible to identify a sufficiently simple and effective test procedure that could replace i.i.d. testing as a new standard method for benchmarking machine learning models in the future.

Designing good o.o.d. tests While a distribution shift (between i.i.d. and o.o.d. data) has a clear mathematical definition, it can be hard to detect in practice [101, 102]. In these cases, training a classifier to distinguish samples in dataset A from samples in dataset A' can reveal a distribution shift. We believe that good o.o.d. tests should fullfill 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*. Typically, the space of all conceivable o.o.d. tests includes numerous uninteresting tests. Thus given limited time and resources, one might want to focus on challenging test cases. As models evolve, generalisation benchmarks will need to evolve as well, which is exemplified by the Winograd Schema Challenge [103]. Initially designed to overcome shortcut opportunities caused by the open-ended nature of the Turing test, this common-sense reasoning benchmark was scrutinised after modern language models started to perform suspiciously well—and it indeed contained more shortcut opportunities than originally envisioned [104], highlighting the need for revised tests. Fortunately, stronger generalisation tests are beginning to gain traction across deep learning. While o.o.d. tests will likely need to evolve alongside the models they aim to evaluate, a few current encouraging examples are listed in Box I. In summary, rigorous generalisation benchmarks are crucial when distinguishing between the intended and a shortcut solution, and it would be extremely useful if a strong generally applicable testing procedure will emerge from this range of approaches.

> **Box I. 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 it 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 [33, 97].
>
> **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 [14].
>
> **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.
>
> **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 [52].
>
> **ObjectNet** introduces the idea of scientific controls into o.o.d. benchmarking, allowing to disentangle the influence of background, rotation and viewpoint [98].
>
> **PACS** and other domain generalisation datasets require extrapolation beyond i.i.d. data per design by testing on a domain different from training data (e.g. cartoon images) [99].
>
> **Shift-MNIST / biased CelebA / unfair dSprites** are controlled toy datasets that introduce correlations in the training data (e.g. 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, 40, 100, 41].

6.3   Shortcuts: why are they learned?

The "Principle of Least Effort"   Why are machines so prone to learning shortcuts, detecting grass instead of cows [9] or a metal token instead of pneumonia [15]? Exploiting those shortcuts seems much *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" [119], the observation that language speakers generally try to minimise the amount of effort involved in 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 [120, 121]. Interestingly, whether a language change makes it easier for the speaker doesn't 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).

Box II. SHORTCUT LEARNING & 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) [105], their inductive bias plays an important role for the types of patterns they prefer to learn given finite capacity and data.

- **Structure: architecture.** Convolutions make it harder for a model to use location—a prior [106] that is so powerful for natural images that even untrained networks can be used for tasks like image inpainting and denoising [107]. In Natural Language Processing, transformer architectures [108] 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 [109].

- **Experience: training data.** As discussed in Section 4.1, shortcut opportunities are present in most data and rarely disappear by adding more data [32, 69, 56, 38, 33]. Modifying the training data to block specific shortcuts has been demonstrated to work for reducing adversarial vulnerability [110] 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]. Regularisation terms that use additional information about the training data have been used to disentangle intended features from shortcut features [39, 111].

- **Learning: optimisation.** Stochastic gradient descent and its variants bias DNNs towards learning simple functions [112, 113, 114, 115]. 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 memorisation [116, 117]. The complex interactions between training method and architecture are poorly understood so far; strong claims can only be made for simple cases [118].

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 optimisation. Often, the training process starts with feeding training data to the model with a fixed architecture and randomly initialised parameters. When the model's prediction is compared to ground truth, the loss function measures the prediction's quality. This supervision signal is used by an optimiser for adapting the model's internal parameters such that the model makes a better prediction the next time. Taken together, these four components (which determine the *inductive bias* of a model) influence how certain solutions are much easier to learn than others, and thus ultimately determine whether a shortcut is learned instead of the intended solution [122]. Box II provides an overview of the connections between shortcut learning and inductive biases.

