When will AI misclassify? Intuiting failures on natural images

Makaela Nartker[*]
Department of Psychological & Brain Sciences,
Johns Hopkins University, Baltimore, MD, USA

Zhenglong Zhou[*]
Department of Psychology, University of Pennsylvania,
Philadelphia, PA, USA

Chaz Firestone
Department of Psychological & Brain Sciences,
Johns Hopkins University, Baltimore, MD, USA

Machine recognition systems now rival humans in their ability to classify natural images. However, their success is accompanied by a striking failure: a tendency to commit bizarre misclassifications on inputs specifically selected to fool them. What do ordinary people know about the nature and prevalence of such classification errors? Here, five experiments exploit the recent discovery of "natural adversarial examples" to ask whether naive observers can predict when and how machines will misclassify natural images. Whereas classical adversarial examples are inputs that have been minimally perturbed to induce misclassifications, natural adversarial examples are simply unmodified natural photographs that consistently fool a wide variety of machine recognition systems. For example, a bird casting a shadow might be misclassified as a SUNDIAL, or a beach umbrella made of straw might be misclassified as a BROOM. In Experiment 1, subjects accurately predicted which natural images machines would misclassify and which they would not. Experiments 2 through 4 extended this ability to *how* the images would be misclassified, showing that anticipating machine misclassifications goes beyond merely identifying an image as nonprototypical. Finally, Experiment 5 replicated these findings under more ecologically valid conditions, demonstrating that subjects can anticipate misclassifications not only under two-alternative forced-choice conditions (as in Experiments 1–4), but also when the images appear one at a time in a continuous stream—a skill that may be of value to human–machine teams. We suggest that ordinary people can intuit how easy or hard a natural image is to classify, and we discuss the implications of these results for practical and theoretical issues at the interface of biological and artificial vision.

Introduction

Look at the images in Figure 1A. Multiple machine vision systems that match or surpass human performance on visual classification benchmarks were supposed to classify both of the images as BUBBLES; however, they misclassified one of the images as a completely different object, and even did so with high confidence. Can you guess which of the two images was misclassified? (And does it help to know that the misclassified image was called a SALT SHAKER instead?)

Sophisticated machine recognition systems—especially deep neural networks (DNNs)—now rival human accuracy on a wide array of visual tasks, matching or surpassing established performance benchmarks for classifying objects, faces, text, scenes, medical scans, traffic signs, and other inputs (LeCun, Bengio, & Hinton, 2015) (for critical discussion, see Raji, Bender, Paullada, Denton, & Hanna, 2021). However, they are also prone to bizarre errors that threaten their promise. An especially striking class of such errors arises from adversarial examples—carefully chosen inputs that an attacker has selected to cause odd and alarming misclassifications (Szegedy et al., 2013). For example, adversarial attacks can cause autonomous vehicles to misread traffic signs (e.g., misperceiving STOP as SPEED LIMIT 45; Eykholt et al., 2018), diagnostic systems to misclassify radiological images (e.g., misdiagnosing BENIGN melanocytic nevi as MALIGNANT; Finlayson et al., 2019), or even smartphones to misinterpret audio signals (e.g., hearing seemingly meaningless static as the command OK GOOGLE, CALL 911; Carlini & Wagner, 2018; Carlini et al., 2016).

Misclassifications of this sort raise deep questions about exactly what DNNs are learning about the image classes they have been shown, with consequences not only in computer science and engineering, but

Citation: Nartker, M., Zhou, Z., & Firestone, C. (2023). When will AI misclassify? Intuiting failures on natural images. *Journal of Vision*, *23*(4):4, 1–15, https://doi.org/10.1167/jov.23.4.4.

https://doi.org/10.1167/jov.23.4.4 Received February 28, 2021; published April 6, 2023 ISSN 1534-7362 Copyright 2023 The Authors

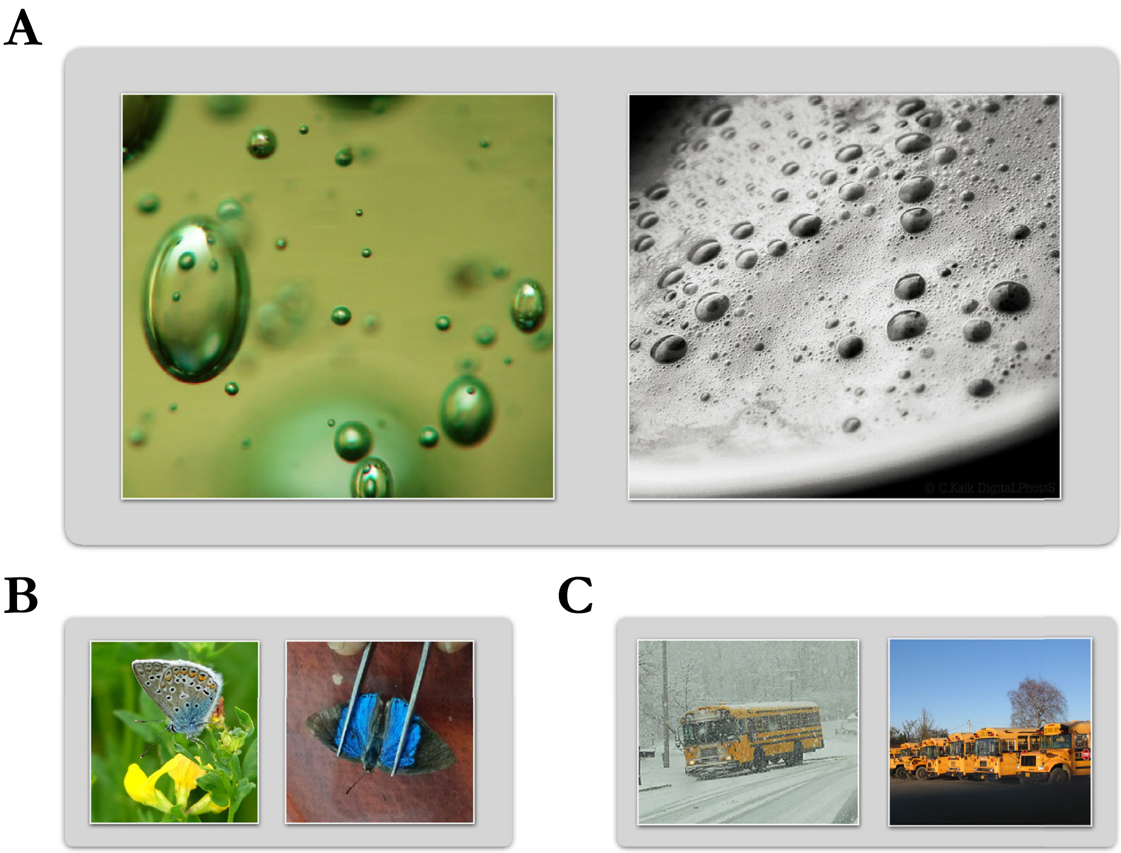

Figure 1. Natural adversarial examples are ordinary, unmodified, natural images that elicit high-confidence misclassifications from machine vision systems. For each pair of images (BUBBLE, BUTTERFLY, SCHOOL BUS), one of the images was misclassified by multiple ResNet50 models as a completely different object (with near-ceiling confidence), whereas the other image was classified correctly. (**A**) The right image was incorrectly classified as a SALT SHAKER. (**B**) The right image was incorrectly classified as a BROOM. (**C**) The left image was incorrectly classified as a SNOWPLOW. The experiments reported here (and illustrated above) ask whether human subjects can anticipate which images were classified incorrectly, and whether they can appreciate the overlap between the images and the classes they were misclassified as.

