Controversial stimuli: Pitting neural networks against each other as models of human cognition

Tal Golan[a,1], Prashant C. Raju[b], and Nikolaus Kriegeskorte[a,c,d,e,1]

[a]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; [b]Department of Computer Science, Columbia University, New York, NY 10027; [c]Department of Psychology, Columbia University, New York, NY 10027; [d]Department of Neuroscience, Columbia University, New York, NY 10027; and [e]Department of Electrical Engineering, Columbia University, New York, NY 10027

Edited by Joshua B. Tenenbaum, Massachusetts Institute of Technology, Cambridge, MA, and accepted by Editorial Board Member Dale Purves September 15, 2020 (received for review November 13, 2019)

Distinct scientific theories can make similar predictions. To adjudicate between theories, we must design experiments for which the theories make distinct predictions. Here we consider the problem of comparing deep neural networks as models of human visual recognition. To efficiently compare models' ability to predict human responses, we synthesize controversial stimuli: images for which different models produce distinct responses. We applied this approach to two visual recognition tasks, handwritten digits (MNIST) and objects in small natural images (CIFAR-10). For each task, we synthesized controversial stimuli to maximize the disagreement among models which employed different architectures and recognition algorithms. Human subjects viewed hundreds of these stimuli, as well as natural examples, and judged the probability of presence of each digit/object category in each image. We quantified how accurately each model predicted the human judgments. The best-performing models were a generative analysis-by-synthesis model (based on variational autoencoders) for MNIST and a hybrid discriminative–generative joint energy model for CIFAR-10. These deep neural networks (DNNs), which model the distribution of images, performed better than purely discriminative DNNs, which learn only to map images to labels. None of the candidate models fully explained the human responses. Controversial stimuli generalize the concept of adversarial examples, obviating the need to assume a ground-truth model. Unlike natural images, controversial stimuli are not constrained to the stimulus distribution models are trained on, thus providing severe out-of-distribution tests that reveal the models' inductive biases. Controversial stimuli therefore provide powerful probes of discrepancies between models and human perception.

visual object recognition | deep neural networks | optimal experimental design | adversarial examples | generative modeling

Convolutional deep neural networks (DNNs) are currently the best image-computable models of human visual object recognition (1–3). To continue improving our computational understanding of biological object recognition, we must efficiently compare different DNN models in terms of their predictions of neuronal and behavioral responses of human and nonhuman observers. Adjudicating among models requires stimuli for which models make distinct predictions.

Here we consider the problem of adjudicating among models on the basis of their behavior: the classifications of images. Finding stimuli over which high-parametric DNN models disagree is complicated by the flexibility of these models. Given a sufficiently large sample of labeled training images, a wide variety of high-parametric DNNs can learn to predict the human-assigned labels of out-of-sample images. By definition, models with high test accuracy will mostly agree with each other on the classification of test images sampled from the same distribution the training images were sampled from.

Even when there is a considerable difference in test accuracy between two models, the more accurate model is not necessarily more human-like in the features that its decisions are based on. The more accurate model might use discriminative features not used by human observers. DNNs may learn to exploit discriminative features that are completely invisible to human observers (4, 5). For example, consider a DNN that learns to exploit camera-related artifacts to distinguish between pets and wild animals. Pets are likely to have been photographed by their owners with cellphone cameras and wild animals by photographers with professional cameras. A DNN that picked up on camera-related features might be similar to humans in its classification behavior on the training distribution (i.e., highly accurate), despite being dissimilar in its mechanism. Another model that does not exploit such features might have lower accuracy, despite being more similar to humans in its mechanism. To reveal the distinct mechanisms, we need to move beyond the training distribution.

There is mounting evidence that even DNN models that exhibit highly human-like responses when tested on in-distribution stimuli often show dramatic deviations from human responses when tested on out-of-distribution (OOD) stimuli (6). Prominent examples include images from a different domain [e.g., training a DNN on natural images and testing on silhouettes (7, 8)], as well as images degraded by noise or distortions (9–11), filtered (4), retextured (12), or adversarially perturbed to bias a DNN's classifications (13). Assessing a model's ability to predict human responses to OOD stimuli provides a severe test of the model's inductive bias, i.e., the explicit or implicit assumptions that allow it to generalize from training stimuli to novel stimuli. To correctly predict human responses to novel stimuli, a model has to have an inductive bias similar to that employed by humans. Universal function approximation by itself is insufficient. Previous studies have formally compared the responses of models and humans to distorted (9, 10) and adversarially perturbed images (14, 15), demonstrating the power of testing for OOD generalization. However, such stimuli are not guaranteed

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Brain Produces Mind by Modeling," held May 1–3, 2019, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019, colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/brain-produces-mind-by.

Author contributions: T.G., P.C.R., and N.K. designed research; T.G. and P.C.R. performed research; and T.G. and N.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.B.T. is a guest editor invited by the Editorial Board.

[1] To whom correspondence may be addressed. Email: n.kriegeskorte@columbia.edu or tal.golan@columbia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1912334117/-/DCSupplemental.

First published November 23, 2020.

to expose differences between different models, because they are not designed to probe the portion of stimulus space where the decisions of different models disagree.

Controversial Stimuli. Here we suggest testing and comparing DNN models of vision on controversial stimuli. A controversial stimulus is a sensory input (here, an image) that elicits clearly distinct responses among two or more models. Collecting human responses to stimuli that are controversial between two models gives us great power to adjudicate between the models. The human responses are guaranteed to provide evidence against at least one of the models, since they cannot agree with both models.

Once we define a controversiality score, we can search for such stimuli in large corpora or, more flexibly, synthesize them by optimization (Fig. 1). Stimulus synthesis need not be limited to any particular stimulus prior. If the candidate models differ mostly in how they classify in-distribution examples, an appropriate synthesis procedure, guided by the models' responses, will push the resulting controversial stimuli toward the training distribution. However, if out-of-distribution stimuli evoke considerably different responses among the candidate models, then stimulus synthesis can find them.

Controversial Stimuli vs. Adversarial Examples. Controversial stimuli generalize the notion of adversarial examples. An adversarial example is a stimulus controversial between a model and an oracle that defines the true label. A stimulus that is controversial between two models must be an adversarial example for at least one of them: Since the models disagree, at least one of them must be incorrect (no matter how we choose to define correctness). However, an adversarial example for one of two models may not be controversial between them: Both models may be similarly fooled (13, 16, 17). Controversial stimuli provide an attractive alternative to adversarial examples for probing models because they obviate the need for ground-truth labels during stimulus optimization. When adversarially perturbing an image, it is usually assumed that the perturbation will not also affect the true label (in most cases, the class perceived by humans). This assumption necessarily holds only if the perturbation is too small to matter (e.g., as in ref. 13). When the bound on the perturbation is large or absent, human observers and the targeted model might actually agree on the content of the image (14), making the image a valid example of another class. Such an image does not constitute a successful adversarial attack. The validity and power of a controversial stimulus, by contrast, are guaranteed given that the stimulus succeeds in making two models disagree.

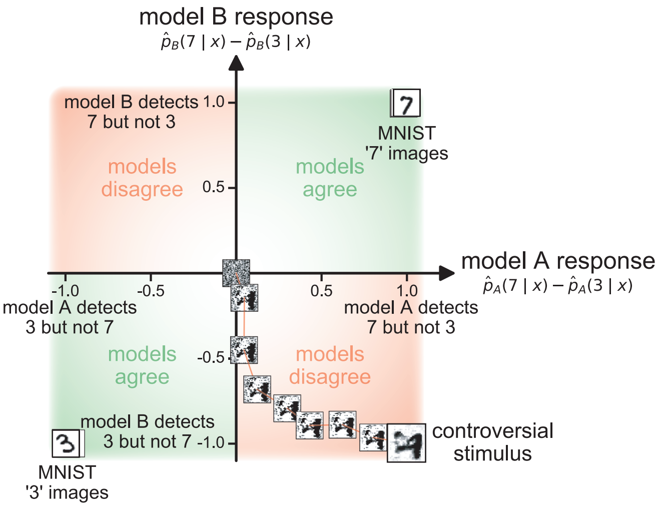

Fig. 1. Synthesizing a single controversial stimulus. Starting from an initial noise image, one can gradually optimize an image so two (or more) object recognition models disagree on its classification. Here, the resulting controversial stimulus (*Bottom Right*) is classified as a 7 by model A and as a 3 by model B. Testing such controversial stimuli on human observers allows us to determine which of the models has decision boundaries that are more consistent with the human decision boundaries. Often, "natural" examples (here 50 randomly selected test MNIST examples) cause no or minimal controversy among models and therefore lack the power to support efficient comparison of models with respect to human perception. Model A here is the Capsule Network reconstruction readout, and model B is small VGG$^-$. The stimulus synthesis optimization path (373 steps long) was sampled at nine roughly equidistant points.

Previous Work. Our approach is conceptually related to maximum differentiation (MAD) competition (18). MAD competition perturbs a source image in four directions: increasing the response of one model while keeping the response of the other fixed, decreasing the response of one model while keeping the response of the other fixed, and the converse pair. In contrast, a single controversial stimulus manipulates two (or more) models in opposite directions. Yet crudely speaking, our approach can be viewed as a generalization of MAD competition from univariate response measures (e.g., perceived image quality) to multivariate response measures (e.g., detected object categories) and from local perturbation of natural images to unconstrained search in image space.

Results

We demonstrate the approach of controversial stimuli on two relatively simple visual recognition tasks: the classification of hand-written digits [the MNIST dataset (19)] and the classification of 10 basic-level categories in small natural images [the CIFAR-10 dataset (20)]. From an engineering perspective, both tasks are essentially solved, with multiple, qualitatively different machine-learning models attaining near-perfect performance. However, this near-perfect performance on in-distribution examples does not entail that any of the existing models solve MNIST or CIFAR-10 the way humans do.

Synthesizing Controversial Stimuli. Consider a set of candidate models. We want to define a controversiality score for an image x. This score should be high if the models strongly disagree on the contents of this image.

Ideally, we would take an optimal experimental-design approach (21, 22) and estimate, for a given image, how much seeing the response would reduce our uncertainty about which model generated the data (assuming that one of the models underlies the observed human responses). An image would be preferred according to the expected reduction of the entropy of our posterior belief. However, this statistically ideal approach is difficult to implement in the context of high-level vision and complex DNN models without relying on strong assumptions.

Here we use a simple heuristic approach. We consider one pair of models (A and B) at a time. For a given pair of classes, y_a and y_b (e.g., the digits 3 and 7, in the case of MNIST), an image is assigned with a high controversiality score $c_{A,B}^{y_a,y_b}(x)$ if it is recognized by model A as class y_a and by model B as class y_b. The following function achieves this:

$$c_{A,B}^{y_a,y_b}(x) = \mathbf{min}\,\{\hat{p}_A(y_a \,|\, x), \hat{p}_B(y_b \,|\, x)\}, \tag{1}$$

where $\hat{p}_A(y_a \,|\, x)$ is the estimated conditional probability that image x contains an object of class y_a according to model A, and **min** is the minimum function. However, this function assumes that a model cannot simultaneously assign high probabilities to both class y_a and class y_b in the same image. This assumption is true for models with softmax readout. To make the controversiality score compatible also with less restricted (e.g., multilabel sigmoid) readout, we used the following function instead:

$$c_{A,B}^{y_a,y_b}(x) = \mathbf{min}\,\{\hat{p}_A(y_a \mid x), 1 - \hat{p}_A(y_b \mid x), \hat{p}_B(y_b \mid x), 1 - \hat{p}_B(y_a \mid x)\}. \qquad [2]$$

If the models agree over the classification of image x, then $\hat{p}_A(y_a \mid x)$ and $\hat{p}_B(y_a \mid x)$ will be either both high or both low, so either $\hat{p}_A(y_a \mid x)$ or $1 - \hat{p}_B(y_a \mid x)$ will be a small number, pushing the minimum down.

As in activation–maximization (23), we can use gradient ascent to generate images. Here we maximize Eq. **2** by following its gradient with respect to the image (estimated numerically for experiment 1 and symbolically for experiment 2). To increase the efficiency of the optimization and to avoid precision-related issues, the optimization was done on Eq. **4** (*Materials and Methods*), a numerically favorable variant of Eq. **2**. We initialized images with uniform white noise and iteratively ascended their controversiality gradient until convergence. A sufficiently controversial resulting image (e.g., $c_{A,B}^{y_a,y_b}(x) \geq 0.75$) is not guaranteed. A controversial stimulus cannot be found, for example, if both models associate exactly the same regions of image space with the two classes. However, if a controversial image is found, it is guaranteed to provide a test stimulus for which at least one of the models will make an incorrect prediction.

Experiment 1: Adjudicating among MNIST Models

Candidate MNIST Models. We assembled a set of nine candidate models, all trained on MNIST (*SI Appendix*, Table S1 and section A). The nine models fall into five families: 1) discriminative feedforward models, an adaptation of the VGG architecture (24) to MNIST, trained on either the standard MNIST dataset ("small VGG"; *SI Appendix*, section A.1) or on a version extended by nondigit images ("small VGG$^-$"; *SI Appendix*, Fig. S1); 2) discriminative recurrent models, the Capsule Network (25) ("CapsuleNet") and the Deep Predictive Coding Network (26) ("Wen-PCN-E4"); 3) adversarially trained discriminative models, DNNs trained on MNIST with either ℓ_∞ ("Madry ℓ_∞") or ℓ_2 ("Madry ℓ_2") norm-bounded perturbations (27); 4) a reconstruction-based readout of the Capsule Network (28) ("CapsuleNet Recon"); and 5) class-conditional generative models, models classifying according to a likelihood estimate for each class, obtained from either a class-specific, pixel-space Gaussian kernel density estimator ("Gaussian KDE") or a class-specific variational autoencoder (VAE), the "Analysis by Synthesis" model (29) ("Schott ABS").

Many DNN models operate under the assumption that each test image is paired with exactly one correct class (here, an MNIST digit). In contrast, human observers may detect more than one class in an image or, alternatively, detect none. To capture this, the outputs of all of the models were evaluated using multilabel readout, implemented with a sigmoid unit for each class, instead of the usual softmax readout. This setup handles the detection of each class as a binary classification problem (30).

Another limitation of many DNN models is that they are typically too confident about their classifications (31). To address this issue, we calibrated each model by applying an affine transformation to the preactivations of the sigmoid units (the logits) (31). The slope and intercept parameters of this transformation were shared across classes and were fitted to minimize the predictive cross-entropy on MNIST test images. For pretrained models, this calibration (as well as the usage of sigmoids instead of the softmax readout) affects only the models' certainty and not their classification accuracy (i.e., it does not change the most probable class of each image).

Synthetic Controversial Stimuli Reveal Deviations between MNIST Models and Human Perception. For each pair of models, we formed 90 controversial stimuli, targeting all possible pairs of classes. In experiment 1, the classes are the 10 digits. Fig. 2 shows the results of this procedure for a particular digit pair across all model pairs. Fig. 3 shows the results across all digit pairs for four model pairs.

Viewing the resulting controversial stimuli, it is immediately apparent that pairs of discriminative MNIST models can detect incompatible digits in images that are meaningless to us, the human observers. Images that are confidently classified by DNNs, but unrecognizable to humans are a special type of an adversarial example [described by various terms including "fooling images" (32), "rubbish class examples" (16), and "distal adversarial examples" (29)]. However, instead of misleading one model (compared to some standard of ground truth), our controversial stimuli elicit disagreement between two models. For pairs of discriminatively trained models (Fig. 3 *A* and *B*), human classifications are not consistent with either model, providing evidence against both.