7 Beyond shortcut learning

A lack of out-of-distribution generalisation can be observed all across machine learning. Consequently, a significant fraction of machine learning research is concerned with overcoming shortcut learning, albeit not necessarily as a concerted effort. Here we highlight connections between different research areas. Note that an exhaustive list would be out of the scope for this work. Instead, we cover a diverse set of approaches we find promising, each providing a unique perspective on learning beyond shortcut learning.

Domain-specific prior knowledge Avoiding reliance on unintended cues can be achieved by designing architectures and data-augmentation strategies that discourage learning shortcut features. If the orientation of an object does not matter for its category, either data-augmentation or hard-coded rotation invariance [123] can be applied. This strategy can be applied to almost any well-understood transformation of the inputs and finds its probably most general form in auto-augment as an augmentation strategy [124]. Extreme data-augmentation strategies are also the core ingredient of the most successful semi-supervised [125] and self-supervised learning approaches to date [126, 127].

Adversarial examples and robustness Adversarial attacks are a powerful analysis tool for worst-case generalisation [8]. Adversarial examples can be understood as counterfactual explanations, since they are the smallest change to an input that produces a certain output. Achieving counterfactual explanations of predictions aligned with human intention makes the ultimate goals of adversarial robustness tightly coupled to causality research in machine learning [128]. Adversarially robust models are somewhat more aligned with humans and show promising generalisation abilities [129, 130]. While adversarial attacks test model performance on model-dependent worst-case noise, a related line of research focuses on model-independent noise like image corruptions [51, 52].

Domain adaptation, -generalisation and -randomisation These areas are explicitly concerned with out-of-distribution generalisation. Usually, multiple distributions are observed during training time and the model is supposed to generalise to a new distribution at test time. Under certain assumptions the intended (or even causal) solution can be learned from multiple domains and environments [131, 39, 111]. In robotics, domain randomisation (setting certain simulation parameters randomly during training) is a very successful approach for learning policies that generalise to similar situations in the real-world [70].

Fairness Fairness research aims at making machine decisions "fair" according to a certain definition [132]. Individual fairness aims at treating similar individuals similarly while group fairness aims at treating subgroups no different than the rest of the population [133, 134]. Fairness is closely linked to generalisation and causality [135]. Sensitive group membership can be viewed as a domain indicator: Just like machine decisions should not typically be influenced by changing the domain of the data, they also should not be biased against minority groups.

Meta-learning Meta-learning seeks to learn how to learn. An intermediate goal is to learn representations that can adapt quickly to new conditions [136, 137, 138]. This ability is connected to the identification of causal graphs [139] since learning causal features allows for small changes when changing environments.

Generative modelling and disentanglement Learning to generate the observed data forces a neural network to model every variation in the training data. By itself, however, this does not necessarily lead to representations useful for downstream tasks [140], let alone out-of-distribution generalisation. Research on disentanglement addresses this shortcoming by learning generative models with well-structured latent representations [141]. The goal is to recover the true generating factors of the data distribution from observations [142] by identifying independent causal mechanisms [128].

8 Conclusion

> *"The road reaches every place, the short cut only one"*
> — James Richardson [143]

Science aims for understanding. While deep learning as an engineering discipline has seen tremendous progress over the last few years, deep learning as a scientific discipline is still lagging behind in terms of understanding the principles and limitations that govern how machines learn to extract patterns from data. A deeper understanding of how to overcome shortcut learning is of relevance beyond the current application domains of machine learning and there might be interesting future opportunities for cross-fertilisation with other disciplines such as Economics (designing management incentives that do not jeopardise long-term success by rewarding unintended "shortcut" behaviour) or Law (creating laws without "loophole" shortcut opportunities). Until the problem is solved, however, we offer the following four 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 generalisation failures. "Connecting the dots" between affected areas is likely to facilitate progress, and making progress can generate highly valuable impact across various applications domains.

(2) Interpreting results carefully
Discovering a shortcut often reveals the existence of an easy solution to a seemingly complex dataset. We argue that 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. generalisation
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. generalisation 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 (optimisation), as well as a thorough understanding of the interactions between these factors.