also neuroscience, psychology, and even philosophy (Buckner, 2020; Firestone, 2020; Serre, 2019; Yamins & DiCarlo, 2016; Yuille & Liu, 2021). However, beyond these theoretical issues, such extreme classification errors also have practical consequences, including for users of technologies that rely on these new advances: If autonomous vehicles are prone to misclassifying overturned schoolbuses or graffitied traffic signs, they may behave dangerously and unpredictably; if automated radiological systems misread medical images with slight perturbations, they may misdiagnose patients with critical illnesses; if meaningless noise is parsed by home assistants as valid voice commands, such systems could surrender financial records and other sensitive data without permission from the user; and so on for many other industries touched by automated classification. For these reasons, it would be valuable to know whether and to what extent humans can anticipate or predict such misclassifications, as a growing literature now investigates (whether directly or indirectly; Dujmović, Malhotra, & Bowers, 2020; Elsayed et al., 2018; Harding, Rajivan, Bertenthal, & Gonzalez, 2018; Lepori & Firestone, 2022; Zhou & Firestone, 2019).

At the same time, however, certain features of adversarial examples have made it unclear how worried one should be about their real-world impact. Most notably, because classical adversarial images (including all of those studied in the above-referenced work) are *manipulated* to cause misclassifications (e.g., by perturbing an existing natural image [Szegedy et al., 2013], or by creating wholly synthetic inputs derived from noise [Nguyen, Yosinski, & Clune, 2015]), most real-world machine vision applications are unlikely to actually encounter these stimuli "in the wild"—except in the very rare and specific case of being targeted by a malicious actor. Although such targeting is certainly *possible* (as demonstrated by Kurakin, Goodfellow, & Bengio, 2016; Morgulis, Kreines, Mendelowitz, & Weisglass, 2019), it is noteworthy that, at least to our knowledge, a targeted adversarial attack has not (yet) been responsible for any major mishap in a real-life high-stakes setting. It is thus genuinely unclear whether and how ordinary users of machine recognition technologies should be concerned about the misclassifications arising from such specific and targeted attacks. Might there be other sources of machine misclassification that pose an equal (or even greater) threat?

Natural adversarial examples: A problem and an opportunity

Recently, it was demonstrated that, beyond *generating* stimuli to fool machines (as in classical adversarial examples), it is also possible to adversarially *search* large spaces of natural images to find ordinary, unmodified, real-world images that just so happen to elicit completely false and high-confidence misclassifications from machine recognition systems (Hendrycks et al., 2019). This search process proceeds by retrieving millions of user-labeled images from online databases, feeding each one into multiple image-recognizing DNNs, and then simply removing those images that fail to fool the models. The resulting images have been called "natural adversarial examples" (Hendrycks et al., 2019), and they include images such as a bird casting a shadow (misclassified as a SUNDIAL) or a beach umbrella made of straw (misclassified as a BROOM). Indeed, all of the misclassified images in Figure 1 are natural adversarial examples: normal, unmodified, real-world images that nevertheless robustly fool machines.

Natural adversarial examples are in many ways quite different than more traditional adversarial examples; indeed, because they are just ordinary unmodified images, it is not clear that these two kinds of inputs form a coherent or unified class in the first place.[1] However, regardless of how best to taxonomize them, these inputs amplify many of the threats originally raised by adversarial misclassification.

First, because they are just ordinary real-world images, they may be more likely to appear as inputs to (and cause trouble for) actual machine vision applications in realistic settings. (Indeed, several car crashes involving autonomous vehicles seem to have arisen from misclassifications not unlike those discussed here, such as a vehicle that fails to see a truck lying on its side as an obstacle and so plows straight into it.) Second, because natural adversarial examples are drawn from the very same distribution as the training sets of many or most DNNs (*without* any intervention from an attacker perturbing them), there is every reason to think they should be classified accurately, making the failures they elicit all the more surprising. Third, and finally, natural adversarial examples have proven to stump multiple machine-recognition models beyond the ResNet-50 models that were initially used to find them, including AlexNet, VGG-19, and DenseNet-121 (Hendrycks et al., 2019). In other words, rather than exploiting the vagaries and vulnerabilities of a single model, they are genuinely and generally difficult for many types of models to classify (such that they exhibit adversarial transfer; Tramèr, Papernot, Goodfellow, Boneh, & McDaniel, 2017).

For these same reasons, however, natural adversarial examples also present a research opportunity: As images that systematically fool machines, they may serve as a testbed for human intuition about visual processing. If subjects can predict which images are likely to elicit these machine failures, this may suggest 1) that at least some of the errors that machine classification systems make are intuitive to humans too (in ways that may generalize to visual processing writ large), and 2) that human users of automated classification technologies may be in a position to anticipate the (mis)behavior of such systems. Yet, although a growing body of work examines human perception of classical adversarial examples (Dujmović et al., 2020; Elsayed et al., 2018; Harding et al., 2018; Lepori & Firestone, 2022; Zhou & Firestone, 2019), natural adversarial examples have not received this same empirical attention (although see Chandrasekaran, Yadav, Chattopadhyay, Prabhu, & Parikh, 2017; Bos, Glasgow, Gersh, Harbison, & Lyn Paul, 2019).

The present experiments: When will AI misclassify natural images?

Here, we fill this gap by examining human intuition about machine misclassification of natural adversarial examples. If ordinary people can anticipate such misclassifications (to at least some degree), this result could suggest that human users of machine vision technologies (such as vehicles with auto-pilot, or computer-aided detection systems in radiology) may be in a position to intuit and anticipate their failures (and perhaps know when to intervene—e.g., by taking control of the wheel, or discounting computer-aided detection system marks that are likely false alarms). To this end, we showed human subjects samples of these images, as well as ordinary ImageNet images from the same categories that machines typically classify correctly; we then asked subjects to predict which image(s) a machine vision system would misclassify. Later experiments add variations, such as exploring what strategies are helpful to subjects in making these decisions.

Experiment 1: When will AI misclassify?

Our first experiment asked whether ordinary human subjects can predict which natural images will be incorrectly classified by machines, relative to chance performance. We showed subjects natural adversarial examples (which are classified incorrectly), as well as ordinary ImageNet images (Figure 2A). Could naive observers tell which were which?

Open science practices

All of the studies reported here adhere to principles of open science, including pre-registration and public availability of data and materials. An archive of our stimuli, experiment code, data, analyses, and pre-registrations is available at https://osf.io/y72qe/.