One may hypothesize that the poor behavior of discriminative models when presented with images falling into none of the classes results from the lack of training on such examples. However, the small VGG$^-$ model, trained with diverse nondigit examples, still detected digits in controversial images that are unrecognizable to us (Fig. 3*A*).

There were some qualitative differences among the stimuli resulting from targeting pairs of discriminative models. Images targeting one of the two discriminative recurrent DNN models, the Capsule Network (25) and the Predictive Coding Network (26), showed increased (yet largely humanly unrecognizable) structure (e.g., Fig. 3*B*). When the discriminative models pitted against each other included a DNN that had undergone ℓ_2-bounded adversarial training (27), the resulting controversial stimuli showed traces of human-recognizable digits (Fig. 2; Madry ℓ_2). These digits' human classifications tended to be

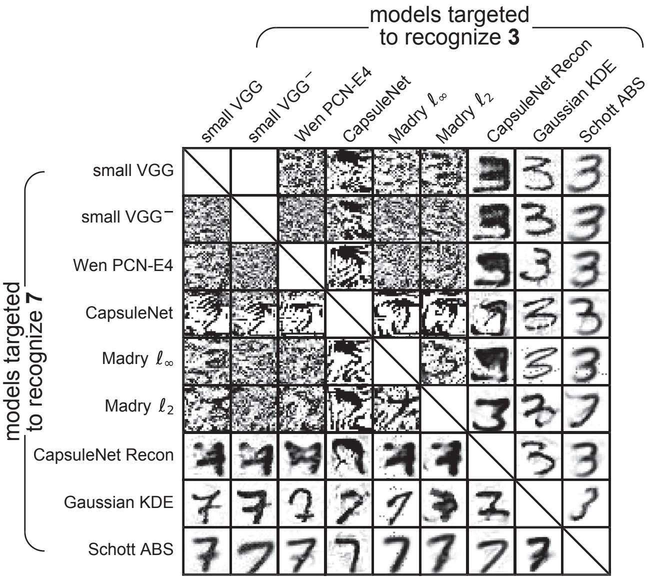

Fig. 2. Synthetic controversial stimuli for one digit pair and all pairs of MNIST models (experiment 1). All these images result from optimizing images to be recognized as containing a 7 (but not a 3) by one model and as containing a 3 (but not a 7) by the other model. Each image was synthesized to target one particular model pair. For example, the bottom-left image (seen as a 7 by us) was optimized so that a 7 will be detected with high certainty by the generative ABS model and the discriminative small VGG model will detect a 3. All images here achieved a controversiality score (Eq. **2**) greater than 0.75.

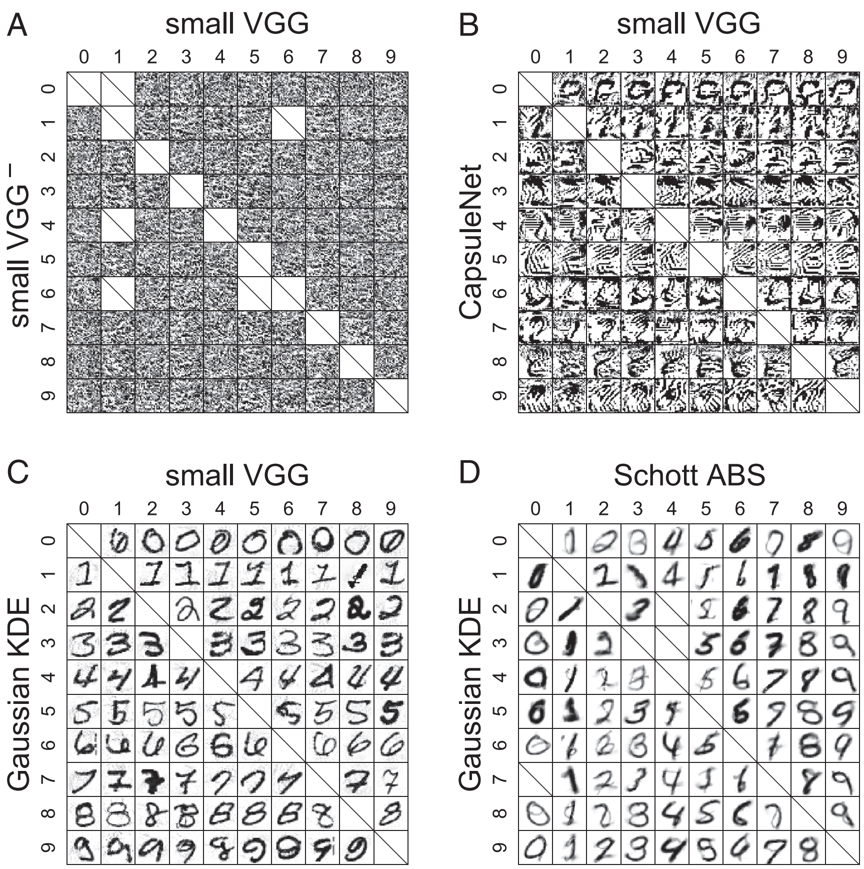

Fig. 3. (*A*–*D*) Synthetic controversial stimuli for all digit pairs and four different MNIST model pairs (experiment 1). The rows and columns within each panel indicate the targeted digits. For example, the top-right image in *D* was optimized so that a 9 (but not a 0) will be detected with high certainty by the Schott ABS model and a 0 (but not a 9) will be detected with high certainty by the Gaussian KDE model. Since this image looks like a 9 to us, it provides evidence in favor of Schott ABS over Gaussian KDE as a model of human digit recognition. Missing (crossed) cells are either along the diagonal (where the two models would agree) or where our optimization procedure did not converge to a sufficiently controversial image (a controversiality score of at least 0.75). See *SI Appendix*, Fig. S2 for all 36 model pairs.

consistent with the classifications of the adversarially trained discriminative model (see ref. 33, for a discussion of ℓ_2 adversarial training and perception).

And yet, when any of the discriminative models was pitted against either the reconstruction-based readout of the Capsule Network or either of the generative models (Gaussian KDE or ABS), the controversial image was almost always a human-recognizable digit consistent with the target of the reconstruction-based or generative model (e.g., Fig. 3*C*). Finally, synthesizing controversial stimuli to adjudicate between the three reconstruction-based/generative models produced images whose human classifications are most similar to the targets of the ABS model (e.g., Fig. 3*D*).

The ABS model is unique in having one DNN per class, raising the question of whether this, rather than its generative nature, explains its performance. However, imitating this structure by training 10 small VGG models as 10 binary classifiers did not increase the human consistency of the small VGG model (*SI Appendix*, Fig. S3). Another possibility is that a higher-capacity discriminative model with more human-like visual training on natural images might perform better. However, MNIST classification using visual features extracted from the hidden layers of an Imagenet-trained VGG-16 did not outperform the ABS model (*SI Appendix*, Fig. S4). Finally, the advantage of the ABS model persisted also when the optimization was initialized from MNIST test examples instead of random noise images (*SI Appendix*, Fig. S5).

Human Psychophysics Can Formally Adjudicate among Models and Reveal Their Limitations. Inspecting a matrix of controversial stimuli synthesized to cause disagreement among two models can provide a sense of which model is more similar to us in its decision boundaries. However, it does not tell us how a third, untargeted model responds to these images. Moreover, some of the resulting controversial stimuli are ambiguous to human observers. We therefore need careful human behavioral experiments to adjudicate among models.

We evaluated each model by comparing its judgments to those of human subjects and compared the models in terms of how well they could predict the human judgments. For experiment 1, we selected 720 controversial stimuli (20 per model-pair comparison; *SI Appendix*, section D) as well as 100 randomly selected MNIST test images. We presented these 820 stimuli to 30 human observers, in a different random order for each observer. For each image, observers rated each digit's probability of presence from 0 to 100% on a five-point scale (*SI Appendix*, Fig. S6*A*). The probabilities were not constrained to sum to 1, so subjects could assign high probability to multiple digits or zero probability to all of them for a given image. There was no objective reference for correctness of the judgments, and no feedback was provided.

For each human subject s_i and model M, we estimated the Pearson linear correlation coefficient between the human and model responses across stimuli and classes,

$$r(M, s_i) = \frac{\sum\limits_{x,y} \left(\hat{p}_{s_i}(y \,|\, x) - \bar{\hat{p}}_{s_i}\right)\left(\hat{p}_M(y \,|\, x) - \bar{\hat{p}}_M\right)}{\sqrt{\sum\limits_{x,y} \left(\hat{p}_{s_i}(y \,|\, x) - \bar{\hat{p}}_{s_i}\right)^2}\sqrt{\sum\limits_{x,y} \left(\hat{p}_M(y \,|\, x) - \bar{\hat{p}}_M\right)^2}}, \qquad [3]$$

where $\hat{p}_{s_i}(y \,|\, x)$ is the human-judged probability that image x contains class y, $\hat{p}_M(y \,|\, x)$ is the model's corresponding judgment, $\bar{\hat{p}}_{s_i}$ is the mean probability judgement of subject s_i, and $\bar{\hat{p}}_M$ is the mean probability judgment of the model. The overall score of each model was set to its mean correlation coefficient, averaged across all subjects: $\bar{r}_M = \frac{1}{n}\sum\limits_i r(M, s_i)$, where n is the number of subjects.

Given the intersubject variability and decision noise, the true model (if it were included in our set) cannot perfectly predict the human judgments. We estimated a lower bound and an upper bound on the maximal attainable performance (the noise ceiling; *SI Appendix*, section F). The lower bound of the noise ceiling ("leave one subject out"; black bars in Fig. 4 *A* and *B*) was estimated as the mean across subjects of the correlation between each subject's response pattern and the mean response pattern of the other subjects (34). The upper bound of the noise ceiling ("best possible model"; dashed lines in Fig. 4 *A* and *B*) is the highest across-subject-mean correlation achievable by any possible set of predictions.

The results of experiment 1 (Fig. 4*A*) largely corroborate the qualitative impressions of the controversial stimuli, indicating that the deep generative ABS model (29) is superior to the other models in predicting the human responses to the stimulus set. Its performance is followed by that of the Gaussian KDE, the reconstruction-based readout of the Capsule Network, and the ℓ_2 adversarially trained model. The other models (all discriminative) performed significantly worse. All models were significantly below the lower bound of the noise ceiling (the black bar in Fig. 4*A*), indicating that none of the models fully explained the explainable variability in the data.

We also evaluated the models separately for controversial stimuli and natural stimuli (i.e., MNIST test images; *SI Appendix*, Fig. S8*C*). The ABS and Gaussian KDE models were not as good as the discriminative models in predicting the human responses to the natural MNIST test images, indicating that the discriminative models are better at achieving human-like responses within the MNIST training distribution.

Experiment 2: Adjudicating among CIFAR-10 Models

The MNIST task has two obvious disadvantages as a test case: 1) its simplicity compared to visual object recognition in natural images and 2) the special status of handwritten characters, which are generated through human movement. In experiment 2, we applied the method of controversial stimuli to a set of models designed to classify small natural images from the CIFAR-10 image set. The purely generative ABS model is reported to fail to scale up to CIFAR-10 (29). We therefore included the Joint Energy Model (JEM) (35), which implements a hybrid discriminative–generative approach to CIFAR-10 classification.

Candidate CIFAR-10 Models. We assembled a set of seven CIFAR-10 candidate models (*SI Appendix*, Table S2 and section B). The seven models fall into five model families largely overlapping with the model families tested in experiment 1: 1) discriminative feedforward models, a VGG-16 (24) first trained on ImageNet and then retrained on upscaled CIFAR-10 ("fine-tuned VGG-16") and a Wide-Resnet trained exclusively on CIFAR-10 (36) ("Wide-Resnet"); 2) a discriminative recurrent model, a CIFAR-10 variant of the Deep Predictive Coding Network (26) ("Wen-PCN-A6"); 3) adversarially trained discriminative models, Resnet-50 DNNs trained on CIFAR-10 with either ℓ_∞ ("Engstrom ℓ_∞") or ℓ_2 ("Engstrom ℓ_2") norm-bounded perturbations (37); 4) a class-conditional generative model, the pixel-space Gaussian kernel density estimator ("Gaussian KDE"); and 5) a hybrid discriminative–generative model, the Joint Energy Model (35) ("Grathwol JEM-20").

The hybrid JEM has the same WRN-28-10 architecture (36) as the discriminative Wide-Resnet model mentioned above, but its training combines a discriminative training objective (minimizing

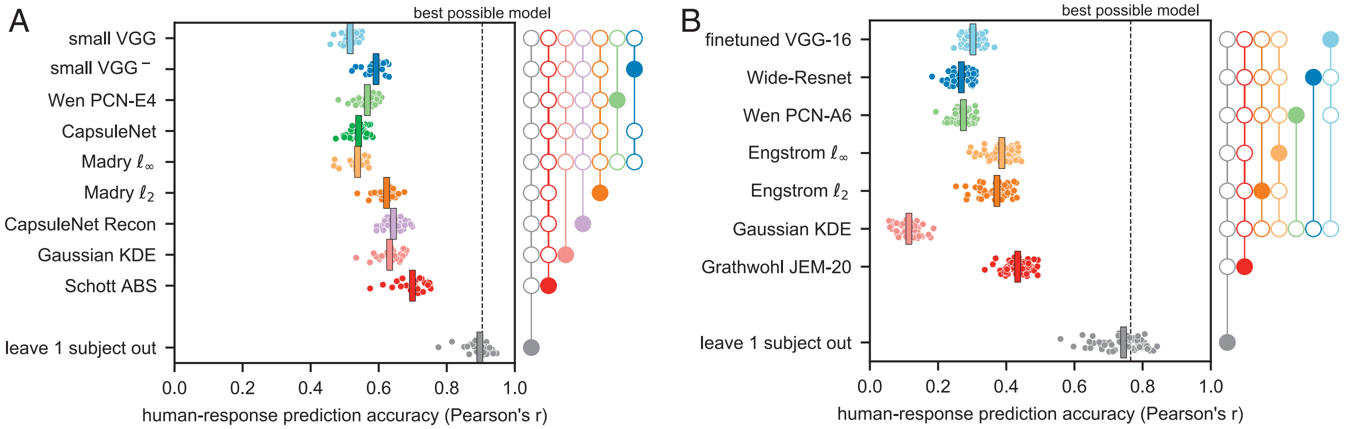

Fig. 4. The performance of the candidate MNIST (*A*) and CIFAR-10 (*B*) models in predicting the human responses to the entire stimulus set. Each dot marks the correlation coefficient between the responses of one individual human participant and one model (Eq. **3**). The vertical bars mark across-subject means ($\bar{r}_M$). The gray dots mark the correlation between each participant's responses and the mean response pattern of the other participants. The mean of the gray dots (a black bar) is the lower bound of the noise ceiling. The dashed line ("best possible model") marks the highest across-subject mean correlation achievable by any single model (upper bound of the noise ceiling). Significance indicators (*A* and *B*, *Right*): A solid dot connected to a set of open dots indicates that the model aligned with the solid dot has significantly higher correlation than any of the models aligned with the open dots ($P < 0.05$, subject-stimulus bootstrap). Testing controlled the familywise error rate at 0.05, accounting for the total number of model-pair comparisons (45 for experiment 1, 28 for experiment 2). For equivalent analyses with alternative measures of human-response prediction accuracy, see *SI Appendix*, Fig. S7. See *SI Appendix*, Fig. S8 for the models' prediction accuracy evaluated separately on controversial and natural stimuli and *SI Appendix*, Fig. S9 for an evaluation on different ratios of controversial to natural stimuli. The deep generative model (ABS, experiment 1) and the deep hybrid model (JEM-20, experiment 2) (both in red) explain human responses to the combined set of controversial and natural stimuli better than all of the other candidate models. And yet, none of the models account for all explainable variance: Predicting each subject from the subject's peers' mean response pattern achieves significantly higher accuracy.

the classification error) with a generative training objective. The generative objective treats the LogSumExp of the DNN's logits as an unnormalized image likelihood estimate and encourages high-likelihood assignments to in-distribution images. Including the generative objective in the training improves the model's robustness to adversarial attacks (35). The model's robustness can be further improved by refining the input-layer representation during inference, nudging it to have higher likelihood. We have tested the JEM model with 20 refinement steps (hence we refer to it here as "JEM-20").