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 generalisation that is able to replace the current i.i.d. tests.

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.; 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.

References

[1] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034 (2015).

[2] Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484 (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. *arXiv:1711.05225* (2017).

[5] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* (2018).

[6] Rolnick, D. *et al.* Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433* (2019).

[7] Reichstein, M. *et al.* Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195 (2019).

[8] Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv:1312.6199* (2013).

[9] Beery, S., Van Horn, G. & Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, 456–473 (2018).

[10] Rosenfeld, A., Zemel, R. & Tsotsos, J. K. The elephant in the room. *arXiv preprint arXiv:1808.03305* (2018).

[11] Heuer, H., Monz, C. & Smeulders, A. W. Generating captions without looking beyond objects. *arXiv preprint arXiv:1610.03708* (2016).

[12] Buolamwini, J. & Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91 (2018).

[13] Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. `https://reut.rs/2Od9fPr` (2018).

[14] Niven, T. & Kao, H.-Y. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355* (2019).

[15] Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine* **15**, e1002683 (2018).

> **This study highlights the importance of testing model generalisation in the medical context.**

[16] Bickel, S., Brückner, M. & Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research* **10**, 2137–2155 (2009).

[17] Schölkopf, B. *et al.* On causal and anticausal learning. In *International Conference on Machine Learning*, 1255–1262 ([Sl: sn], 2012).

[18] Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2011).

> **This study provides a comprehensive overview of dataset biases in computer vision.**

[19] Branwen, G. The neural net tank urban legend. `https://www.gwern.net/Tanks` (2011).

[20] Pfungst, O. *Clever Hans: (the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology* (Holt, Rinehart and Winston, 1911).

[21] Scouller, K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education* **35**, 453–472 (1998).

[22] Shane, J. Do neural nets dream of electric sheep? (2018). URL https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep.

[23] Wichmann, F. A., Drewes, J., Rosas, P. & Gegenfurtner, K. R. Animal detection in natural scenes: Critical features revisited. *Journal of Vision* **10**, 6–6 (2010).

[24] Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (ACM, 2016).

[25] Zhu, Z., Xie, L. & Yuille, A. L. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596* (2016).

[26] Wang, J. *et al.* Visual concepts and compositional voting. *arXiv preprint arXiv:1711.04451* (2017).

[27] Dawson, M., Zisserman, A. & Nellåker, C. From same photo: Cheating on visual kinship challenges. In *Asian Conference on Computer Vision*, 654–668 (Springer, 2018).

[28] Biederman, I. *On the semantics of a glance at a scene* (Hillsdale, NJ: Erlbaum, 1981).

[29] Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology* **14**, 143–177 (1982).

[30] Oliva, A. & Torralba, A. The role of context in object recognition. *Trends in Cognitive Sciences* **11**, 520–527 (2007).

[31] Castelhano, M. S. & Heaven, C. Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review* **18**, 890–896 (2011).

[32] Jo, J. & Bengio, Y. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561* (2017).

[33] Ilyas, A. *et al.* Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175* (2019).

> **This study shows how learning imperceptible predictive features leads to adversarial vulnerability.**

[34] Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *Intelligent Systems* (2009).

[35] Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67–82 (1997).

[36] Brendel, W. & Bethge, M. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations* (2019).

[37] Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology* **14**, e1006613 (2018).

[38] Geirhos, R. *et al.* ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations* (2019).

> **This article shows how shortcut feature combination strategies are linked to distortion robustness.**

[39] Heinze-Deml, C. & Meinshausen, N. Conditional variance penalties and domain shift robustness. *arXiv: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 *International Conference on Learning Representations* (2019).

[41] Jacobsen, J.-H., Behrmann, J., Zemel, R. & Bethge, M. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations* (2019).

[42] Kamin, L. J. Predictability, surprise, attention, and conditioning. *Punishment and aversive behavior* 279–96 (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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436 (IEEE, 2015).

[46] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. Natural adversarial examples. arXiv preprint arXiv:1907.07174 (2019).

[47] Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv:1805.12177* (2018).

[48] Wang, M. & Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153 (2018).