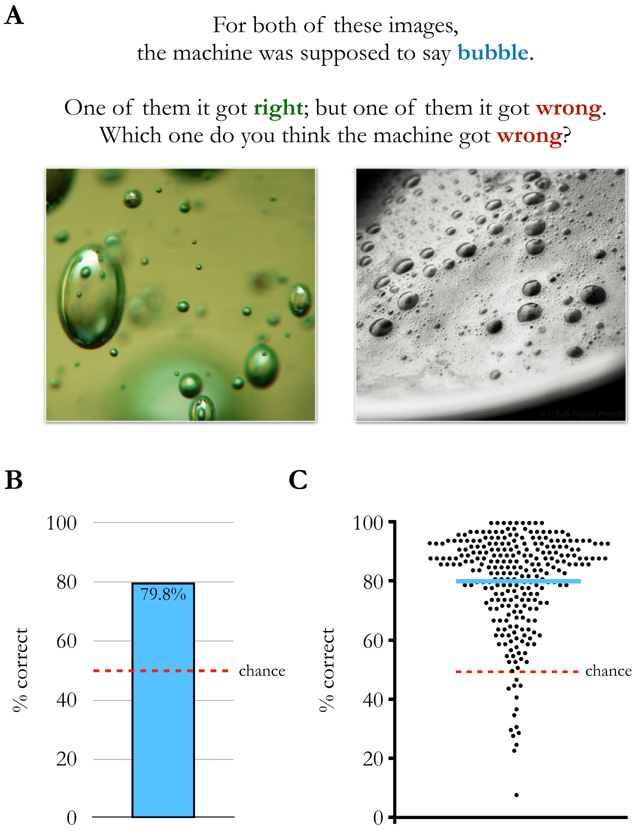

Figure 2. Method and results of Experiment 1. (**A**) On each trial, subjects saw two images, and were asked to guess which of the images was misclassified. (**B**) Subjects correctly identified the misclassified image on 79.8% of trials. (**C**) A beeswarm plot (where each dot represents one image) reveals that subjects' success was not driven only by a small subset of images but rather was fairly widespread throughout the imageset.

Methods

Subjects

Experiment 1 recruited 200 subjects from Amazon Mechanical-Turk. (For a discussion of this subject pool's reliability, see Crump, McDonnell, & Gureckis, 2013). We initially chose this sample to match previous work on human perception of adversarial examples (Zhou & Firestone, 2019). This sample size was pre-registered. All subjects gave informed consent and were monetarily compensated for their participation.

Stimuli

ImageNet-A (the library containing natural adversarial examples; Hendrycks et al., 2019) contains 7,500 images across 200 ImageNet categories. However, not all of these images are appropriate for the present experiments: Whereas many of these categories are familiar and commonplace (e.g., BUBBLE, SCHOOL BUS), others are quite obscure and not familiar to most people (e.g., YELLOW LADY'S SLIPPER [a particular species of orchid], or JUNCO [a small bird related to sparrows]); these latter categories are unsuitable for evaluating human prediction of machine misclassification, because an ordinary human subject cannot reasonably be expected to tell whether a machine misclassified a JUNCO without knowing what a JUNCO is in the first place.

Thus, to ensure that we used images from familiar categories, we selected the 50 image classes that returned the greatest number of Google hits at the time of our selecting them (August 2019). This approach indeed produced familiar and intuitive categories, including not only BUBBLE and SCHOOL BUS, but also ACOUSTIC GUITAR, BEE, HOT DOG, SANDAL, WINE BOTTLE, and so on. Still, each of these categories contains approximately 50 to 100 natural adversarial examples. So, to create a stimulus set for our experiment, we selected the 5 images from each of these 50 categories (250 images total) that produced the highest confidence misclassifications as reported in the ImageNet-A dataset; these 250 images served as the "misclassified" stimuli. Finally, we drew 5 random ImageNet images from each of the same 50 classes; these 250 images served as the control stimuli. Importantly, this procedure ensured that all images were chosen by an objective process rather than by examining the images ourselves, which protected against experimenter bias in image selection (for discussion, see Funke et al., 2021).

Procedure

In the experiment itself, subjects read the following prompt:

> *We have a machine that can look at a picture and tell us what it is. Most of the time, it tells us the right answers. But sometimes it makes mistakes. We don't know why the machine makes the mistakes it makes; do you?*
>
> *Your job in this experiment is to predict when our machine will make a mistake. On each trial, you'll see two images that will look to you like a familiar object. But, for one of them, the machine gave a different answer than it was supposed to. For each pair of images, we want you to predict which image the machine got wrong.*

Each trial (out of 50 total; 1 per image class) included one misclassified image (from ImageNet-A) and one control image (from ImageNet), each chosen randomly from the five images per class (Figure 2A). Subjects were given the following information about the images (here using the images in Figure 1A as an example):

> *For both of these images, the machine was supposed to say* **bubble***. One of them it got* **right***; but one of them it got* **wrong***. Which image do you think the machine got* **wrong***?*

A response was "correct" if subjects selected the ImageNet-A image, and "incorrect" if they selected the ImageNet image. Subjects were not given feedback about the correctness of their responses, and had unlimited time to respond.

To ensure that subjects were engaged and understood the instructions, we also included three pairs of "catch" images (appearing randomly within the experiment): a baseball, American football, and soccer ball, along with highly distorted versions of those images. We expected subjects to select the distorted images as being more likely to elicit misclassifications, and we excluded any subjects who did not do so (as well as any subjects who didn't provide a complete dataset). Readers can experience the task for themselves at https://perceptionresearch.org/naturalAdversarial/e1.

Results and discussion

We found strong evidence that humans can anticipate machine misclassifications in this task. Subjects ($N = 165$ after exclusions) correctly identified which images were misclassified by machines on 79.8% of trials, well above chance-level accuracy of 50%, $t(164) = 37.15$, $p < .001$ (Figure 2B). Moreover, subjects' success was not driven only by a small subset of images but rather was fairly widespread throughout the imageset (Figure 2C). This result suggests that naive human observers have at least some ability to predict which images are easy or difficult for machines to classify, even in less constrained scenarios than explored before (since the subjects in our experiments did not know which labels were among the relevant alternatives, nor which misclassification the machine made). These data thus provide early evidence that humans can intuit certain kinds of machine (mis)behaviors in naturalistic settings.

Experiment 2: How will AI misclassify?

The previous experiment suggested that people have some ability to anticipate when machines will misclassify images. But where do these intuitions come from? One source is surely an evaluation of *representativeness* — the sense that the image in question suitably resembles (or fails to resemble) its target class. However, an intriguing possibility is that subjects can appreciate not only why a particular image might not receive its intended category label, but also why *another* category label might be attractive—e.g., not only that the right image in Figure 1A won't be classified as a BUBBLE, but also that it has the features of a rival class that might make it prone to *mis*classification (rather than simply poor or unconfident classification). If true, this latter explanation would suggest that machines' failures on natural adversarial examples are not completely incomprehensible to humans, but might instead be caused by recognition of meaningful features that happen to be representative of a different image class (e.g., recognizing that an umbrella made of straw shares features with straw brooms).

To explore the possibility that humans can appreciate the overlap between natural adversarial examples and their adversarial target classes, Experiment 2 used a similar design to the previous experiment but with the addition of a "hint" on half of the trials. (Given our desire to see if these hints improved performance, we also modified the design to include six images on each trial—one adversarial image and five control images—so that baseline accuracy would be lower and there would be more room to observe an improvement.)