As in experiment 1, we used sigmoid readout to allow for more flexible responses, such as detecting multiple or none of the categories. Since the candidate models had a wide range of test accuracies (*SI Appendix*, Table S2), the sigmoid readout was calibrated for each model such that negative examples would be assigned a median probability of 0.1 and positive examples a median probability of 0.9.

Synthetic Controversial Stimuli Reveal Deviations between CIFAR-10 Models and Human Perception. Examples of the resulting controversial stimuli appear in Fig. 5 and *SI Appendix*, Fig. S10. When DNNs trained with a nonadversarial discriminative objective (i.e., the finetuned VGG-16, the discriminatively trained Wide-Resnet, and the Predictive Coding Network) are paired with each other, the resulting controversial stimuli do not appear to humans to contain objects of any of the categories. These results bear strong resemblance to those in experiment 1. In contrast to experiment 1, however, the target categories for the Gaussian KDE were, by and large, not discernible to humans, indicating that this shallow-generative model, which worked surprisingly well on MNIST, does not scale up to CIFAR-10. Pitting the Gaussian KDE against the JEM-20 model (*SI Appendix*, Fig. S10*C*) produced almost naturally looking images, in which the target categories of JEM-20 are discernible. In some of these images, low-level features suggestive of the target category of the Gaussian KDE can also be recognized. Also, the target categories of the adversarially trained models were more discernible than in experiment 1 (*SI Appendix*, Fig. S10 *A* and *B*). Finally, pitting the JEM-20 model against one of the adversarially trained models (*SI Appendix*, Fig. S10*D*) often produced images in which the target category for JEM-20 was discernible. In some images, however, the human-perceptible category was the target of the adversarially trained DNN or both or neither of the categories were perceptible. These ambiguities suggest deviations of both JEM-20 and the adversarially trained DNNs from human perception and emphasize the importance of quantitative behavioral experiments.

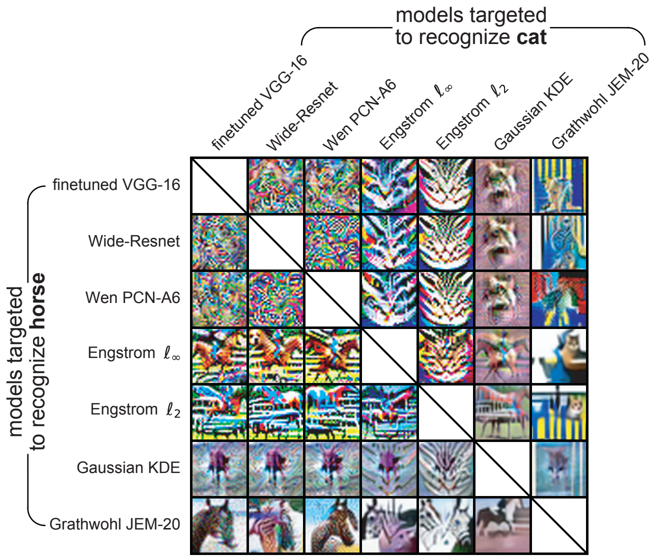

Fig. 5. Synthetic controversial stimuli contrasting the seven different CIFAR-10-classifying models. Each stimulus results from optimizing an image to be detected as a cat (but not as a horse) by one model and as a horse (but not as a cat) by another model. For example, the image at *Bottom Left* (seen as a horse by us) was optimized so that the hybrid discriminative–generative JEM-20 model will detect a horse and the discriminative, finetuned VGG-16 model will detect a cat. All images here achieved a controversiality score (Eq. **2**) greater than 0.75. The images are shown in upsampled format as presented to the human subjects. See *SI Appendix*, Fig. S11 for all class combinations.

We ran a behavioral experiment similar to experiment 1, presenting 420 controversial stimuli (20 per model-pair comparison) as well as 60 randomly selected CIFAR-10 test images. We ran two replications of the experiment on 30 subjects each, using a new, independent batch of controversial stimuli for each replication. The results pooled over both replications (60 subjects) are presented in Fig. 4*B*, whereas the (virtually identical) results of each individual replication are presented in *SI Appendix*, Fig. S12.

On average across the stimulus set, JEM-20 was significantly more accurate at predicting the human perceptual judgments than all other models. Similarly to experiment 1, none of the models reached the lower bound of the noise ceiling (the leave-one-subject-out estimate). The two adversarially trained models (trained on ℓ_∞ and ℓ_2 bounded perturbations) were second to the JEM-20 model in their human-response prediction accuracy. Next was the finetuned VGG-16 model and then the discriminatively trained Wide-Resnet and the Predictive Coding Network. The Gaussian KDE had the lowest human-response prediction accuracy.

Measuring the human response-prediction accuracy separately for controversial stimuli (*SI Appendix*, Fig. S8*B*) showed no significant difference between the JEM-20 model and the adversarially trained DNNs. For the natural images, however, the JEM-20 model significantly outperformed the adversarially trained DNNs (*SI Appendix*, Fig. S8*D*). The model that best predicted the human responses to the natural images was the finetuned ImageNet-trained VGG-16, indicating that no single model in our candidate set was uniformly dominant, as would be expected of the true model.

Discussion

In this paper, we introduce the method of synthetic controversial stimuli, and we demonstrate its utility for adjudicating among DNNs as models of human recognition in the context of two simple visual recognition tasks, MNIST and CIFAR-10. Controversial stimuli reveal model differences and empower us to find failure modes, capitalizing on the fact that if two models disagree, at least one of them must be wrong.

The method of controversial stimuli can be useful to two groups of scientists. The first group is cognitive computational neuroscientists interested in better understanding perceptual processes, such as object recognition, by modeling them as artificial neural networks. The second group is computer scientists interested in comparing the robustness of different DNN models to adversarial attacks.

Controversial Stimuli Offer a Severe Test for DNNs as Brain-Computational Models. Natural stimuli will always remain a necessary benchmark for models of perception. Moreover, at the moment it is still feasible to compare and reject DNNs as models of human vision on the basis of their classifications of natural, nonmodel-driven stimuli (e.g., refs. 38–42). As DNN models become better at fitting the training distribution, such informative errors on natural examples are expected to diminish. Scientists designing experiments comparing the human consistency of models can search for natural controversial stimuli to increase the power of model comparison. However, even for the models we have today, natural stimuli (including controversial ones) do not provide a severe test. In particular, a mechanistically incorrect model with many parameters that has been trained on natural images can achieve high performance at predicting human-assigned labels of images sampled from the same distribution. Synthetic controversial stimuli that are not limited to the training distribution provide a severe test of a model's inductive bias because they require the model to generalize far beyond the training distribution. Synthetic controversial stimuli ensure that we do not favor models that are higher-capacity function approximators regardless of their functional consistency with human perception. Here, the considerably different model rank orders observed when considering synthetic controversial stimuli and when considering natural stimuli (*SI Appendix*, Figs. S8 and S9) indicate that these two benchmarks shed light on complementary facets of model–human consistency.

Controversial Stimuli Generalize Adversarial Attacks. Engineers use adversarial examples to test the robustness of models. Adversarial examples can be viewed as a special case of controversial stimuli. An ideal adversarial example is controversial between the targeted model and ground truth. In principle, therefore, adversarial examples require the evaluation of ground truth in the optimization loop. However, the evaluation of ground truth is often difficult, because it may be costly to compute or may require human judgment. In practice, adversarial attacks usually use a stand-in for ground truth, such as the assumption that the true label of an image does not change within a pixel-space ℓ_p ball of radius ϵ.

Controversial stimulus synthesis enables us to compare two models in terms of their robustness without needing to evaluate or approximate the ground truth within the optimization loop. We require only a single ground-truth evaluation once the optimization is completed to determine which of the models responded incorrectly. Hence, controversial stimuli enable us to use more costly and compelling evaluations of ground truth (e.g., human judgments or a computationally complex evaluation function), instead of relying on a surrogate measure.

The most common surrogate measure for ground truth is ϵ robustness. A model is said to be ϵ robust if perturbations of the image confined to some distance in image space (defined by an ℓ_p norm) do not change the model's classification. The notion of ϵ robustness has led to analytical advances and enables adversarial training (5, 27). However, since ϵ robustness is a simple surrogate for a more complicated ground truth, it does not preclude the existence of adversarial examples and so does not guarantee robustness in a more general sense. This is particularly evident in the case of object recognition in images, where the ground truth is usually human categorization: A model can be ϵ robust for a large ϵ and yet show markedly human-inconsistent classifications, as demonstrated by controversial stimuli (here), distal adversaries (29), and "invariance attacks" (43), in which a human subject manually changes the true class of an image by making modifications confined to an ℓ_p ball in image space. The motivating assumption of ϵ robustness is that the decision regions are compact and their boundaries are far from the training examples. This does not hold in general. Controversial stimuli allow us to find failure modes in two or more models by studying differences in their decision boundaries instead of relying on assumptions about the decision boundaries.

Controversial Stimuli: Current Limitations and Future Directions. Like most works using pretrained models (1, 2), this study operationalized each model as a single trained DNN instance. In this setting, a model predicts a single response pattern, which should be as similar as possible to the average human response. To the extent that the training of a model results in instances that make idiosyncratic predictions, the variability across instances will reduce the model's performance at predicting the human responses. An alternative approach to evaluating models would be to use multiple instances for each model (44), considering each DNN instance as an equivalent of an individual human brain. In this setting, each model predicts a distribution of input–output mappings, which should be compared to the distribution of stimulus–response mappings across the human population. Instance-specific idiosyncrasies may then be found to be consistent (or not) with human idiosyncratic responses.

Another limitation of our current approach is scaling up: Synthesizing controversial stimuli for every pair of classes and every pair of models is difficult for problems with a large number of classes or models. A natural solution to this problem would be subsampling, where we do not synthesize the complete cross-product of class pairs and model pairs.

Future research should also explore whether it is possible to replace the controversiality index with an optimal experimental design approach, jointly optimizing a stimulus set to reduce the expected entropy of our posterior over the models. Finally, adaptive measurement between or within experimental sessions could further increase the experimental efficiency.

Generative Models May Better Capture Human Object Recognition. One interpretation of the advantage of the best-performing models (the VAE-based analysis by synthesis model in experiment 1 and the Joint Energy Model in experiment 2) is that, like these two models, human object recognition includes elements of generative inference. There has recently been considerable progress with DNNs that can estimate complex image distributions (e.g., VAEs and normalizing-flow models). However, such approaches are rarely used in object recognition models, which are still almost always trained discriminatively to minimize classification error. Our direct testing of models against each other suggests that DNN classifiers that attempt to learn the distribution of images (in addition to being able to classify) provide better models of human object recognition.

However, none of the tested models approached the noise ceiling, and while the ABS and JEM models performed better than all of the other models on average, they were worse than some of the discriminative models when the natural examples were considered in isolation (*SI Appendix*, Fig. S8 *C* and *D*). Each of these two outcomes indicates that none of the models were functionally equivalent to the process that generated the human responses.

Generative models do not easily capture high-level, semantic properties of images (45, 46). In particular, this problem is evident in the tendency of various deep generative models to assign high likelihood to out-of-distribution images that are close to the mean low-level statistics of the in-distribution images (45). Hybrid (discriminative–generative) approaches such as the joint energy model (35) are a promising middle ground, yet the particular hybrid model we tested (JEM-20) was still far from predicting human responses accurately. An important challenge is to construct a generative or hybrid model that 1) reaches the noise ceiling in explaining human judgments, 2) scales up to real-world vision (e.g., ImageNet), and 3) is biologically plausible in both its architecture and training. The method of controversial stimuli will enable us to severely test such future models and resolve the question of whether human visual judgments indeed employ a process of generative inference, as suggested by our results here.

Materials and Methods

Further details on training/adaptation of candidate models, stimulus optimization and selection, human testing, and noise-ceiling estimation appear in *SI Appendix*.

Controversial Stimuli Synthesis. Each controversial stimulus was initialized as a randomly seeded, uniform noise image ($x \sim \mathcal{U}(0, 1)$, where 0 and 1 are the image intensity limits). To efficiently optimize the controversiality score (Eq. **2**), we ascended the gradient of a more numerically favorable version of this quantity:

$$\tilde{c}_{A,B}^{y_a,y_b}(x) = \boldsymbol{LSE}_{\alpha}^{-} \{l_A(y_a \mid x), -l_A(y_b \mid x), l_B(y_b \mid x), -l_B(y_a \mid x)\}, \qquad [4]$$

where $\boldsymbol{LSE}_{\alpha}^{-} = -\log \sum_i exp^{-\alpha x_i}$ (an inverted LogSumExp, serving as a smooth minimum), α is a hyperparameter that controls the LogSumExp smoothness (initially set to 1), and $l_A(y \mid x)$ is the calibrated logit for class y (the input to the sigmoid readout). Experiment-specific details on stimulus optimization appear in *SI Appendix*, sections C.1 and C.2.

Human Subjects. Ninety participants took part in the online experiments and were recruited through prolific.co. All participants provided informed consent at the beginning of the study, and all procedures were approved by the Columbia Morningside ethics board.

Statistical Inference. Differences between models with respect to their human response prediction accuracy were tested by bootstrapping-based hypothesis testing. For each bootstrap sample (100,000 resamples), subjects and stimuli were both randomly resampled with replacement. Stimuli resampling was stratified by stimuli conditions (one condition per model pair, plus one condition of natural examples). For each pair of models M_1 and M_2, this bootstrapping procedure yielded an empirical sampling distribution of $\bar{r}_{M_1} - \bar{r}_{M_2}$, the difference between the models' prediction accuracy levels. Percentages of bootstrapped accuracy differences below (or above) zero were used as left-tail (or right-tail) P values. These P values were Holm–Šídák corrected for multiple pairwise comparisons and for two-tailed testing.

Data Availability. Optimization source code, synthesized images, and anonymized, detailed behavioral testing results are available at GitHub, https://github.com/kriegeskorte-lab/PNAS_2020_Controversial_Stimuli.

ACKNOWLEDGMENTS. This material is based upon work supported by the National Science Foundation under Grant 1948004. T.G. acknowledges Edmond and Lily Safra Center for Brain Sciences (ELSC) Postdoctoral Fellowships for Training Abroad. NVIDIA donated a Titan Xp used in this research. We thank Máté Lengyel for a helpful discussion and Raphael Gerraty, Heiko Schütt, Ruben van Bergen, and Benjamin Peters for their comments on the manuscript.