[49] Alcorn, M. A. *et al.* Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019).

[50] Dodge, S. & Karam, L. Human and DNN classification performance on images with quality distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)* **16**, 7 (2019).

[51] Geirhos, R. *et al.* Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (2018).

[52] Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations* (2019).

[53] Michaelis, C. *et al.* Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484* (2019).

[54] Minderer, M., Bachem, O., Houlsby, N. & Tschannen, M. Automatic shortcut removal for self-supervised representation learning. *arXiv preprint arXiv:2002.08822* (2020).

[55] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. & Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6904–6913 (2017).

[56] Gururangan, S. *et al.* Annotation artifacts in Natural Language Inference data. *arXiv preprint arXiv:1803.02324* (2018).

> **This article highlights how Natural Language Inference models learn heuristics that exploit superficial cues.**

[57] Kaushik, D. & Lipton, Z. C. How much reading does reading comprehension require? A critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926* (2018).

[58] Geva, M., Goldberg, Y. & Berant, J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898* (2019).

[59] Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. & Van Durme, B. Hypothesis only baselines in Natural Language Inference. *arXiv preprint arXiv:1805.01042* (2018).

[60] Kavumba, P. *et al.* When choosing plausible alternatives, Clever Hans can be clever. *arXiv preprint arXiv:1911.00225* (2019).

[61] McCoy, R. T., Pavlick, E. & Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in Natural Language Inference. *arXiv preprint arXiv:1902.01007* (2019).

[62] Agrawal, A., Batra, D. & Parikh, D. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356* (2016).

[63] Belinkov, Y. & Bisk, Y. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* (2017).

[64] Jia, R. & Liang, P. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[65] Glockner, M., Shwartz, V. & Goldberg, Y. Breaking NLI systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266* (2018).

[66] Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* **1** (2019).

[67] Murphy VII, T. The first level of Super Mario Bros. is easy with lexicographic orderings and time travel. *The Association for Computational Heresy (SIGBOVIK) 2013* 112 (2013).

[68] Amodei, D. *et al.* Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[69] Lehman, J. *et al.* The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453* (2018).

> **This paper provides a comprehensive collection of anecdotes about shortcut learning / reward hacking in Reinforcement Learning and beyond.**

[70] Tobin, J. *et al.* Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 23–30 (IEEE, 2017).

[71] Akkaya, I. *et al.* Solving Rubik's Cube with a robot hand. *arXiv:1910.07113* (2019).

[72] Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

> **This study shows how algorithms amplify social biases to boost performance.**

[73] Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence* **1**, 174 (2019).

[74] Hashimoto, T. B., Srivastava, M., Namkoong, H. & Liang, P. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010* (2018).

[75] Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).

[76] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830* (2019).

[77] Marr, D. *Vision: A computational investigation into the human representation and processing of visual information* (W.H. Freeman and Company, San Francisco, 1982).

[78] Borowski, J. *et al.* The notorious difficulty of comparing human and machine perception. In *NeurIPS Shared Visual Representations in Human and Machine Intelligence Workshop* (2019).

> **The case studies presented in this article highlight the difficulty of interpreting machine behaviour in the presence of shortcut learning.**

[79] Buckner, C. The Comparative Psychology of Artificial Intelligences (2019). URL http://philsci-archive.pitt.edu/16034/.

> **This opinionated article points out important caveats when comparing human to machine intelligence.**

[80] Morgan, C. L. Introduction to Comparative Psychology. (rev. ed.). *New York: Scribner* (1903).

[81] Hays, J. & Efros, A. A. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)* **26**, 4 (2007).

[82] Hays, J. & Efros, A. A. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8 (IEEE, 2008).

[83] Jasani, B., Girdhar, R. & Ramanan, D. Are we asking the right questions in MovieQA? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0 (2019).

[84] Cichy, R. M. & Kaiser, D. Deep neural networks as scientific models. *Trends in Cognitive Sciences* (2019).

[85] Ghahramani, Z. Panel of workshop on advances in Approximate Bayesian Inference (AABI) 2017 (2017). URL https://www.youtube.com/watch?v=x1UByHT60mQ&feature=youtu.be&t=37m44s.