On each trial, subjects were told not only that one of the images was misclassified, but also what the misclassified image was misclassified as. For example, on a given trial, subjects might have seen five control BUBBLE images and one adversarial BUBBLE image, in which case they would then see the following instructions:

> *For each of these images, the machine was supposed to say* **bubble***. Most of them it got* **right***; but one of them it got* **wrong***. (Here's a hint: For the one it got wrong, it accidentally said* **salt shaker***.) Which image do you think the machine got* **wrong***?*

We reasoned that, if these hints improve subjects' ability to identify which image is adversarial, then

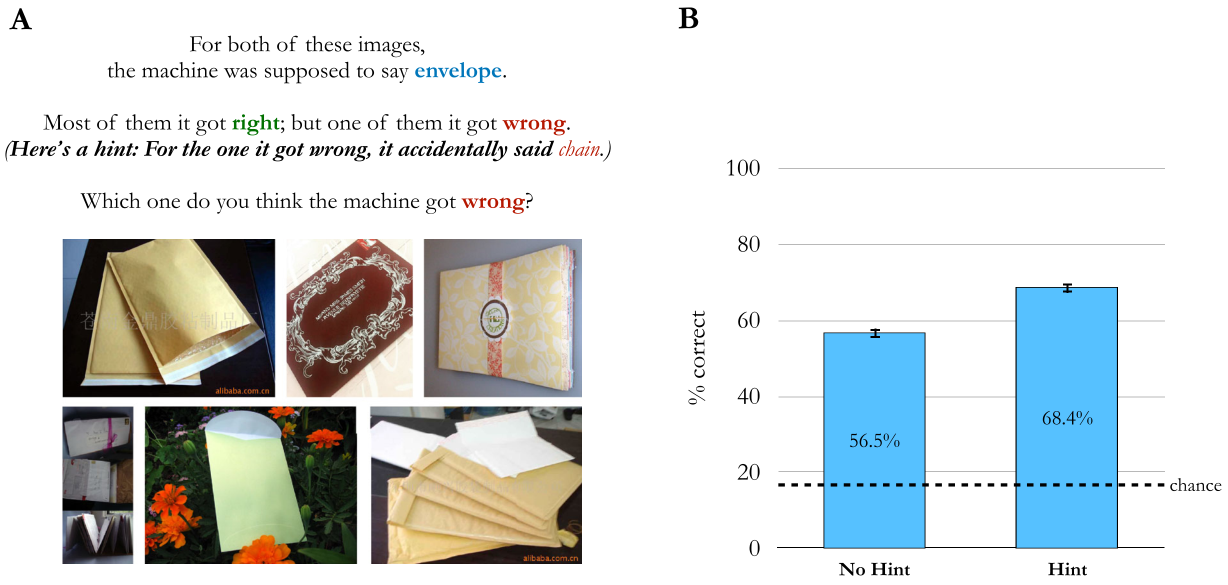

Figure 3. Results of Experiment 2. When asked which of six images was misclassified (e.g., which ENVELOPE image was misclassified), subjects identified the misclassified image on 56.5% of trials (where chance was 16.7%). However, when also informed of the nature of the misclassification (e.g., being asked not only which ENVELOPE image was misclassified, but more specifically which ENVELOPE image was misclassified as a CHAIN), performance improved to 68.4%, indicating that subjects could appreciate the overlap between the natural adversarial image and its target class. Error bars are ± 1 standard error of the difference between means. (The correct answer is the middle image of the top row.)

subjects must be able to appreciate the overlap between the natural adversarial images and their adversarial target classes (in this case, for example, seeing the resemblance between the adversarial BUBBLE image and a SALT SHAKER). By contrast, if subjects are simply deciding which images seem difficult to classify, or which images are poor examples of their parent class, then such hints should not be particularly helpful.

Other than the addition of hints on one-half of the trials, and the presence of six images per trial instead of two, the only other difference between this experiment and Experiment 1 was that the 250 adversarial images from Experiment 1 were decreased to only 50 images (one from each class), and each subject saw all 50 of the images (rather than seeing 50 of 250 total images). We made this modification because many of the adversarial misclassifications included those same obscure categories we eliminated in Experiment 1. (For example, whereas one of the adversarial DUCK images was misclassified as a RABBIT, one of them was misclassified as an OCARINA, which may not be familiar to many subjects.) To choose which one of the five adversarial images per class would appear in the experiment, we again selected the image whose adversarial target class returned the greatest number of Google hits.[2] Readers can experience this task at https://perceptionresearch.org/naturalAdversarial/e2.

Results and discussion

As in Experiment 1, subjects performed well above chance at selecting adversarial images over control images (with an overall accuracy of 62.4%, when chance would be 16.7%). However, performance was better in the presence of hints than without any such hints: 68.4% versus 56.5%, $t(156) = 10.20$, $p < .001$ (Figure 3). This result suggests that humans not only have intuitions about which images will be hard to classify, but also that they can understand which kinds of errors machines are likely to make, and even see the resemblance between misclassified images and the classes they are incorrectly placed into.

Experiment 3: Real hints versus fake hints

The previous two experiments suggest that ordinary observers have some ability to intuit when machines will misclassify natural images, and that they can even appreciate the kinds of misclassifications machines will make, as evidenced by their ability to take advantage of hints about the images' adversarial target classes. However, an alternative explanation of the performance boost provided by these hints is not that subjects could appreciate the resemblance between the images and their adversarial target classes, but rather that the hints simply enhanced the subjects' *motivation*. For example, it might be that, upon receiving a hint, subjects believe the trial should be easier, or they are motivated to study the image more closely, and so on. In that case, it is possible that the role of hints was not so different than a message saying "this is a trial where you should be able to get the right answer," regardless of any contentful information carried by the hint.

To rule out the possibility that a motivation-based confound accounts for these findings, Experiment 3 repeated the design of Experiment 2 but contrasted real hints with fake hints. As before, one-half of trials contained a (real) hint: the true misclassification given by the machine for whichever of the six images on that trial was the natural adversarial example. However, on the other one-half of the trials, instead of no hint at all, subjects received a fake hint: a randomly chosen label from the list of real hints that bore no correspondence to the actual natural adversarial image being shown. Regardless of condition (real hint or fake hint), subjects were asked on each trial to identify which of six images the machine misclassified. We hypothesized that subjects would perform better on trials with real hints than on trials with fake hints, which would indicate that the content of the hint is essential for facilitating better anticipation of machine misbehaviors. Readers can experience this task at https://perceptionresearch.org/naturalAdversarial/e3.

Results and discussion

We found that subjects correctly identified a greater proportion of natural adversarial examples on real hint trials than on fake hint trials. Group-level accuracy on real hint trials was 66.6%, compared with only 44.9% on fake hint trials, $t(138) = 12.06$, $p < .0001$ (Figure 4). (Note that, as in Experiment 2, chance performance was $\frac{1}{6}$, or 16.67%.) This provides evidence that the content of hints can help (or hinder) humans' ability to anticipate how machines misclassify natural images, and addresses the concern that motivation alone can account for the advantage conferred by hints in Experiment 2.

Taken together, Experiments 2 and 3 suggest that ordinary people can anticipate which natural images will be misclassified by machines, and that additional information about *how* the machine might have misclassified can help humans to anticipate machine errors with even greater accuracy. In other words, naive human subjects can tell not only which images are difficult to classify correctly, but can also appreciate the resemblance between misclassified images and the categories they are misclassified *as*.