1. N. Kriegeskorte, Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis.* **1**, 417–446 (2015).
2. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
3. T. C. Kietzmann, P. McClure, N. Kriegeskorte, "Deep neural networks in computational neuroscience" in *Oxford Research Encyclopedia of Neuroscience*, (Oxford University Press, 2019).
4. J. Jo, Y. Bengio, Measuring the tendency of CNNs to learn surface statistical regularities. arXiv:1711.11561 (30 November 2017).
5. A. Ilyas *et al.*, *Adversarial Examples Are Not Bugs, They Are Features in Advances in Neural Information Processing Systems 32*, H. Wallach *et al.*, Eds. (Curran Associates, Inc., 2019), pp. 125–136.
6. R. Geirhos *et al.*, Shortcut learning in deep neural networks. arXiv:2004.07780 [cs, q-bio] (20 May 2020).
7. J. Kubilius, S. Bracci, H. P. Op de Beeck, Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12**, e1004896 (2016).
8. N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
9. S. Dodge, L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions" in *2017 26th International Conference on Computer Communication and Networks* (ICCCN, Vancouver, Canada, 2017), pp. 1–7.
10. R. Geirhos *et al.*, "Generalization in humans and deep neural networks" in *Advances in Neural Information Processing Systems 31*, S. Bengio *et al.*, Eds. (Curran Associates, Inc., 2018), pp. 7538–7550.
11. D. Hendrycks, T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations" in *7th International Conference on Learning Representations* (ICLR, New Orleans, LA, 2019).
12. R. Geirhos *et al.*, "ImageNet-trained CNNs are biased toward texture; Increasing shape bias improves accuracy and robustness" in *7th International Conference on Learning Representations* (ICLR, New Orleans, LA, 2019).
13. C. Szegedy *et al.*, Intriguing properties of neural networks. arXiv:1312.6199 [cs] (19 February 2014).
14. Z. Zhou, C. Firestone, Humans can decipher adversarial images. *Nat. Commun.* **10**, 1334 (2019).
15. G. Elsayed *et al.*, "Adversarial examples that fool both computer vision and time-limited humans" in *Advances in Neural Information Processing Systems 31*, S Bengio *et al.*, Eds. (Curran Associates, Inc., 2018), pp. 3910–3920.
16. I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples" in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA* (2015).
17. Y. Liu, X. Chen, C. Liu, D. Song, "Delving into transferable adversarial examples and black-box Attacks" in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France* (2017).
18. Z. Wang, E. P. Simoncelli, Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *J. Vis.* **8**, 1–13 (2008).
19. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
20. A. Krizhevsky, "Learning multiple layers of features from tiny images" (Tech. Rep., University of Toronto, Toronto, ON, Canada, 2009).
21. D. V. Lindley, On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005 (1956).
22. N. Houlsby, F. Huszár, Z. Ghahramani, M. Lengyel, Bayesian active learning for classification and preference learning. arXiv:1112.5745 (24 December 2011).
23. D. Erhan, Y. Bengio, A. Courville, P. Vincent, "Visualizing higher-layer features of a deep network" (Tech. Rep. 13411, University of Montreal, Montreal, QC, Canada, 2009).
24. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (10 April 2015).
25. S. Sabour, N. Frosst, G. E. Hinton, "Dynamic routing between capsules" in *Advances in Neural Information Processing Systems 30*, I Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), pp. 3856–3866.
26. H. Wen *et al.*, "Deep predictive coding network for object recognition" in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (PMLR, Stockholmsmässan, Stockholm, Sweden, 2018), vol. 80, pp. 5266–5275.
27. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, "Toward deep learning models resistant to adversarial attacks" in *6th International Conference on Learning Representations* (ICLR, Vancouver, Canada, 2018).
28. Y. Qin *et al.*, "Detecting and diagnosing adversarial images with class-conditional capsule reconstructions" in *8th International Conference on Learning Representations* (ICLR, 2020).
29. L. Schott, J. Rauber, M. Bethge, W. Brendel, "Toward the first adversarially robust neural network model on MNIST" in *7th International Conference on Learning Representations* (ICLR, New Orleans, LA, 2019).
30. M.-L. Zhang, Z.-H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. Data Eng.* **18**, 1338–1351 (2006).
31. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, "On calibration of modern neural networks" in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. The, Eds. (Association for Computing Machinery, 2017), vol. 70, pp. 1321–1330
32. A. Nguyen, J. Yosinski, J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images" in *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR, 2015).
33. D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, "Robustness may Be at odds with accuracy" in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA* (2019).
34. H. Nili *et al.*, A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, 1–11 (2014).
35. W. Grathwohl *et al.*, "Your classifier is secretly an energy based model and you should treat it like one" in *7th International Conference on Learning Representations* (ICLR, New Orleans, LA, 2019).
36. S. Zagoruyko, N. Komodakis, Wide residual networks. arXiv:1605.07146 (14 June 2017).
37. L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, *Robustness* (Python Library, 2019).
38. K. M. Jozwik, N. Kriegeskorte, K. R. Storrs, M. Mur, Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* **8**, 1726 (2017).
39. R. Rajalingham *et al.*, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
40. J. C. Peterson, J. T. Abbott, T. L. Griffiths, Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* **42**, 2648–2669 (2018).
41. R. M. Battleday, J. C. Peterson, T. L. Griffiths, Capturing human categorization of natural images at scale by combining deep networks and cognitive models. arXiv:1904.12690 (26 April 2019).
42. M. Schrimpf *et al.*, Brain-Score: Which artificial neural network for object recognition is most brain-like?. bioRxiv:407007 (5 September 2018).
43. J. H. Jacobsen, J. Behrmannn, N. Carlini, F. Tramèr, N. Papernot, Exploiting excessive invariance caused by norm-bounded adversarial robustness. arXiv:1903.10484 [cs, stat] (25 March 2019).
44. J. Mehrer, C. J. Spoerer, N. Kriegeskorte, T. C. Kietzmann, Individual differences among deep neural network models. bioRxiv:2020.01.08.898288 (9 January 2020).
45. E. Nalisnick, A. Matsukawa, Y. Whye Teh, D. Gorur, B. Lakshminarayanan, "Do deep generative models know what they don't know?" in *7th International Conference on Learning Representations* (ICLR, New Orleans, LA 2019).
46. E. Fetaya, J. H. Jacobsen, W. Grathwohl, R. Zemel, "Understanding the limitations of conditional generative models" in *The International Conference on Learning Representations* (ICLR, 2020).

PNAS
www.pnas.org

Supplementary Information for

Controversial stimuli: pitting neural networks against each other as models of human recognition

Tal Golan, Prashant C. Raju and Nikolaus Kriegeskorte

Tal Golan and Nikolaus Kriegeskorte.
E-mail: tal.golan@columbia.edu, n.kriegeskorte@columbia.edu

This PDF file includes:

Supplementary text
Figs. S1 to S12
Tables S1 to S2
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1912334117

Supporting Information Text

A. Candidate MNIST models (Experiment 1). Most of the tested models (Table S1) were based on official pre-trained versions (1–4), unmodified except for the readout layer. Here we describe the models we trained from scratch or more deeply altered.

A.1. Small VGG. Starting from the VGG-16 architecture (architecture D in Table 1 of reference 5), we downsized its input to the 28×28 pixels MNIST format, removed the deepest three convolutional layers and replaced the three fully-connected layers with a single, 512-unit fully-connected layer, feeding a ten-sigmoid readout layer. All weights were initialized using the Glorot uniform initializer, as implemented in Keras. Batch normalization was applied between the convolution and the ReLU operations in all convolutional layers. The model was trained with Adagrad (learning rate $= 10^{-3}$, $\epsilon = 10^{-8}$, decay=0) for 20 epochs using a mini-batch size of 128. The epoch with best validation performance (evaluated on 5000 MNIST held-out training examples) was used.

A.2. Reconstruction-based readout of the Capsule Network. In the training procedure of the original Capsule network (2), the informativeness of the class-specific activation vectors ('DigitCaps') is promoted by minimizing the reconstruction error of a decoder that is trained to read out the input image from the vector activation related to each example's correct class. (6, 7) suggested using this reconstruction error during inference, flagging examples with high reconstruction error (conditioned on their inferred class) as potentially adversarial. While rejecting suspicious images and avoiding their classification is a legitimate engineering solution, for a vision model we require that class conditional probabilities ($\hat{p}(y \mid x)$) will always be available. Hence, instead of using the reconstruction error as a rejection criterion, we used it as a classification signal. We modified the same official pre-trained Capsule Network (https://github.com/Sarasra/models/tree/master/research/capsules) used in our testing of the original Capsule Network such that for each image during inference, the decoder network produced ten class-specific input image reconstructions. The ten class-conditional mean squared reconstruction-errors were fed into ten sigmoids, whose response was calibrated as described in the results section. To eliminate the bias of this error measure towards blank images, we normalized the reconstruction error of each class by dividing it by the mean squared difference between the input image and the average image of all MNIST training examples (averaged across classes).

A.3. Gaussian KDE. For each class y, we formed a Gaussian KDE model, $\hat{p}(x \mid y) = \frac{1}{n\sigma_y}\sum_{i=1}^{n} K\left(\frac{x-x_i^y}{\sigma_y}\right)$ where σ_y is a class-specific bandwidth hyper-parameter, $K(\cdot)$ is a multivariate Gaussian likelihood with unit covariance, and $\{x_i^y\}$ are all MNIST training examples labeled as class y. σ_y was chosen independently for each class from the range $[10^{-2}, .., 10^{0}]$ (100 logarithmic steps) to maximize the likelihood of held-out 500 training examples. The ten resulting log-likelihoods were fed as penultimate activations to a sigmoid readout layer, calibrated as described in the results section.

B. Candidate CIFAR-10 models (Experiment 2). As in Experiment 1, most of the tested CIFAR-10 models (Table S2) were based on official pre-trained versions (1, 8), unmodified except for the readout layer. Models we trained from scratch or more deeply altered are described below.

B.1. Finetuned VGG-16. We initiated this model from TorchVision's (9) batch-normalized implementation of VGG-16 (5), pretrained on ImageNet. We added a differentiable bilinear upsampling operation from 32×32 to 128×128 pixels before the first layer, replaced the DNN's final readout layer with ten sigmoids, one per CIFAR-10 class, and retrained the network on the CIFAR-10 training dataset to minimize the cross-entropy loss between the sigmoid outputs and CIFAR-10 one-hot labels for up to 100 epochs. We used stochastic gradient descent (PyTorch's implementation) with a learning rate of 0.001 and a momentum of 0.9. 5000 training examples (500 per class) were held out to serve as a validation dataset. We applied early stopping and selected the best epoch in terms of validation loss.

B.2. Gaussian KDE. Fitting the pixel-space, class-specific CIFAR-10 Gaussian KDE followed the same procedure described for the MNIST Gaussian KDE (see SI subsection A.3), using a vectorized form of the $32 \times 32 \times 3$ RGB image representation as features. σ_y was chosen independently for each class from the range $[10^{-2}, .., 10^{2}]$ (501 logarithmic steps).

B.3. Joint Energy Model. We used the official code and pretrained model (https://github.com/wgrathwohl/JEM) with two modifications: we increased the number of inference-time logit-refinement steps to 20 and reduced the stochasticity of this refinement by increasing the number of sampling chains from 5 to 30.

B.4. Wide-Resnet. We trained the discriminative control of the JEM model ('Wide-Resnet') by executing the training code of JEM after adjusting the generative objective weight, '`p_x_weight`', from 1.0 to 0.0. During inference, this model did not perform logit refinement.

C. Synthesis of controversial stimuli: experiment-specific details.

C.1. Controversial-stimulus synthesis in Experiment 1 (MNIST). Each controversial stimulus was initialized to a floating-point 28×28 uniform white random noise matrix ($x \sim \mathcal{U}(0.0, 1.0)$, where 0.0 and 1.0 correspond to MNIST's 0 and 255 integer intensity levels, respectively). While for most candidate models included in Experiment 1, one can derive an analytical gradient of Eq. 4, this is not possible for the ABS model since its inference is based on latent space optimization. Hence, following (4)'s approach to forming adversarial examples, we used numerical differentiation for all models. In each optimization iteration, we used the symmetric finite difference formula $\frac{f(x+h)-f(x-h)}{2h}$ to estimate the gradient of Eq. 4 with respect to the image. An indirect benefit of this approach is that one can set h to be large, trading gradient precision for better handling rough cost-landscapes. For each image, we began optimizing using $h = 1$ (clipping $(x + h)$ and $(x - h)$ to stay within the the grayscale intensity range). Once the optimization converged to a local maximum, we halved h and continued optimizing. We kept halving h upon convergence until final convergence with $h = 1/256$. We then increased the LSE hyperparameter α to 10 and reset h to equal 1.0 again, repeating the procedure (but without resetting the optimized image). A third and final optimization epoch used $\alpha = 100$.

In each optimization iteration, once a gradient estimate was determined we used a line search for the most effective step size: We evaluated the effect of the maximal step in the direction of the gradient that did not cause intensity clipping, as well as $[2^{-1}, 2^{-2}, ..., 2^{-8}]$ of this step size.

When the optimization converged to an image that had a controversiality score (Eq. 2) of less than 0.85 we repeated the optimization procedure with a different initial random image, up to five attempts.

For analytically differentiable MNIST models, we found that this more involved (and more computationally intensive) approach to image optimization resulted in less convergence to poor local maxima (i.e., images with low controversiality) compared to standard gradient ascent using symbolic differentiation.

C.2. Controversial-stimulus synthesis in Experiment 2 (CIFAR-10). Since all of the candidate CIFAR-10 models were differentiable, we applied a symbolic-differentiation-based stimulus synthesis procedure in Experiment 2. Unlike the MNIST case, optimizing controversiality with the symbolic gradients of CIFAR-10 models rarely led to convergence to poor local maxima, potentially indicating a smoother cost landscape associated with the CIFAR-10 models.

The optimized image was parameterized as an unconstrained floating point 32×32×3 matrix x_0. This representation was transformed to an intensity image x (with pixel intensities constrained between 0.0 and 1.0) by means of the sigmoid function, $x = \frac{1}{1+\exp(-x_0)}$. The image x was fed to the target models, and eventually presented to the human subjects. x was initilized as a uniform random noise image ($x \sim \mathcal{U}(0.0, 1.0)$) and the corresponding initial x_0 was set by the inverse sigmoid transform, $x_0 = \log(\frac{x}{1-x})$.

Using Eq. 4 as our optimization objective, we applied Adam (10) ($\alpha = 0.1$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) to the unbounded image representation x_0. The smooth-minimum sharpness parameter α in Eq. 4 was initially set to 1.0. The optimizer was run to convergence and then resumed (without resetting the image) with a sharpness parameter of 10.0, and then finally with a sharpness parameter of 100.0. The convergence criterion was an improvement of less than 0.1% in the maximal controversiality score in the last 50 time steps compared to the maximal controversiality score in the time steps that preceded this window. When the resulting image had a controversiality score (Eq. 2) of less than 0.85, we repeated the optimization procedure with a different initial random image (up to five attempts). In most cases, even a single repetition was not needed.

D. Selection of controversial stimuli for human testing. For each model pair, we selected 20 controversial stimuli for human testing (out of up to 90 we produced). Using integer programming (IBM DOcplex), we searched for the set of 20 images with the highest total controversiality score, under the constraint that each class is targeted exactly twice per model. This was done separately for Experiment 1 and Experiment 2.

E. Human testing. The two human experiments were conducted using a custom javascript interface. In addition to collecting the perceptual judgments, we monitored reaction times to detect too quick responses. Trials completed in less than 100 ms were rejected post hoc, treating the corresponding perceptual judgments as missing values. No participant took part in more than one experiment.

E.1. Experiment 1. 30 participants (17 women, mean age = 29.3) participated in Experiment 1.

We monitored the participants' performance through three measures: their accuracy on the 100 MNIST images, their reaction times, and the reliability of their responses to 108 controversial images (3 per model pair) that were displayed again at the end of the experiment. While the participants' performance on these measures varied, we found no basis for rejecting the data produced by any participant due to evident low effort or negligence.

E.2. Experiment 2. A total of 60 participants (25 women, mean age = 26.1) participated in the two replications of Experiment 2.

Since CIFAR-10 categories do not naturally map to response keys as MNIST categories do, we altered Experiment 1's graphical user interface to saliently display the mapping of categories to response keys (Fig. S6B). This mapping was randomized for each participant.