[86] Marton, F. & Säaljö, R. On qualitative differences in learning—II Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology* **46**, 115–127 (1976).

[87] Biggs, J. Individual differences in study processes and the quality of learning outcomes. *Higher Education* **8**, 381–394 (1979).

[88] Chin, C. & Brown, D. E. Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching* **37**, 109–138 (2000).

> **This article from the field of Education reflects upon ways to achieve a better overlap between educational objectives and the way students learn.**

[89] Marcus, G. F. Rethinking eliminative connectionism. *Cognitive Psychology* **37**, 243–282 (1998).

[90] Kilbertus, N., Parascandolo, G. & Schölkopf, B. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524* (2018).

[91] Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

[92] Lapuschkin, S. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* **10**, 1096 (2019).

> **This study highlights how shortcut learning can lead to deceptively good results on standard metrics.**

[93] Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. arXiv preprint arXiv:1604.00289 (2016).

[94] Chollet, F. The measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019).

[95] Crosby, M., Beyret, B. & Halina, M. The Animal-AI Olympics. *Nature Machine Intelligence* **1**, 257–257 (2019).

[96] Juliani, A. *et al.* Obstacle tower: A generalization challenge in vision, control, and planning. *arXiv preprint arXiv:1902.01378* (2019).

[97] Engstrom, L. *et al.* A discussion of 'adversarial examples are not bugs, they are features'. *Distill* (2019).

[98] Barbu, A. *et al.* ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 9448–9458 (2019).

[99] Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 5542–5550 (2017).

[100] Creager, E. *et al.* Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589* (2019).

[101] Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451* (2018).

[102] Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do ImageNet classifiers generalize to ImageNet? *arXiv preprint arXiv:1902.10811* (2019).

[103] Levesque, H., Davis, E. & Morgenstern, L. The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning* (2012).

[104] 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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3373–3378 (2019).

[105] Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366 (1989).

[106] d'Ascoli, S., Sagun, L., Bruna, J. & Biroli, G. Finding the needle in the haystack with convolutions: On the benefits of architectural bias. *arXiv preprint arXiv:1906.06766* (2019).

[107] Ulyanov, D., Vedaldi, A. & Lempitsky, V. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454 (2018).

[108] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008 (2017).

[109] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 41–50 (2019).

[110] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations* (2018).

[111] Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[112] Wu, L., Zhu, Z. & E, W. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239* (2017).

[113] De Palma, G., Kiani, B. T. & Lloyd, S. Deep neural networks are biased towards simple functions. *arXiv preprint arXiv:1812.10156* (2018).

[114] Valle-Perez, G., Camargo, C. Q. & Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations* (2019).

[115] Sun, K. & Nielsen, F. Lightlike neuromanifolds, Occam's Razor and deep learning. *arXiv preprint arXiv:1905.11027* (2019).

[116] Arpit, D. *et al.* A closer look at memorization in deep networks. In *International Conference on Machine Learning* (2017).

[117] Li, Y., Wei, C. & Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595* (2019).

[118] Bartlett, P. L., Long, P. M., Lugosi, G. & Tsigler, A. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300* (2019).

[119] Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley press, 1949).

[120] Ohala, J. J. The phonetics and phonology of aspects of assimilation. *Papers in Laboratory Phonology* **1**, 258–275 (1990).

[121] Vicentini, A. The economy principle in language. *Notes and Observations from early modern English grammars. Mots, Palabras, Words* **3**, 37–57 (2003).

[122] Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a less artificial intelligence. *Neuron* **103**, 967–979 (2019).

[123] Cohen, T. & Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2990–2999 (2016).

[124] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 113–123 (2019).

[125] Berthelot, D. *et al.* Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249* (2019).

[126] Hjelm, R. D. *et al.* Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).

[127] Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[128] Schölkopf, B. Causality for machine learning. *arXiv preprint arXiv:1911.10500* (2019).

[129] Schott, L., Rauber, J., Bethge, M. & Brendel, W. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations* (2019).