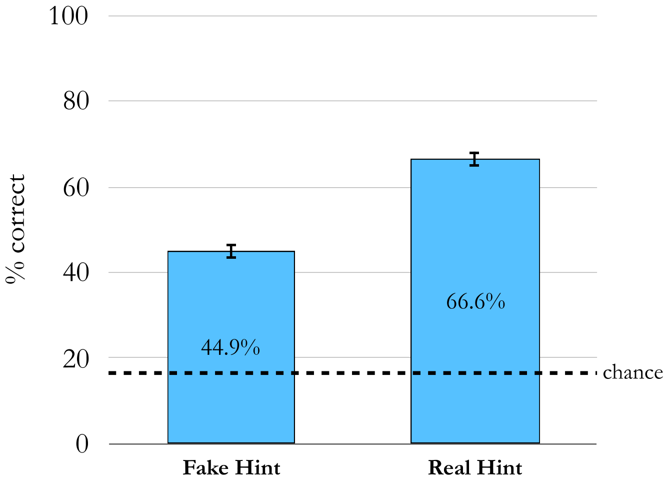

Figure 4. Results of Experiment 3. When given a real hint (e.g., being asked which ENVELOPE image was misclassified as a CHAIN, or which BUBBLE image was misclassified as a SALT SHAKER), subjects identified the misclassified image on 66.6% of trials, replicating the pattern found in Experiment 2. However, when given a fake hint (e.g., being asked which ENVELOPE image was misclassified as a BEE, or which BUBBLE image was misclassified as a BROOM [when in fact no ENVELOPE or BUBBLE images were misclassified that way]), performance was much worse. This suggests that subjects were able to appreciate the overlap between the natural adversarial images and their adversarial target classes. Error bars are ±1 standard error of the difference between means.

Experiment 4: The role of (un)representativeness

Experiment 1 revealed that ordinary human subjects can anticipate when natural images will be misclassified by leading machine recognition systems, and Experiments 2 and 3 provided evidence that this success is driven in part by appreciating the overlap between the image and a rival class. Thus, although judging an image to be unrepresentative of its target class is surely a (and perhaps *the*) primary driver of successful performance in our task, these results raise the possibility that subjects understand not only why the target label is less attractive but also why an alternative category label might be more attractive.

Experiment 4 explored this explanation for subjects' success in a different way, by explicitly investigating the role of representativeness. We split subjects into two groups, one of which was given the same task as in Experiment 1 (i.e., to guess which image was misclassified), and the other of which was not told anything at all about misclassification but instead was just asked which image was a *worse example* of its category. We expected subjects in both groups to perform quite well, given the clear role of representativeness in performing this task. However, we also wondered if the two conditions would produce different degrees of success. If subjects anticipate misclassifications *only* by judging an image to be unrepresentative of its target class, then such judgments should essentially be interchangeable with judgments of (un)representativeness. But if subjects' performance goes beyond mere judgments of (un)representativeness, then instructions to pick the misclassified image should lead subjects to perform better.

Methods

Experiment 4 recruited 200 subjects from Prolific. We chose this platform moving forward given evidence that its data quality had surpassed that of Amazon Mechanical Turk (Eyal, David, Andrew, Zak, & Ekaterina, 2022). Subjects in the control (misclassified) condition ($N = 100$ before exclusions) were given the following instructions, which were identical to the instructions given to subjects in Experiment 1:

> *We have a machine that can look at a picture and tell us what it is. Most of the time, it tells us the right answers. But sometimes it makes mistakes. We don't know why the machine makes the mistakes it makes; do you?*
>
> *Your job in this experiment is to predict when our machine will make a mistake. On each trial, you'll see a few images that will look to you like a familiar object. But, for one of them, the machine gave a different answer than it was supposed to. For each set of images, we want you to predict which image the machine got* **wrong**.
>
> [...]
>
> *For each of these images, the machine was supposed to say* **bubble**. *Most of them it got* **right**; *but one of them it got* **wrong**. *Which image do you think the machine got* **wrong**?

Another group of subjects ($N = 100$ before exclusions) in the worst example condition were instead instructed to identify the *worst example* of the image category:

> *We have some images of familiar objects. But some of these images are "bad examples"—they don't really look like what you'd expect when you think of that object. Can you tell us which ones are the worst examples?*
>
> *On each trial, you'll see a few image examples of a familiar object. For each set of images, we want you to select the* **worst example** *of that object.*
>
> [...]
>
> *These images are examples of* **bubble**. *Which image do you think is the* **worst** *example?*

The instructions for the worst example condition were intended to encourage subjects to choose what they consider to be the least representative example of the category.

Apart from the instruction change for one group of subjects, the design and trial structure was the same as Experiment 2: Subjects completed three catch trials and 50 experimental trials where six images were presented on each trial, one of which was a natural adversarial image. If subjects' selections in this task are based primarily on the sense that the image is a poor example of its category, then subjects in the worst example condition should choose the natural adversarial images just as often as subjects in the Misclassified condition. However, if human anticipation of machine misclassification goes beyond mere judgments of representativeness, then the Misclassified condition might show better performance than the worst example condition.

Readers can experience this task at https://perceptionresearch.org/naturalAdversarial/e4.

Results and discussion

Chance performance in this task is 16.67%. We found that subjects ($N = 184$ after exclusions) in both conditions performed well above-chance, indicating that both approaches (including representativeness) are effective for identifying natural adversarial images. However, those subjects who were instructed to choose the misclassified image selected the natural adversarial images more often than those who were instructed to choose the worst example of the image category, 59.7% vs. 56.3% of trials, $t(182) = 2.48$, $p = .014$. Clearly, both sets of instructions produced similar degrees of performance, suggesting that unrepresentativeness is quite a reliable guide to machine misclassification. However, together with Experiments 2 and 3 (which showed that subjects can appreciate the resemblance between these images and their rival classes), these results suggest that subjects can anticipate misclassifications by considering multiple sources of information, including both how (un)representative an image is of its class and also whether its features overlap with another class (although still other sources of information not tested here may be useful too).

Experiment 5: One at a time

The results presented thus far suggest that ordinary people can predict, to some degree, when machine vision systems will misclassify images in constrained settings—that is, in N-alternative-forced-choice tasks where a misclassified image is always presented next to at least one other correctly classified image of the same category. Yet, in real-world settings where predicting machine misclassifications might prove useful, it is more likely that people would be asked to detect potentially misclassified images in a continuous stream of individual images. Further, given the excellent image classification abilities of state-of-the-art models, the potentially misclassified images might be relatively rare. Can people still intuit machine misclassifications when natural adversarial images are both relatively infrequent and presented alone on a trial?

Methods

Experiment 5 recruited 100 subjects (equivalent to the sample size of each condition in Experiment 4). Each subject completed 300 trials, and each trial contained a single image from the set used in Experiments 2, 3, and 4 (250 from ImageNet and 50 from the Natural Adversarial Images dataset). Thus, the prevalence of adversarial images was low (16.67%). Subjects were reminded of the low prevalence of machine-misclassified images on every trial. Presented with each image were the following instructions:

> *The machine was supposed to say* **bubble** . *If you had to guess, did the machine get it* **right** *or* **wrong**? *Remember: Approximately 80% to 85% are right, and 15% to 20% are wrong!*

Subjects clicked a button labeled "right" if they thought the machine classified the image correctly, and "wrong" if they thought the machine classified the image incorrectly. In contrast with each of the previous experiments, there were six catch trials that appeared randomly throughout the experiment. Only the three undistorted images of a baseball, American football, and soccer ball used for catch trials in the previous experiments were used here. On three of the catch trials, these images were shown with the correct label (e.g., the baseball image was shown with text that read, "The machine was supposed to say *baseball*."). Subjects should have said that the machine got these images "right," because they are clear examples of the category that machines should have no trouble classifying. On the other three catch trials, those same images were shown with an incorrect label (e.g., the baseball image was shown with text that read, "The machine was supposed to say soccer ball."). Subjects should have said that the machine got these images "wrong," because they are not examples of the given category. Only subjects who answered at least five of the six catch trials correctly were included in the analysis.