As in Experiment 1, we monitored the performance of the human subjects through three measures: their accuracy on the 60 CIFAR-10 test images, their reaction times, and the reliability of their responses to 42 controversial images (2 per model pair) that were displayed again at the end of the experiment. The data of two participants were excluded due to suspected low effort performance (very short completion time and all-zero ratings for several natural CIFAR-10 test examples). These two participants are not included in the abovementioned total count of participating subjects.

We found in pilot runs that a minority of the participants interpreted the task in an overly conservative way, assigning all-zero responses (i.e., no hint of an object) to all images that were not natural. We eliminated this kind of response pattern by including the following instruction:

```
    In each of the images that you will see, there will be hints for at least one of the ten object categories.
If you see anything that reminds one of the ten object categories, rate the relevant category with at least 25%.
```

```
We will give a 5 USD worth bonus payment to any participant who will do well in detecting objects that are especially
hard to recognize.
```

This bonus was paid after the experiment to half of the participants according to an objective criterion. Since the bonus was awarded offline, it did not serve as feedback (which was intentionally absent in both experiments).

F. Noise-ceiling estimates.

F.1. Lower bound (leave-one-subject-out). Here we further detail the calculation of the leave-one-subject-out estimate, which serves as a lower bound on the noise ceiling (i.e., the true model should be at least as accurate as this estimate). To calculate this estimate, we held out one subject s_i at a time and averaged the response patterns of all of the other subjects: If $\hat{p}_{s_i}(y \mid x)$ is the probability judgment provided by subject s_i for image x and category y, then the leave-one-subject-out prediction for this subject, image and category is given by $\hat{q}_{s_i}(y \mid x) = \frac{1}{n-1} \sum\limits_{j \neq i} \hat{p}_{s_j}(y \mid x)$. Considering all stimuli and categories, one can represent subject s_i's response pattern and the corresponding leave-one-subject-out predicted response pattern as two vectors of matching lengths, $\boldsymbol{p}_{s_i}$ and $\boldsymbol{q}_{s_i}$. These vectors would have 8200 elements for the responses of an Experiment 1 subject (820 images × 10 response categories) or 4800 elements for the responses of an Experiment 2 subject (480 images × 10 response categories). The linear correlation between $\boldsymbol{p}_{s_i}$ and $\boldsymbol{q}_{s_i}$ measures how well can the subject's response pattern be predicted from the mean response pattern of her/his peers:

$$r(S, s_i) = \frac{\sum\limits_{x,y} \left(\hat{p}_{s_i}(y \mid x) - \bar{\hat{p}}_{s_i}\right)\left(\hat{q}_{s_i}(y \mid x) - \bar{\hat{q}}_{s_i}\right)}{\sqrt{\sum\limits_{x,y} \left(\hat{p}_{s_i}(y \mid x) - \bar{\hat{p}}_{s_i}\right)^2} \sqrt{\sum\limits_{x,y} \left(\hat{q}_{s_i}(y \mid x) - \bar{\hat{q}}_{s_i}\right)^2}}, \qquad \text{[S.1]}$$

where $\bar{\hat{p}}_{s_i}$ is the average probability judgment of subject s_i and $\bar{\hat{q}}_{s_i}$ is the average probability judgment of the corresponding leave-one-subject-out predicted response pattern.

The resulting leave-one-subject-out correlation coefficients for each held-out subject were plotted as gray dots at the bottom of Fig. 4A and Fig. 4B. The mean leave-one-subject correlation coefficient (averaged across subjects) was marked as a vertical bar in these two figures and was statistically tested against the mean model-human correlation coefficients of each candidate model. For analyses that included model recalibration, an inverse sigmoid transform was applied to the $\boldsymbol{q}_{s_i}$ vectors to produce logits, which were then tuned in the same fashion the models' logits were recalibrated, sharing the same scale and shift parameters across different held-out subjects, exactly matching the level of flexibility in fitting each model's predictions to the human data by scaling and shifting the model's logits.

F.2. Upper bound ('best possible model'). The upper bound on the noise ceiling (marked as 'best possible model' in Fig. 4A and Fig. 4B) was determined by optimization. We initiated a vector of logits (one value per image-category combination, e.g., 8200 elements for Experiment 1, or 4800 elements for Experiment 2) as a zero vector and optimized this vector with L-BFGS so that the across-subject mean of the correlation coefficients between each subject and a sigmoid transform of this vector is maximized. The resulting mean correlation coefficient reflects the inherent limitation of predicting variable individual response patterns by a single response pattern.

Note that for the simple case of no missing values, the vector that maximizes the across-subject average correlation is directly obtainable by transforming each individual response vector to a z-scored vector and then averaging the resulting z-scored vectors across subjects (11, supplementary materials). Here, we used the more general optimization approach since the rejection of trials with too short reaction times led to missing values which complicate the analytical derivation of the correlation-maximizing response-vector.

G. Additional software tools used. TensorFlow (12) (Experiment 1), Keras (13) (Experiment 1), and PyTorch (14) (Experiments 1 & 2) were used for DNN training and testing; psiTurk (15) was used for as the backend of the online experiments; Numpy (16), XArray (17), pandas (18), scikit-learn (19), and statsmodels (20) were used for data analysis; matplotlib (21) and seaborn (22) were used for visualization.

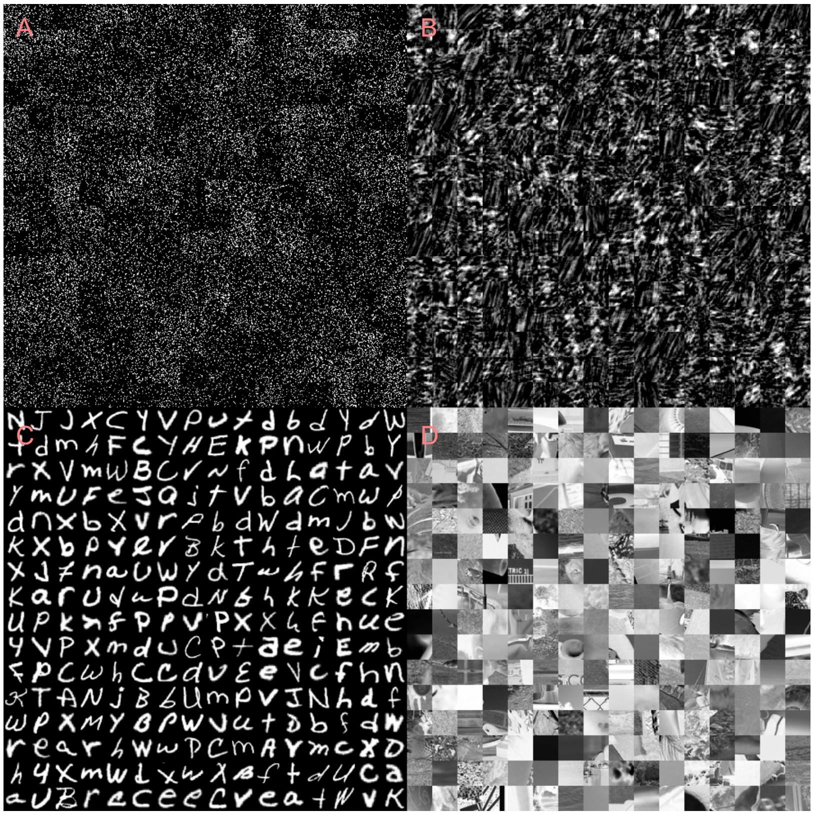

Fig. S1. A random sample of the negative examples (non-digits) used as a background class for the small VGG$^-$ model. (A) pixel-scrambled MNIST images. (B) Fourier-phase scrambled MNIST-images. (C) EMNIST letters (23), excluding the letters o,s,z,l,i,q and g. (D) patches cropped from natural images. The small VGG$^-$ model was trained on the MNIST dataset plus a dataset of a similar size per each of these four non-digit classes (so the digit images were only a fifth of the training set). The non-digit class labels' were coded as all-zero rows in the one-hot coding.

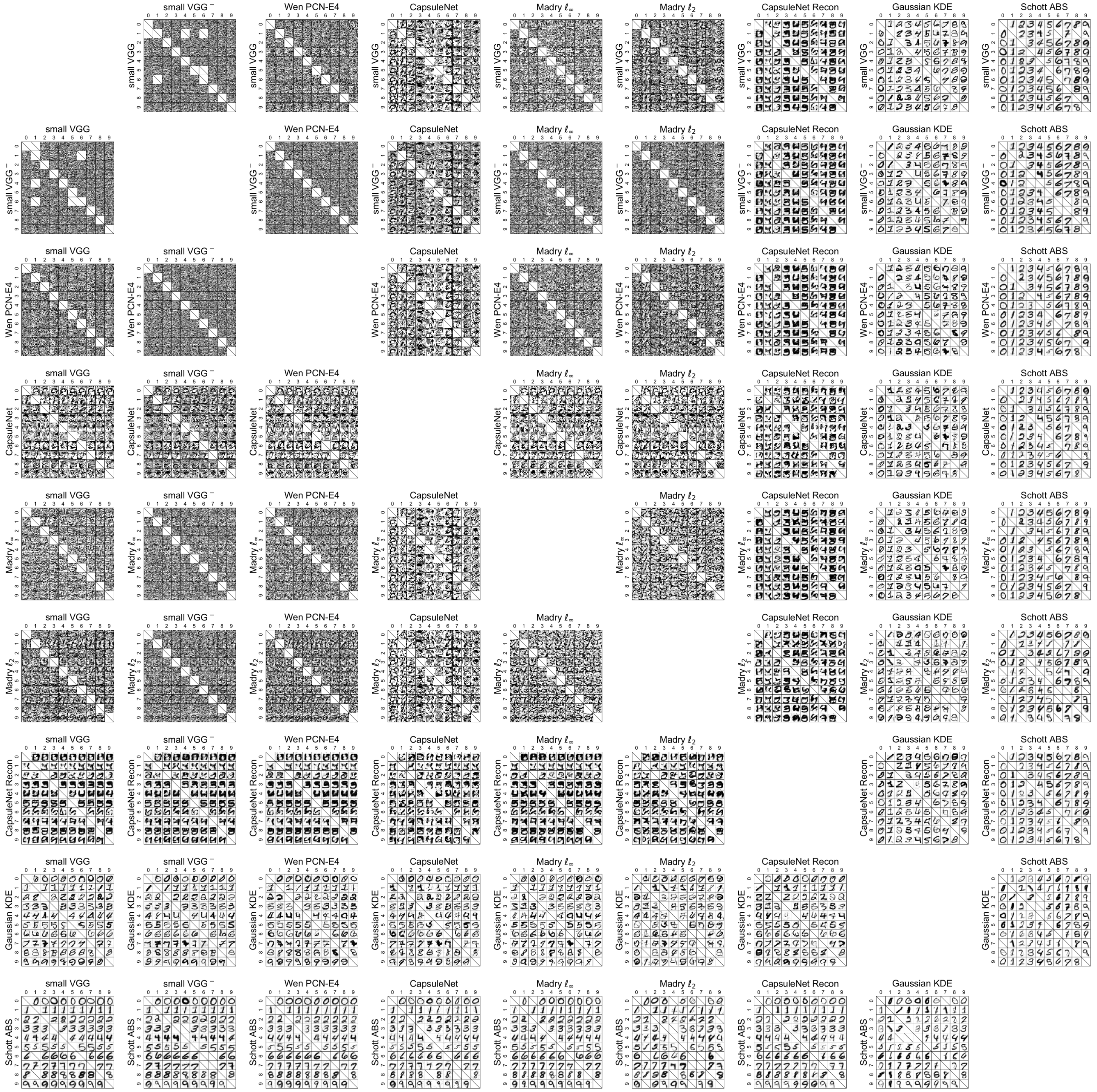

Fig. S2. The entire set of controversial stimuli we synthesized for Experiment 1 (MNIST), organized by model pairs. Each panel indicates a targeted model pair and the rows and columns within each panel indicate the targeted class pairs. For example, consider the bottom-left image in the bottom-left panel. This image (seen to us as a 9) is detected as a 0 by the small VGG model and as a 9 by the Schott ABS model. Missing (crossed) cells are either along the diagonal (where the two models would agree) or where our optimization procedure did not converge to a sufficiently controversial image (a controversiality score of at least 0.75). Best viewed digitally.

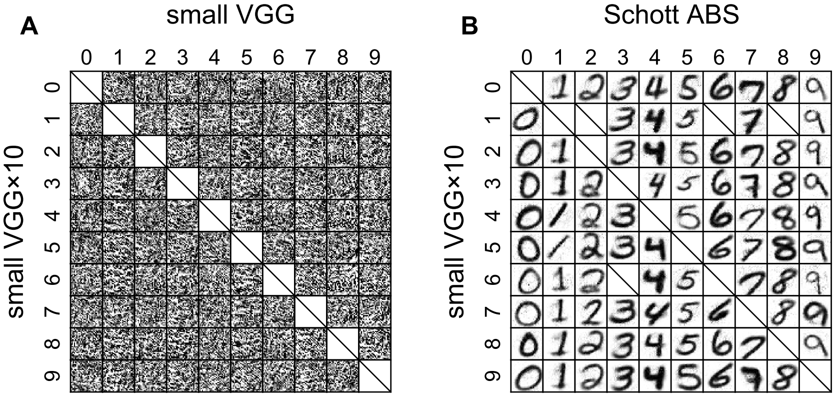

Fig. S3. A set of separate digit-specific discriminative DNNs do not match the set of digit-specific VAEs (the Schott ABS model) in human consistency. One potential alternative explanation of the advantage of the Schott ABS model is that the ABS model has one DNN per MNIST class, whereas the other candidate MNIST models we tested have a single DNN that learns to detect each of the ten classes. To test this alternative explanation, we trained a variant of the small VGG model where ten small VGG DNNs were independently trained on binary one-vs-all classification tasks (i.e., the first DNN is trained to detect the digit 0 vs. the other digits, the second DNN is trained to detect the digit 1 vs. the other digits, and so on). Other than this modification, the training procedure of each digit-specific DNN was identical to the small-VGG training procedure specified in SI subsection A.1. The scalar sigmoid outputs of the ten DNNs, each representing the detection of one particular digit, were concatenated to form a joint ten-unit output layer. The resulting model ('small VGG×10') had MNIST test error of 0.64%, slightly worse than the single DNN small VGG model (0.47%). We synthesized controversial stimuli targeting this model vs. the single DNN small VGG (panel A), as well as this model vs. the Schott ABS model (panel B). **As can be seen above, there is no qualitative indication that having one discriminative DNN per class improves the model's human consistency.** One can form controversial stimuli for the single-DNN small VGG model and the small VGG×10 model that do not resemble digits (panel A). Furthermore, the human recognition of the controversial stimuli for the small VGG×10 and the Schott ABS models is completely aligned with the Schott ABS model (panel B). In fact, we hypothesize that not sharing parameters between classes is a shortcoming rather than a strength of the Schott ABS model, since parameter sharing enables reusing visual features across different classes.