[130] Engstrom, L. *et al.* Learning perceptually-aligned representations via adversarial robustness. *arXiv:1906.00945* (2019).

[131] Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 947–1012 (2016).

[132] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226 (2012).

[133] Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, 325–333 (2013).

[134] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323 (2016).

[135] Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076 (2017).

[136] Schmidhuber, J. Evolutionary principles in self-referential learning. *On learning how to learn: The meta-meta-... hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich* **1**, 2 (1987).

[137] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, 1842–1850 (2016).

[138] Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning* (2017).

[139] Bengio, Y. *et al.* A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912* (2019).

[140] Fetaya, E., Jacobsen, J.-H., Grathwohl, W. & Zemel, R. Understanding the limitations of conditional generative models. In *International Conference on Learning Representations* (2020).

[141] Higgins, I. *et al.* Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations* (2017).

[142] Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430 (2000).

[143] Richardson, J. *Vectors: aphorisms & ten-second essays* (Ausable Press, 2001).

Appendix

A   Toy example: method details

The code to reproduce our toy example (Figure 2) is available from `https://github.com/rgeirhos/shortcut-perspective`. Two easily distinguishable shapes (star and moon) were placed on a 200×200 dimensional 2D canvas. The training set is constructed out of 4000 images, where 2000 contain a star shape and 2000 a moon shape. The star shape is randomly placed in the top right and bottom left quarters of the canvas, whereas the moon shape is randomly placed in the top left and bottom right quarters of the canvas. At test time the setup is nearly identical, 1000 images with a star and 1000 images with a moon are presented. However, this time the position of star and moon shapes are randomised over the full canvas, i.e. in test images stars and moons can appear at any location.

We train two classifiers on this dataset: a fully connected network as well as a convolutional network. The classifiers are trained for five epochs with a batch size of 100 on the training set and evaluated on the test set. The training objective is standard crossentropy loss and the optimizer is Adam with a learning rate of 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1e - 08$. The fully connected network was a three-layer ReLU MLP (multilayer perceptron) with 1024 units in each layer and two output units corresponding to the two target classes. It reaches 100% accuracy at training time and approximately chance-level accuracy at test time (51.0%). The convolutional network had three convolutional layers with 128 channels, a stride of 2 and filter size of 5×5 interleaved with ReLU nonlinearities, followed by a global average pooling and a linear layer mapping the 128 outputs to the logits. It reaches 100% accuracy on train and test set.

B   Image rights & attribution

Figure 1 consists of four images from different sources. The first image from the left was taken from `https://aiweirdness.com/post/171451900302/do-neural-nets-dream-of-electric-sheep` with permission of the author. The second image from the left was generated by ourselves. The third image from the left is from ref. [15]. It was released under the CC BY 4.0 license as stated here: `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683` and adapted by us from Figure 2B of the corresponding publication. The image on the right is Figure 1 from ref. [64]. It was released under CC BY 4.0 license as stated here: `https://www.aclweb.org/anthology/D17-1215/`(at the bottom) and retrieved by us from .

The image from Section 4.1 was adapted from Figure 1 of ref. [9] with permission from the authors (image cropped from original figure by us). The image from Section 4.2 was adapted from Figure 1 of ref. [38] with permission from the authors (image cropped from original figure by us). The image from Section 4.3 was adapted from Figure 1 of ref. [45] with permission from the authors (image cropped from original figure by us).

Figure 4 consists of a number of images from different sources. The first author of the corresponding publication is mentioned in the figure for identification. The images from ref. [8] were released under the CC BY 3.0 license as stated here: `https://arxiv.org/abs/1312.6199` and adapted by us from Figure 5a of the corresponding publication (images cropped from original figure by us). The images from ref. [50] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [49] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [38] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [41] were adapted from Figure 1 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [36] were adapted from Figure 5 of the corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [9] were adapted from Figure 1 of the

corresponding paper with permission from the authors (images cropped from original figure by us). The images from ref. [45] were adapted from Figure 1 and Figure 2 of the corresponding paper with permission from the authors (images cropped from original figures by us).