Given the low prevalence of adversarial trials compared to nonadversarial trials (1/6 vs. 5/6), it is not appropriate to simply report raw accuracies and compare them with chance performance. For example, a subject who answers "wrong" for every image will perform at 100% on adversarial trials but 0% on normal trials and 16.67% overall; similarly, a subject who answers "right" for every image will perform at 0% on the adversarial trials but 100% on normal trials, for an overall accuracy of 83.33%; and even a subject who answers *randomly* (and whose performance is therefore 50%) may seem to have performed well above-chance on adversarial trials (performing at 50% accuracy despite the low prevalence of adversarial images). Thus, we report both raw accuracies and signal detection theory metrics which take into account the base rates of answering right and wrong, and allow computation of sensitivity (d') independent from bias. We predicted that subjects would have greater than zero sensitivity to the adversarial and non-adversarial nature of the stimuli.

Readers can experience this task at https://perceptionresearch.org/naturalAdversarial/e5.

Results and discussion

Subjects ($N = 69$ after exclusions) performed correctly on 79% of trials, answering 'right' for 84% of real images (which appeared at 83.3% prevalence) and 'wrong' for 57% of natural adversarial images (which appeared at 16.7% prevalence). Subjects on average said the machine got 21% of all the images wrong, roughly in line with the prevalence range given in the instructions (15%–20%). The false alarm rate (saying that the machine got a nonadversarial image wrong) was just 16%.

These rates of success and failure are also subject to signal detection analyses, which demonstrate that subjects detected the adversarial images in this task with sensitivity significantly greater than zero, average $d' = 1.26$, $t(68) = 48.3$, $p < .0001$. The detection rate for misclassifications could be even higher, considering subjects in Experiment 5 were conservative in saying that the machine got an image wrong ($c = 0.41$), consistent with behavior in other low prevalence search tasks (where miss rate tends to increase as target prevalence decreases Wolfe et al., 2007). For the kinds of real-world tasks we are imagining here—for example, a driver monitoring visual inputs to their semiautonomous vehicle—the costs of false alarms are fairly low (because additional human intervention is not especially problematic or costly), whereas the costs of misses can be quite high (because failing to intervene in cases of misclassification could result in an accident). Given this, encouraging a more liberal criterion may result in humans catching even more misclassified images, at an acceptable cost of increasing false alarms (and see the General Discussion for more on the practical relevance of these results.)

Supplementary analysis: "Truly" adversarial examples

The previous experiments show that subjects can appreciate when and how machine vision systems that are otherwise very accurate will misclassify natural images. However, it is possible that not all of the natural adversarial examples that appear in the previous experiments (which were drawn from Hendrycks et al., 2019) are actually appropriate to support these claims. In particular, while it is indeed the case that all of the images included in ImageNet-A (the database of natural adversarial examples) were given labels by machines that differed from human-chosen labels (thus making them misclassifications), careful examination of the images reveals that some misclassifications were more wrongheaded than others.

Figure 5 shows a variety of images included in the previous experiments. On one hand, it is clear that many of these images truly are embarrassing misclassifications, such as a BUTTERFLY being mistaken for a BROOM, or four hands each wearing a WATCH being mistaken for a SANDAL (Figure 5A); these images clearly meet the intuitive criteria for adversarial examples. But other misclassifications are simply not as concerning, and could plausibly be made by human observers too. For example, Figure 5B–D shows that ImageNet-A includes as misclassifications an ACOUSTIC GUITAR mistaken for a BANJO, two housecats on a COUCH being mistaken for a LYNX, or beads of water on a SPIDER

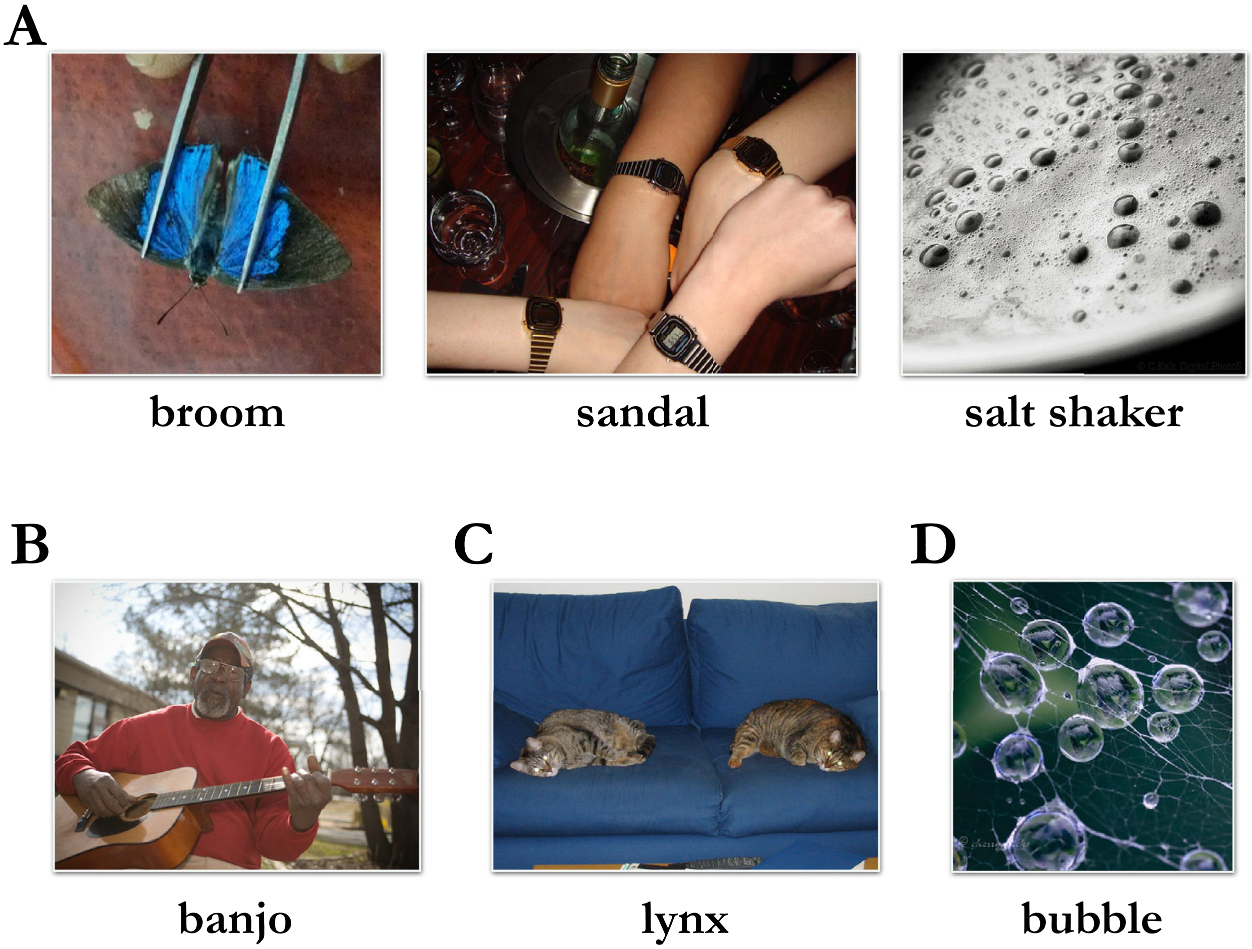

Figure 5. Not all natural adversarial examples are created equal. In some cases, images are misclassified as completely different objects, as in (**A**), which shows a BUTTERFLY misclassified as a BROOM. However, other misclassifications in the natural adversarial imageset seem less problematic: (**B**) shows an ACOUSTIC GUITAR misclassified as a BANJO; (**C**) shows two housecats on a COUCH misclassified as a LYNX; and (**D**) shows beads of water on a SPIDER WEB misclassified as a BUBBLE. These latter examples, although technically errors, are perhaps not so alarming as traditional adversarial misclassification (and, for similar reasons, it may be not be as impressive for humans to anticipate such misclassifications.) A supplementary analysis showed that, even considering only those images falling into category A, all of the effects observed remain, with the exception of Experiment 4 where the difference between worst example and misclassified conditions becomes marginal ($p = .051$).