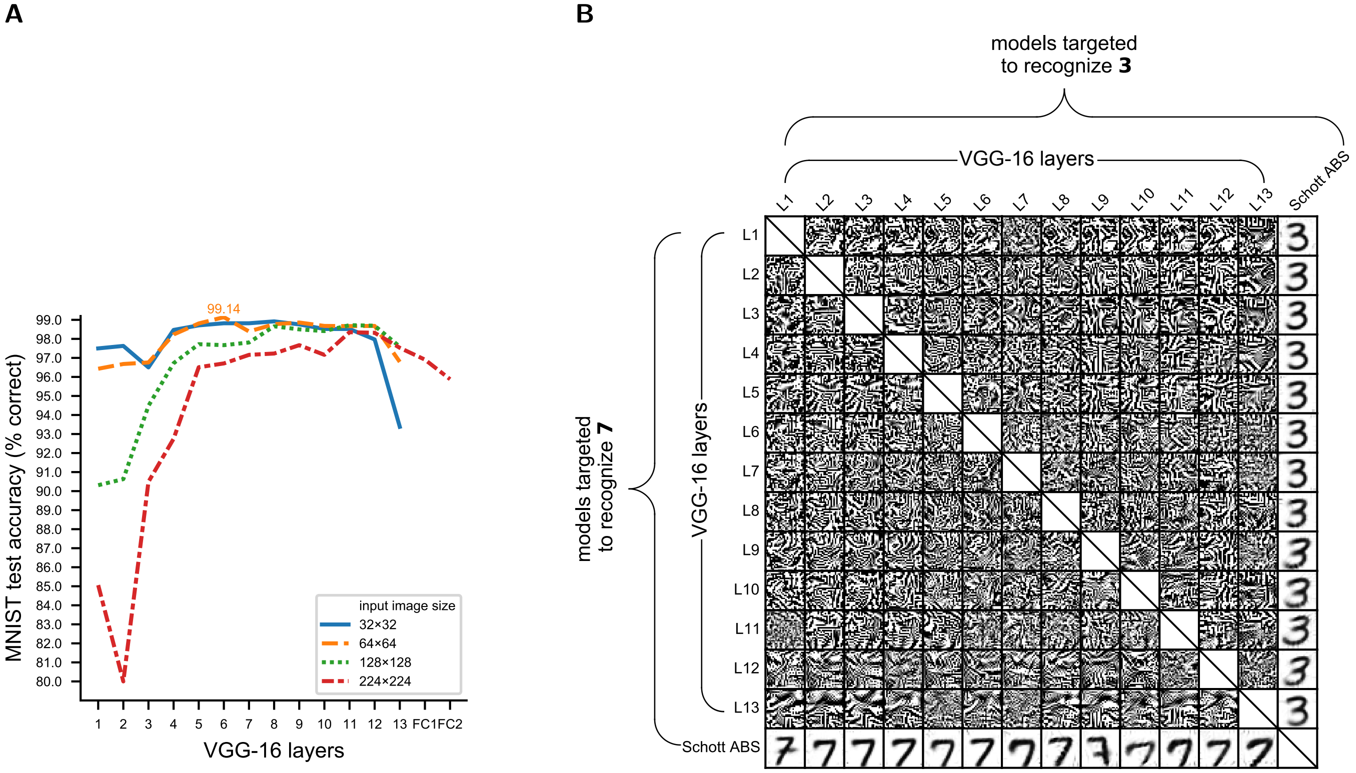

Fig. S4. MNIST linear classifiers reading out hidden-layer features of an ImageNet-trained VGG-16 do not exhibit human-consistent classification. Discriminative DNNs trained on object recognition in natural images (e.g., the ImageNet dataset) were found to have hidden-layer features that can serve as a surprisingly good basis for predicting neural responses in mid-level ventral visual stream (e.g., 24–29), as well as for predicting behavioral similarity judgments (e.g., 30, 31). These findings suggest an alternative explanation of the human-inconsistency of the discriminative models in Experiment 1 (e.g., the small VGG model): perhaps discriminative models can perform well in the MNIST controversial stimuli benchmark if they are first trained on recognizing objects in natural images, and only then learn to classify MNIST as a transfer learning task. After all, the 'visual diet' humans are exposed to from birth consists mostly of natural images. To test for this hypothesis, we trained multi-label linear classifiers on the MNIST task using the activations of the convolutional layers of an Imagenet-trained VGG-16 (5) as fixed features.

(A) **MNIST test accuracy of multi-label linear classifiers, each operating on the activations of one VGG-16 convolutional layer as a fixed feature set.** Each classifier is trained to classify MNIST digits from the activations of a single layer. Different lines represent classifiers with different degrees of input-image upscaling. Convolutional layers can be evaluated with input images of arbitrary size (as long as they are big enough to fit the kernels), so our rescaled input images were not embedded in blank margins. Fully connected layers (which do not share this flexibility) were tested only with the largest input-image scale (224×224 pixels), on which the model was originally trained. Following this accuracy comparison between different input-image scales and layers, we chose for further testing (panel B) the classifiers that use VGG-16's convolutional layers (i.e., layers 1-13) with a 64×64 pixels input-image size (orange dashed line in panel A).

(B) **Controversial stimuli (targeting '3' vs. '7' classification), pitting the linear classifiers trained on the features of the different VGG-16 convolutional layers against each other or against the Schott ABS model. The greater human-consistency of the Schott ABS model is evident, even though this model does not enjoy the advantage of using Imagenet-driven features.** Furthermore, controversial stimuli for pairs of classifiers based on different VGG-16 layers do not resemble human-recognizable digits.

While we cannot preclude that features gained from learning to recognize objects in natural images might be necessary for achieving human-consistent responses in the MNIST task, the controversial stimuli above indicate that in the discriminative-training context, using such features is an insufficient condition for model-human consistency.

Classifier training details: We used a pretrained VGG-16 model (Keras implementation) and prepended a bilinear upsampling operation to the model's normalized input, transforming MNIST images from the dimensionality of 28×28×1 to 32×32×3, 64×64×3, 128×128×3 or 224×224×3. We then trained an MNIST linear multi-label classifier separately for each layer using the layer's ReLU activations as fixed features. Linear classification was implemented as a new fully-connected layer projecting the ReLU activations of one particular hidden-layer to ten units. A sigmoid activation function was then applied to each of the ten units, rendering this setup equivalent to one-vs-all logistic regression. We adjusted the weights of the new fully-connected layer to minimize the cross-entropy of the sigmoids and a one-hot label representation using Adam (10) (Keras implementation). We used a learning rate of 0.001, reduced by a factor of 0.2 after every three epochs of no training loss reduction, down to a minimum learning rate of 10^{-6}. Early stopping was applied based on held-out validation data.

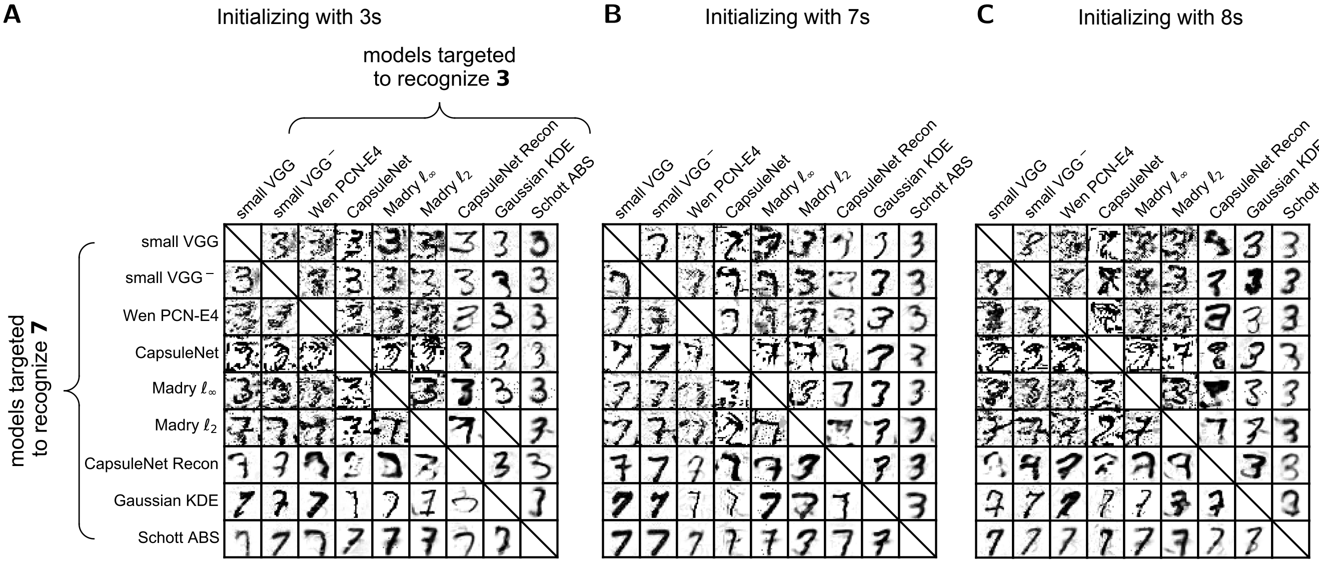

Fig. S5. Initializing controversial stimuli with digit images rather than with random-noise images does not change the observed hierarchy of human-consistency among the tested MNIST models. Each controversial stimulus was initialized with an MNIST test example randomly selected from either the 3 (A), 7 (B) or 8 (C) class. These images were then optimized using the same procedure employed in Experiment 1 to evoke a 3 (but not 7) detection in the models listed above each panel and at the same time, a 7 (but not 3) detection in the models listed to the left of each panel. **The greater alignment of the Schott ABS (4) model with human perception compared to the other candidate models is reproduced for each of the three different initializations: Most of the images in the rightmost column of each panel look like the digit 3, a judgment consistent with the target class of the ABS model for these images. Similarly, most of the images in the bottom row of each panel look like the digit 7, again a human perceptual judgment consistent with the target class of the ABS model and inconsistent with the other models.**

When pairs of standard discriminative DNNs are targeted, the initial image has a considerable effect on the resulting controversial stimulus (e.g., controversial stimuli in the top-left corner of each panel look like 3s, 7s, or 8s, reflecting their seed images). This sensitivity to the starting condition is caused by the non-robustness of these models. Since the classifications of these models can be easily changed by small perturbations, potential controversial stimuli densely populate the image space and hence the stimulus optimization does not need to travel far from the initial image to achieve high controversiality.

While this analysis is provided here as a control for the effect of initialization, we generally recommend initializing controversial stimuli with random-noise images. Controversial stimuli optimized from non-random initial images may inherit human-discernible features from the seed image. The resulting controversial stimuli may thus contain natural features, which are not driven by any of the targeted models (as in the top left part of each panel above, where standard discriminative models are paired). In contrast, when a controversial stimulus is initialized with a random noise image, any discernible feature is model-driven rather than experimenter-driven, eliminating the potential confounding factor of perceivable visual content originating from the initializing image. Furthermore, while there is no evidence that gradient obfuscation (32) plays a role in the optimization of the controversial stimuli above, initializing controversial stimuli with random images provides a strong protection against this issue. A model with obfuscated gradients will not be able to drive a random noise image towards a high-confidence detection of its target class. Such an outcome transparently indicates an optimization issue rather than give a false impression of model robustness.

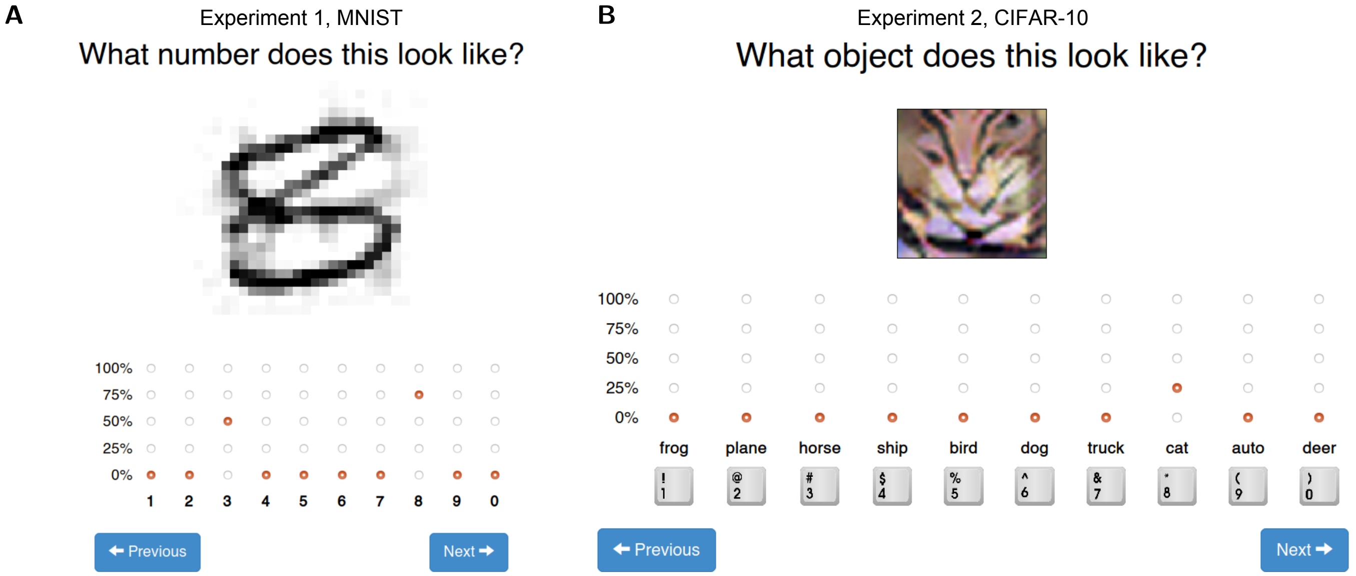

Fig. S6. Trials in the human online experiments. The subjects had to rate the presence of each class from 0% to 100%. These ratings do not need to sum to 100%. Image presentation order was randomized for each subject. The 'Previous' button enabled subjects to go back one trial to correct their responses. (A) Experiment 1 (comparing MNIST models). The images were upsampled using nearest-neighbor interpolation. (B) Experiment 2 (comparing CIFAR-10 models). The images were upsampled using Lanczos interpolation. The mapping from the CIFAR-10 categories to the ten numerical response keys was randomized for each subject.

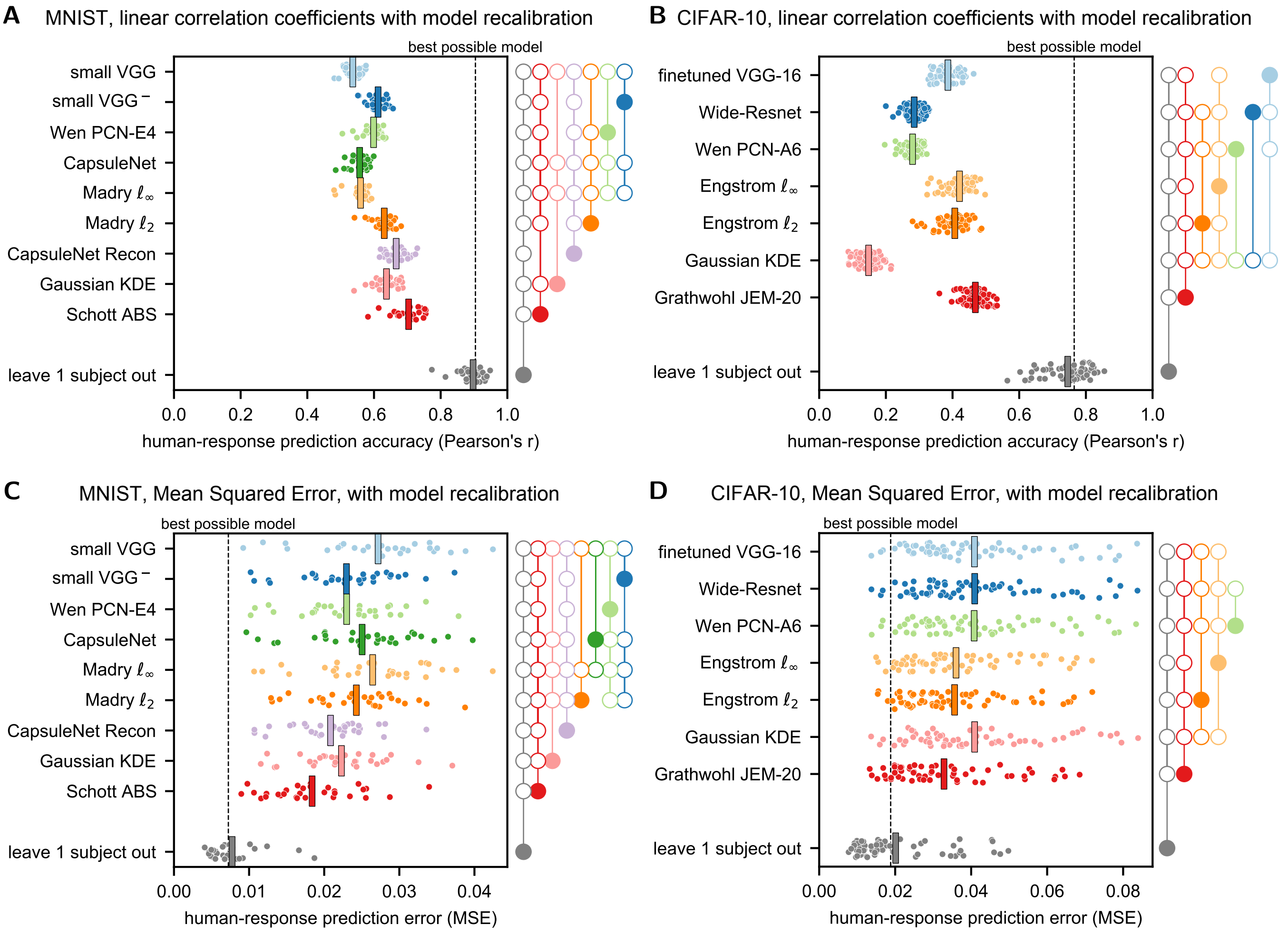

Fig. S7. Alternative measures of the model-human prediction accuracy show a similar rank ordering of models. (A) Linear correlation coefficients between each candidate MNIST models and the human responses to all of the test stimuli (Experiment 1), after recalibrating each model independently to maximize its mean-across-subjects human response prediction accuracy. The logits of each model (i.e., the inputs to the ten readout sigmoids, see main text) were recalibrated by a linear transformation. The slope and intercept of this transformation were adjusted to maximize each model's mean correlation coefficient. The slope and intercept parameters were unique to each model but did not vary across classes and subjects (i.e., there were exactly two free parameters per model). Note that this procedure may introduce a small optimistic bias to all of the correlation measures. (B) An equivalent analysis of the CIFAR-10 models' predictions (Experiment 2). (C) Measuring Mean Squared Error (MSE) instead of linear correlation for the MNIST candidate models. Model recalibration was applied here as well, fitting the slope and intercept of each model's transformed logits to minimize the grand-average MSE across subjects. Unlike the linear correlation measure, the MSE measure does not reduce individual differences in the baseline and scale of human behavioral ratings. Hence the greater dispersion of the subjects' MSEs compared to the subjects' correlation coefficients. (D) An equivalent analysis for the CIFAR-10 candidate models.
The MSE between a model M and subject S_i was calculated as $\frac{1}{|X||C|}\sum_{x,y}\left(\hat{p}_{s_i}(y \mid x)-\hat{p}_M(y \mid x)\right)^2$ where $\hat{p}_{s_i}(y \mid x)$ is the human-judged probability that image x contains class y, $\hat{p}_M(y \mid x)$ is the model's corresponding recalibrated judgment, $|X|$ is the number of stimuli (820 for Experiment 1, 480 for Experiment 2) and $|C|$ is the number of classes (ten for both experiments).

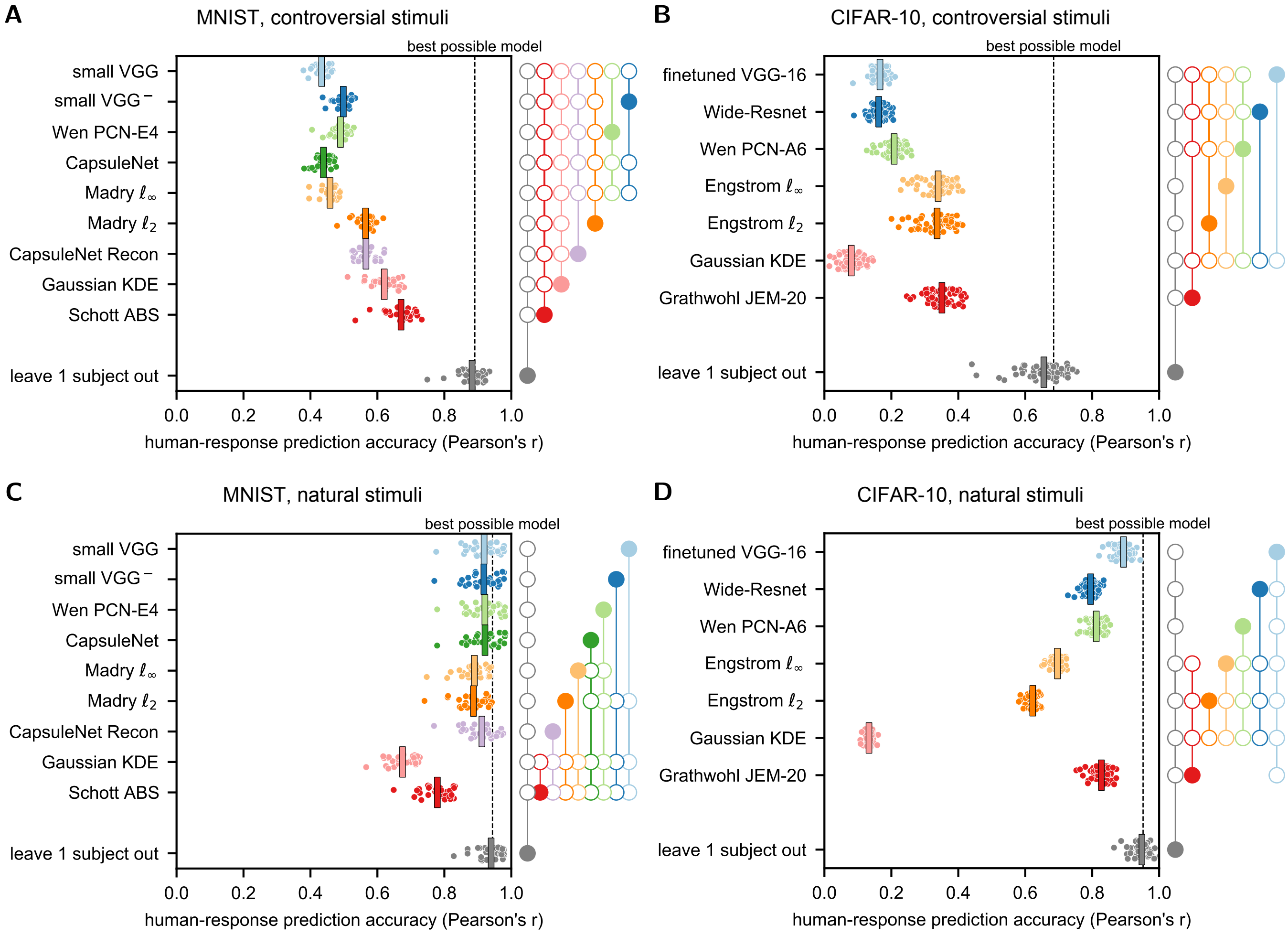

Fig. S8. The performance of the candidate models in predicting the human responses to controversial stimuli (panel A, Experiment 1; panel B, Experiment 2) and to natural stimuli (panel C, Experiment 1; panel D, Experiment 2). Unlike Fig. 4 (main text), where the model-human response correlation coefficients were calculated across all of the stimuli included in the behavioral experiments, here the correlation coefficients were calculated separately for controversial and natural stimuli. The Schott ABS model (4), which excels in the controversial stimuli benchmark (panel A), lags behind all of the discriminative MNIST models in predicting the responses to MNIST 'natural' test examples (panel C). In contrast, the hybrid discriminative-generative JEM-20 (33) performs on par of its discriminative counterpart ('Wide-Resnet', employing the same architecture) in predicting the human response to test CIFAR-10 images (panel D). At the same time, JEM-20 is at least as good as the adversarially trained models (Engstrom ℓ_∞ and Engstrom ℓ_2, (8)) in predicting the human responses to the controversial stimuli (panel B). This relatively high performance both within and outside the training distribution demonstrates the advantage of a hybrid discriminative-generative modeling approach. However, JEM-20 is still less accurate than the finetuned, ImageNet-trained VGG-16 model in predicting the human responses to natural (test) CIFAR-10 images (panel D).

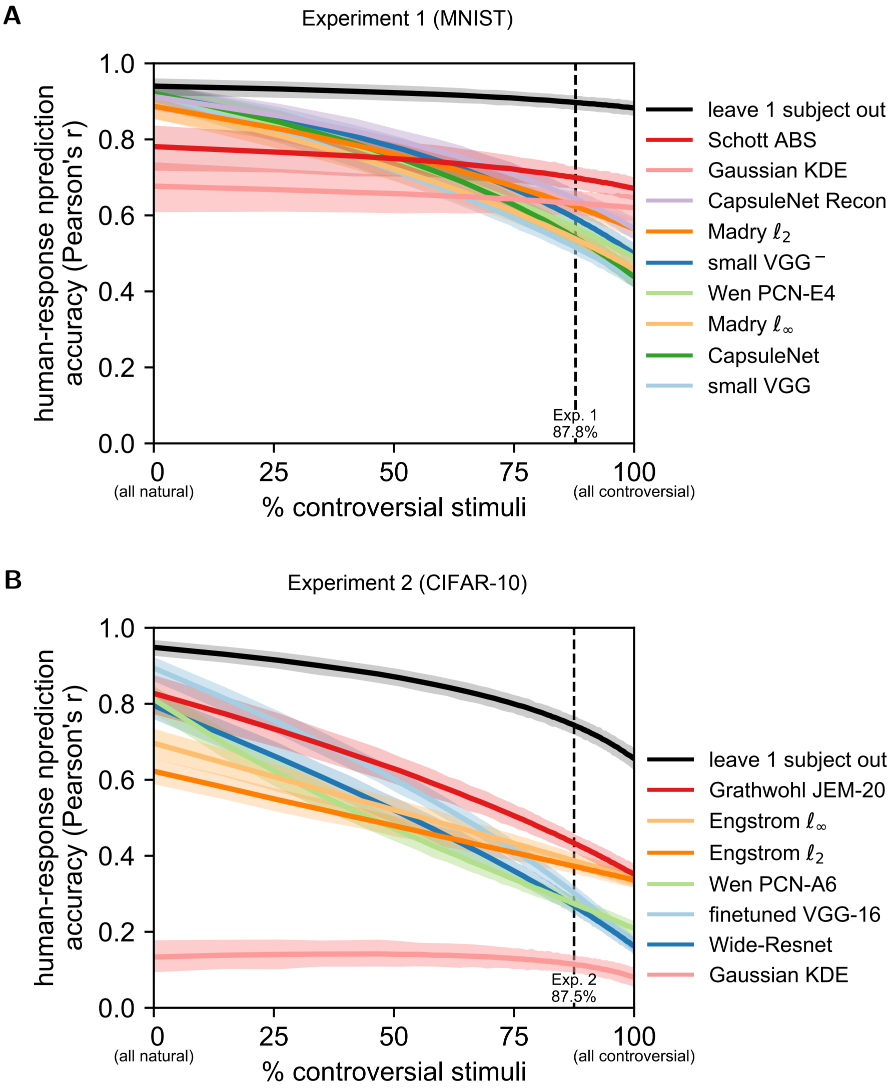

Fig. S9. The effect of the ratio of controversial to natural stimuli on the models' human-response prediction accuracy. The human-response prediction accuracy reported in Fig. 4 is measured across the entire stimulus set presented to the human subjects. Since this stimulus set included both natural and controversial stimuli, the reported prediction accuracy is affected by the proportion of controversial to natural stimuli within this set. Here, we resampled stimuli (and subjects) with replacement to simulate the hypothetical effect of different ratios of controversial to natural stimuli on Experiment 1 (MNIST models, panel A) and on Experiment 2 (CIFAR-10 models, panel B). 1000 datasets were resampled for each point along the x-axis (i.e., for each ratio).

The plots above depict the across-subjects-averaged human-response prediction accuracy of each model as a function of the percentage of controversial stimuli among the resampled stimulus set. The colored lines depict the bootstrap averages of this measure, and the semi-transparent shaded bends correspond to 95% percentile bootstrap confidence intervals. The dashed vertical lines indicate the original ratio of controversial to natural stimuli (i.e., as in Fig. 4). The prediction accuracy to the right of the dashed vertical lines was obtained by gradually decreasing the number of resampled natural stimuli, down to no natural stimuli at all (right end, as in Figs. S8A, B). The prediction accuracy to the left of the dashed lines was obtained by gradually decreasing the number of resampled controversial stimuli, down to no controversial stimuli at all (left end, as in Figs. S8C, D). The resampling of controversial stimuli was stratified by subsets of controversial stimuli targeting different model-pairs.

In Experiment 1 (panel A), the purely generative ABS model beats the discriminative MNIST models only when the controversial stimuli consist of the vast majority of the test stimuli. In contrast, in Experiment 2 (panel B) the hybrid discriminative-generative JEM-20 model beats the other CIFAR-10 models across most of the proportion range.

Note that an ideal model should weakly dominate all other models for any stimulus set; as long as we do not have such an ideal model, the model seen as 'best' depends on our particular sampling of the stimulus space. Considering both natural examples and synthetic controversial stimuli provides a wider perspective on model-human consistency. In particular, synthetic controversial stimuli expose regions in stimulus space where the predictions of some of the models must be wrong. This is demonstrated here by the gap between the noise ceiling (in gray) and the models' prediction accuracy; this gap increases as the controversial stimuli are gradually introduced into the stimulus set.