WEB being mistaken for a BUBBLE. Although technically incorrect, these misclassifications probably do not carry most of the practical and theoretical implications of traditional adversarial misclassifications. For example, mistaking a guitar for a banjo seems unlikely to cause much mischief in a real-world machine vision system (because it is difficult to imagine how such a mistake could have major health, safety, or security implications); and this misclassification is also just not so unreasonable, such that it is not a paradigm case of an unhumanlike classification error of the sort typically taken to reveal significant human-machine differences. However, it is possible that performance in the previous experiments was driven mostly or only by these images, rather than the more canonical adversarial examples in Figure 5A—in which case human prediction of machine misclassification might not be so impressive after all. To rule out this possibility, we conducted a follow-up analysis designed to isolate only those truly adversarial examples, and then reran all of the previous analyses on only that subset of images.

Methods

We created a coding scheme allowing for four categories of images: a) ‘true adversarial examples’, in which an image that a human would typically classify one way is classified completely differently by the machine (e.g., WATCH → SANDAL); b) ‘near misses’, in which the image was misclassified as a rather close competitor (e.g., GUITAR → BANJO); c) ‘wrong object, right answer’, in which the machine gives a different label than a human, but only because it seems to classify a different object in the image (e.g., SPIDER WEB → BUBBLE); and d) ‘wrong object, near miss’, in which the machine seems to be classifying a different object than the intended target (as in b) and gets it nearly right (as in c; e.g., COUCH → LYNX). A more detailed description of these categories, along with multiple examples of each type, is available in our materials archive.

With these criteria established, the three authors of this paper (M.N., Z.Z., C.F.) reviewed all 250 natural adversarial examples used in the previous experiments, and hand-coded each image as belonging to one of these four categories (while remaining blind to how subjects had judged any particular images in the previous experiments). This process showed fairly high inter-rater reliability: All three raters agreed with one another for 80.8% of images (202/250), whereas by chance one would expect all three raters to agree on only 6.25% of images (given four categories and three raters). For each of the remaining 19.2% of images (48/250), at least two of the three raters gave the same rating, with only one rater disagreeing. Thus, to ensure that we identified only those images that unambiguously met the criteria for adversarial examples, we separated out all of those images for which i) all three raters agreed on which category it belonged to, and ii) that category was category A mention above (‘true adversarial examples’). This left 172 of 250 images (69%).

Finally, we simply reran all of the analyses from Experiments 1 to 5, but only over the true natural adversarial examples (of which there were 172 in Experiment 1, and 36 in Experiments 2–5).

Results and discussion

The general pattern of results remained the same in all five studies (though see below for detail), even when only true natural adversarial examples were included in the analysis.

In Experiment 1, accuracy was 79.82% on true natural adversarial examples, whereas it had been 79.77% on the full set of images. In Experiment 2, hints provided a 6% accuracy boost over no hint for true adversarial examples (vs. a 12% accuracy boost for the full set). In Experiment 3, real hints provided a 12% accuracy boost over fake hints (vs. a 21% accuracy boost). In Experiment 4, accuracy was 62.6% on true natural adversarial examples in the Misclassification condition and 59.6% in the worst example condition, whereas it had been 59.7% versus 56.3%, respectively, on the full set. (Although this difference became marginally significant, $p = .051$, it remained in the same direction and may have been due to the lesser statistical power in this experiment compared to every other experiment reported here). Finally, in Experiment 5, d′ was similar on true natural adversarial examples as compared to the whole set ($d' = 1.25$ vs. $d' = 1.26$).

In all five experiments, the primary effect of detecting adversarial images above-chance remained highly significant (all *p*s $< .0001$). Thus, even restricting the analyses to only those images that unambiguously meet the criteria for adversarial examples, subjects still showed an ability to appreciate when and how machines would misclassify these images.

General discussion

How intuitive are the errors made by state-of-the-art machine vision systems? The present results suggest that humans can anticipate several aspects of such errors, even in naturalistic settings with ordinary, unmodified images. The results of Experiment 1 suggest that humans can intuit when machines are likely to misclassify natural images, even without knowing which labels the machine prefers for the misclassified image. And the results of Experiments 2 and 3 suggest that human anticipation of such misclassifications is facilitated by knowing the actual labels machines selected, suggesting that they are able to “see why” the machines made the mistakes they did. Those intuitions seem to be underwritten by subjects’ ability to appreciate the overlap between features of the misclassified images and features of the incorrectly chosen image class: Experiment 4 shows that, although misclassified natural images are often poor examples of their human-assigned category labels, subjects’ anticipation of such misclassifications goes beyond mere (un)representativeness. Finally, Experiment 5 suggests that the human ability to detect machine misclassifications extends to more naturalistic presentations, including when no alternative images are shown and misclassifications are relatively rare.

Practical and theoretical implications

That humans can anticipate machine failures in this way may have implications for several issues at the interface of human and machine vision.

First, and more theoretically, the present results suggest that at least some machine failures are intuitive to naive human subjects, in ways that may speak to broader questions about overlapping processing between these systems. Although there are clear differences between human vision and state-of-the-art machine classification systems (Geirhos et al., 2020; Serre, 2019; Xu & Vaziri-Pashkam, 2022), it remains important to be clear about exactly which differences are meaningful (and what exactly they imply; Firestone, 2020; Funke et al., 2021). Our results suggest that, even in scenarios where a human would not fail like machines do, some of these machine failures are predictable and even intuitive to humans who have no knowledge, training, or expertise in machine vision—in ways that are difficult to account for without embracing at least *some* overlap in how the relevant images are seen. Put differently, even if these results do not reveal a processing similarity between human and machine vision, they perhaps recast the apparent *dis*similarity that these failures had seemed to imply. Relatedly, the present work also opens avenues for future research concerning human prediction of machine behavior in the domain of visual perception. Our studies intentionally recruited naive human observers with no knowledge, training, or expertise about the behavior of machine vision systems; however, future work could explore the ‘upper bound’ of humans’ predictive accuracy, by investigating the role of such expertise—for example, by exposing human subjects to examples of machine misclassifications and asking whether their predictive abilities improve to any interesting extent (Yang, Folke, & Shafto, 2022).

Second, and more practically, these results may have implications for real-world applications of automated visual classification systems (e.g., in vehicles with autopilot, or in computer-aided detection systems in radiology). Many real-world tasks where artificial visual recognition systems are deployed rely on semantic image segmentation models, which are built out of the DNN-based classification algorithms that produced the natural adversarial images dataset we used here (Kaymak & Uçar, 2019). At the same time, these systems may produce misclassifications that differ from the ones we studied here. Although natural adversarial images have been shown to transfer across different families of classification models (Hendrycks et al., 2019), it is an open question how well they transfer to other tasks, and so future work could explore human intuitions for natural image misclassifications made by semantic segmentation models. Another major question regarding such systems is how much human supervision is necessary, and whether that supervision would be effective (Fridman, Ding, Jenik, & Reimer, 2019). Although our experiments do not explore these systems in particular, our results suggest that human users of machine classification systems may have some ability to anticipate the conditions under which they will fail, in ways that may bear on analyses of the ideal amount of human supervision of such systems. Further, it may be that making humans aware of the potential for errors (e.g., the potential for an autonomous vehicle to mistake the broad side of a semitrailer for sky) could make collaborations between humans and machines safer (e.g., in making decisions about when to wrest control of the car from its autopilot function).