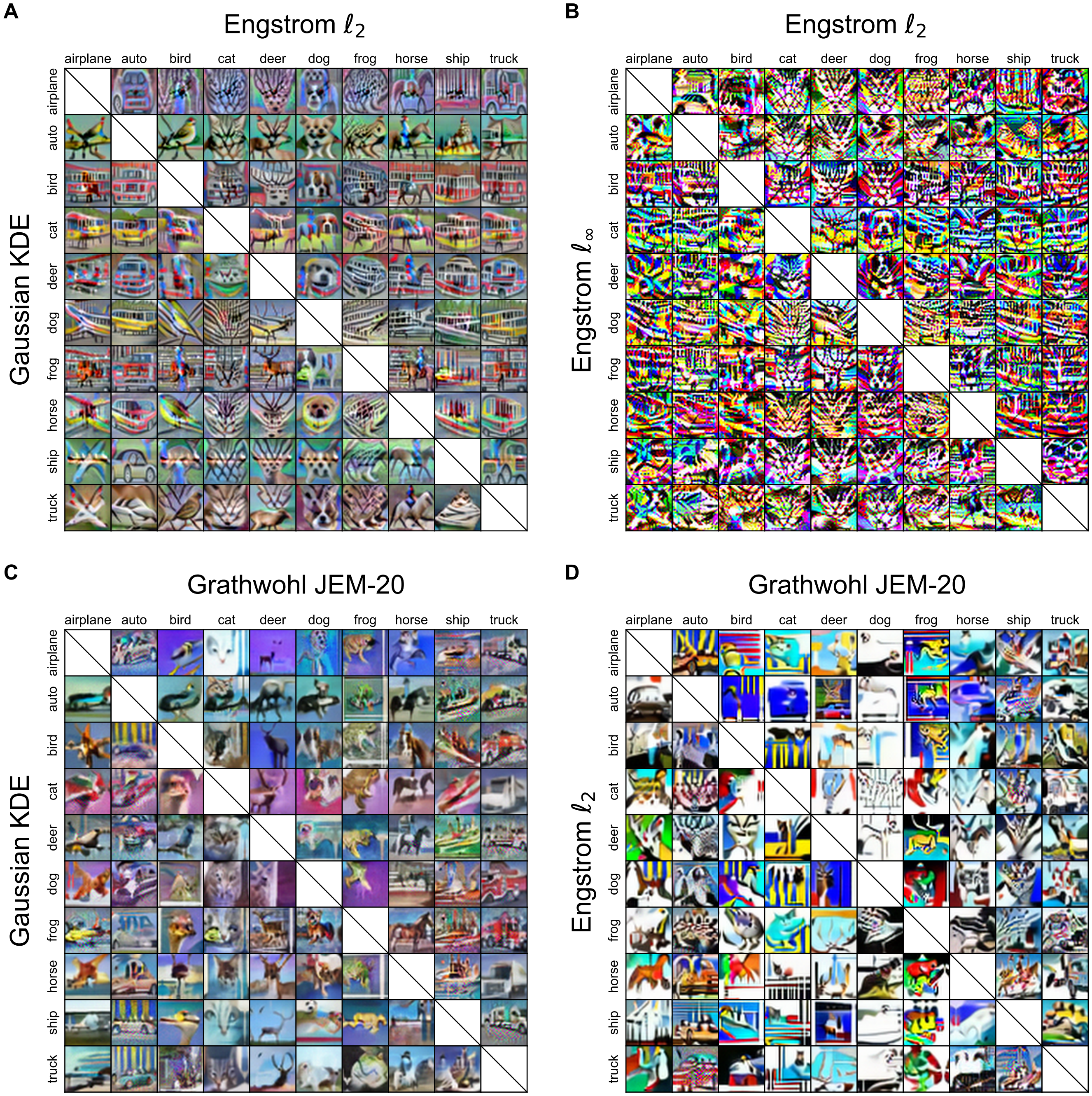

Fig. S10. Controversial stimuli for CIFAR-10 models, here organized to show the results of targeting each possible CIFAR class pair for four different model pairs. The rows and columns within each panel indicate the targeted classes. For example, the top-right image in panel C was optimized to be detected as a truck by the JEM-20 model and as an airplane by the Gaussian KDE model. Missing (crossed) cells are along the diagonal, where the two models would agree. All images here have a controversiality score of at least 0.75. See Fig. S11 for all 21 model pairs.

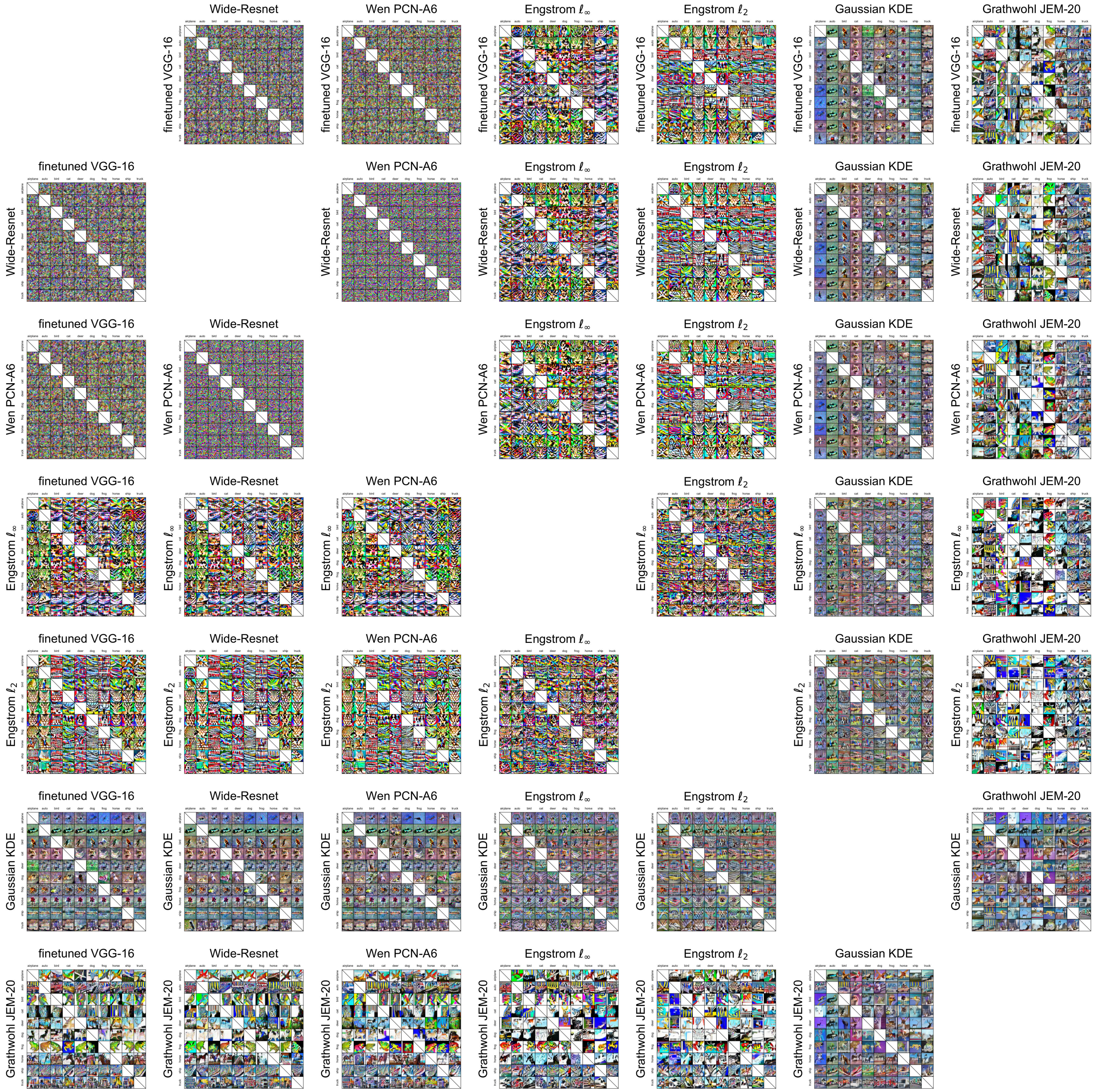

Fig. S11. The entire set of controversial stimuli we synthesized for Experiment 2 (CIFAR-10), organized by model pairs. Each panel indicates a targeted model pair, and the rows and columns within each panel indicate the targeted class pairs. For example, consider the bottom-left image in the bottom-left panel. This image is classified as an airplane by the finetuned VGG-16 model and as a truck by the Grathwohl JEM-20 model. Missing (crossed) cells are either along the diagonal (where the two models would agree) or where our optimization procedure did not converge to a sufficiently controversial image (a controversiality score of at least 0.75). Best viewed digitally.

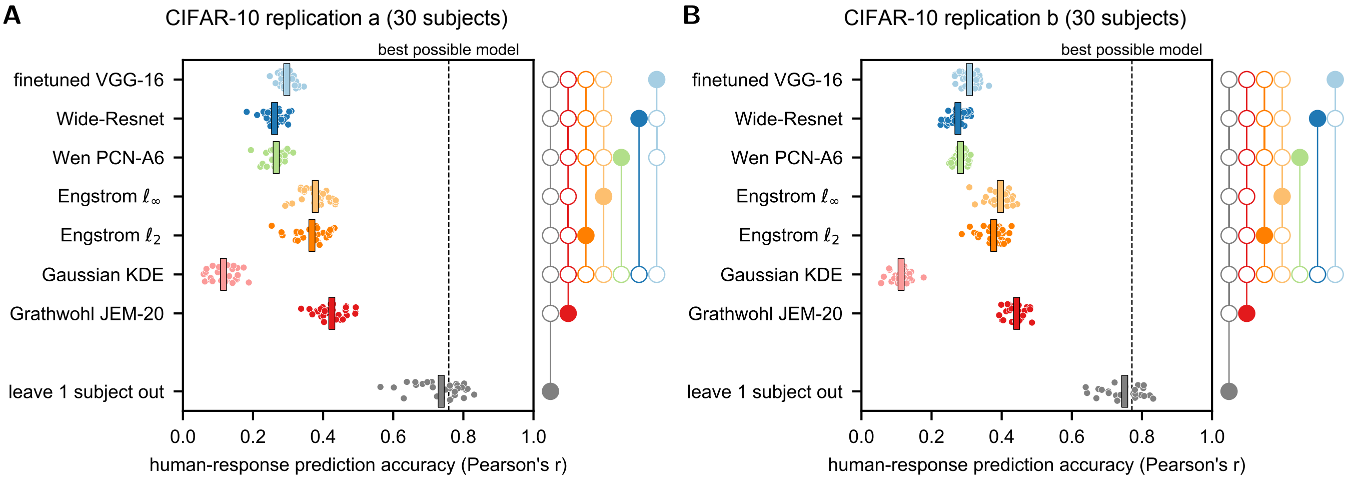

Fig. S12. A side by side comparison of two replications of Experiment 2. Each replication was conducted with an independently synthesized set of 420 controversial stimuli and an independently selected set of 60 CIFAR-10 test examples. 30 subjects participated in each replication and no subject participated in both of them. Main text Fig. 4B shows the same data pooled over the two replications. **The similarity of the outcomes of the two replications demonstrates the statistical reliability of our experimental procedure as a whole.**

Table S1. Candidate MNIST models

			human-response prediction accuracy		
model family	model	MNIST test error	all stimuli	controversial	natural
discriminative feedforward	small VGG (5)*	0.47%	0.516	0.434	0.919
	small VGG^{-} (5)*	0.59%	0.592	0.498	0.918
discriminative recurrent	Wen PCN-E4 (1)	0.42%	0.567	0.490	0.921
	CapsuleNet (2)	0.24%	0.541	0.439	0.921
adversarially trained	Madry ℓ_∞ (3) $(\epsilon = 0.3)$	1.47%	0.538	0.459	0.890
	Madry ℓ_2 (3) $(\epsilon = 2.0)$	1.07%	0.623	0.565	0.887
reconstruction-based	CapsuleNet Recon (7)*	0.29%	0.643	0.566	0.912
generative	Gaussian KDE	3.21%	0.632	0.621	0.675
	Schott ABS (4)	1.00%	0.699	0.671	0.779

MNIST models tested in Experiment 1. 'Human-response prediction accuracy' is the across-subject average of the correlation between model and human judgments ($\bar{r}_M$). The values listed here correspond to the means depicted by vertical bars in Figs. 4A, S8A, and S8C.

$*$ A modified architecture. See SI section A.

Table S2. Canidate CIFAR-10 models

			human-response prediction accuracy		
model family	model	CIFAR-10 test error	all stimuli	controversial	natural
discriminative feedforward	finetuned VGG-16 (5)*	4.37%	0.302	0.166	0.894
	Wide-Resnet (33)	6.22%	0.268	0.162	0.795
discriminative recurrent	Wen PCN-A6 (1)	6.66%	0.274	0.208	0.812
adversarially trained	Engstrom ℓ_∞ (8) $(\epsilon = 8/255)$	12.97%	0.387	0.340	0.696
	Engstrom ℓ_2 (8) $(\epsilon = 1.0)$	18.38%	0.373	0.337	0.622
generative	Gaussian KDE	61.21%	0.114	0.080	0.133
hybrid discriminative-generative	Grathwohl JEM-20 (33)	9.71%	0.434	0.351	0.827

CIFAR-10 models tested in Experiment 2. 'Human-response prediction accuracy' is the across-subject average of the correlation between model and human judgments ($\bar{r}_M$). The values listed here correspond to the means depicted by vertical bars in Figs. 4B, S8B, and S8D.

$*$ A modified architecture. See SI subsection B.1.

† Both models share the Wide-Resnet WRN-28-10 architecture (34), trained without batch normalization, and differ in their training and inference procedures (see SI subsection B.4).

References

1. Wen H, et al. (2018) Deep Predictive Coding Network for Object Recognition in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, eds. Dy J, Krause A. (PMLR, Stockholmsmässan, Stockholm Sweden), Vol. 80, pp. 5266–5275.
2. Sabour S, Frosst N, Hinton GE (2017) Dynamic Routing Between Capsules in *Advances in Neural Information Processing Systems 30*, eds. Guyon I, et al. (Curran Associates, Inc.), pp. 3856–3866.
3. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards Deep Learning Models Resistant to Adversarial Attacks in *International Conference on Learning Representations*.
4. Schott L, Rauber J, Bethge M, Brendel W (2019) Towards the first adversarially robust neural network model on MNIST in *International Conference on Learning Representations*.
5. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
6. Frosst N, Sabour S, Hinton G (2018) DARCCC: Detecting Adversaries by Reconstruction from Class Conditional Capsules. *arXiv:1811.06969 [cs, stat]*. arXiv: 1811.06969.
7. Qin Y, et al. (2020) Detecting and Diagnosing Adversarial Images with Class-Conditional Capsule Reconstructions in *International Conference on Learning Representations*.
8. Engstrom L, Ilyas A, Santurkar S, Tsipras D (2019) Robustness (Python Library).
9. Marcel S, Rodriguez Y (2010) Torchvision the machine-vision package of torch in *Proceedings of the 18th ACM international conference on Multimedia*, MM '10. (Association for Computing Machinery, Firenze, Italy), pp. 1485–1488.
10. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
11. Nili H, et al. (2014) A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology* 10(4):1–11.
12. Martín Abadi, et al. (2015) *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.
13. Chollet F, others (2015) *Keras*.
14. Paszke A, et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library in *Advances in Neural Information Processing Systems 32*, eds. Wallach H, et al. (Curran Associates, Inc.), pp. 8024–8035.
15. McDonnell J, et al. (2012) *psiTurk (Version 1.02) [Software]*. (New York University, New York, NY).
16. Walt Svd, Colbert SC, Varoquaux G (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* 13(2):22–30.
17. Hoyer S, Hamman J (2017) xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software* 5(1).
18. McKinney W (2010) Data Structures for Statistical Computing in Python in *Proceedings of the 9th Python in Science Conference*, eds. Walt Svd, Millman J. pp. 56 – 61.
19. Pedregosa F, et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
20. Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python in *9th Python in Science Conference*.
21. Hunter JD (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9(3):90–95.
22. Waskom M, et al. (2014) Seaborn: statistical data visualization. *URL: https://seaborn. pydata. org/(visited on 2017-05-15)*.
23. Cohen G, Afshar S, Tapson J, van Schaik A (2017) EMNIST: an extension of MNIST to handwritten letters. *arXiv:1702.05373 [cs]*. arXiv: 1702.05373.
24. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111(23):8619–8624.
25. Khaligh-Razavi SM, Kriegeskorte N (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology* 10(11):1–29.
26. Güçlü U, van Gerven MAJ (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of Neuroscience* 35(27):10005–10014.
27. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports* 6:27755.
28. Eickenberg M, Gramfort A, Varoquaux G, Thirion B (2017) Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 152:184 – 194.
29. Wen H, et al. (2017) Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex* 28(12):4136–4160.
30. Peterson JC, Abbott JT, Griffiths TL (2018) Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cognitive Science* 42(8):2648–2669.
31. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 586–595.
32. Athalye A, Carlini N, Wagner D (2018) Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, eds. Dy J, Krause A. (PMLR, Stockholmsmässan, Stockholm Sweden), Vol. 80, pp. 274–283.
33. Grathwohl W, et al. (2019) Your classifier is secretly an energy based model and you should treat it like one in *International Conference on Learning Representations*.
34. Zagoruyko S, Komodakis N (2016) Wide residual networks. *arXiv preprint arXiv:1605.07146*.