A similar approach could complement and extend our own Experiment 5, which aimed to explore the lower bound of these abilities by asking subjects to detect misclassifications when images were presented in a continuous stream at low prevalence (16.67%). A future experiment could lower error prevalence even further and test people in even more naturalistic settings (such as driving simulators) to explore whether the ability to detect machine misclassifications can actually be used by people to avoid dangerous outcomes such as vehicle collisions (perhaps in combination with other relevant factors, such as human trust in the machine’s competence and perceived authority; Gombolay, Gutierrez, Clarke, Sturla, & Shah, 2015; Jian, Bisantz, & Drury, 2000; Khadpe, Krishna, Fei-Fei, Hancock, & Bernstein, 2020). In addition to being potentially useful for online detection of inputs that are likely to cause misclassifications, human intuitions about when and why machines misclassify objects may be useful to scientists, engineers, or regulators focused on the safety of human–AI collaborations. There are a variety of different adversarial attacks (e.g., minimal or synthetic adversarial images) that present different challenges to human collaborators and may require different solutions. For example, synthetically graffitied stop signs that become misclassified as speed limit signs (Eykholt et al., 2018) present a problem that might be best addressed by engineered solutions, such as by cross-checking the car's location against known speed limits or traffic patterns typical of that area. On the other hand, natural adversarial inputs might well be better detected by humans, at least at present. In addition, if humans' signal for potential for misclassification could be sufficiently modeled, that algorithm could be a valuable addition to a suite of adversarial detector algorithms that are currently being designed to work alongside the primary visual recognition system (Meng & Chen, 2017) (although this approach of adding auxiliary detector algorithms has been criticized, as the models are computationally expensive and could open the door to new adversarial attacks themselves; Shumailov, Zhao, Mullins, & Anderson, 2020).

Finally, despite participants' ability to recognize the overlap between natural adversarial examples and machines' (mis)chosen categories, it is unknown whether humans and machines appreciate this overlap for the same reason—that is, whether they agree on which features are diagnostic of the adversarial target class within natural adversarial examples. This line of questioning is important, given that targeted/synthetic adversarial examples have been evaluated as evidence of potential divergence between artificial and biological visual systems (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos et al., 2018, 2020). To gain a more fine-grained insight into human intuition about machine misclassifications, future research may examine participants' ability to identify features in adversarial examples that machines recognize as diagnostic of their incorrectly chosen categories (e.g., by identifying the region of the image most responsible for the misclassification). Such work could further our understanding of the ways in which human and machine visual processing converge or diverge (for discussion, see Firestone, 2020; Funke et al., 2021; Serre, 2019).

Keywords: adversarial images, deep neural networks, natural scene classification

Acknowledgments

Commercial relationships: none.
Corresponding author: Chaz Firestone.
Email: chaz@jhu.edu.
Address: Department of Psychological & Brain Sciences, Johns Hopkins University, 3400 N Charles St., Baltimore, MD, USA.

*MN and ZZ are co-first authors.

Footnotes

[1] Goodfellow et al. (2017) offer the following definition: "Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake." If one considers a curated set of specially chosen natural images to be "intentionally designed," then natural adversarial examples are appropriately named; otherwise, perhaps not. In any case, we are less concerned here with the definition of the term "adversarial examples" and more with the question of machine misclassification, which is certainly raised by this imageset.

[2] A small number of image classes (CHEST, REEL, and JAY) have names that are homonymous with other popular search terms and so returned a large number of hits unrelated to the relevant image category; these classes were thus excluded from this process.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), e1006613.

Bos, N., Glasgow, K., Gersh, J., Harbison, I., & Lyn Paul, C. (2019). Mental models of AI-based systems: User predictions and explanations of image classification results. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*, 184–188.

Brendel, W., Rauber, J., Kurakin, A., Papernot, N., Veliqi, B., Salathé, M., . . . Bethge, M. (2018). Adversarial vision challenge. *arXiv preprint arXiv*:1808.01976.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence, 2*(12), 731–736.

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., . . . Zhou, W. (2016). Hidden voice commands. In *25th USENIX Security Symposium,* 513–530.

Carlini, N., & Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7.

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It takes two to tango: Towards theory of AI's mind. *arXiv preprint arXiv*:1704.00717.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife, 9*, e55978.

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., & Goodfellow, I., et al. (2018). Adversarial examples that fool both computer vision and time-limited humans. *arXiv preprint arXiv*:1802.08195.

Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*, 1643–1662.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1625–1634.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science, 363*(6433), 1287–1289.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America, 117*(43), 26562–26571.

Fridman, L., Ding, L., Jenik, B., & Reimer, B. (2019). Arguing machines: Human supervision of black box AI systems that make life-critical decisions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* pp. 1335–1343.

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision, 21*(3), 16–16.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., & Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*(11), 665–673.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; Increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv*:1811.12231.

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots, 39*(3), 293–312.

Goodfellow, I., Papernot, N., Huang, S., Duan, R., Abbeel, P., & Clark, J. (2017). Attacking machine learning with adversarial examples. *OpenAI Blog,* https://openai.com/research/attacking-machine-learning-with-adversarial-examples.

Harding, S., Rajivan, P., Bertenthal, B. I., & Gonzalez, C. (2018). Human decisions on targeted and non-targeted adversarial samples. In *Proceedings of the Annual Meeting of the Cognitive Sciences Society,* pp. 451–456.

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

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71.

Kaymak, Ç, & Uçar, A. (2019). A brief survey and an application of semantic image segmentation for autonomous driving. *Handbook of Deep Learning Applications,* pp. 161–200.

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–26.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv preprint arXiv*:1607.02533.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 52*(7553), 436–444.

Lepori, M. A., & Firestone, C. (2022). Can you hear me now? Sensitive comparisons of human and machine perception. *arXiv preprint arXiv*:2003.12362.

Meng, D., & Chen, H. (2017). Magnet: A two-pronged defense against adver sarial examples. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,* 135–147.

Morgulis, N., Kreines, A., Mendelowitz, S., & Weisglass, Y. (2019). Fooling a real car with adversarial traffic signs. *arXiv preprint arXiv*:1907.00374.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 427–436.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv*:2111.15366.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Shumailov, I., Zhao, Y., Mullins, R., & Anderson, R. (2020). Towards certiable adversarial sample detection. In *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security,* pp. 13–24.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., ... Fergus, R. (2013).

Intriguing properties of neural networks. *arXiv preprint arXiv*:1312.6199.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). The space of transferable adversarial examples. *arXiv preprint arXiv*:1704.03453.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136* (4), 623.

Xu, Y., & Vaziri-Pashkam, M. (2022). Understanding transformation tolerant visual object representations in the human brain and convolutional neural networks. *NeuroImage,* 119635.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Yang, S. C.-H., Folke, T., & Shafto, P. (2022). A psychological theory of explainability. *arXiv preprint arXiv*:2205.08452.

Yuille, A. L., & Liu, C. (2021). Deep nets: What have they ever done for vision? *International Journal of Computer Vision, 129*(3), 781–802.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1–9.