



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

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 ($\hat{p}(y | x)$) 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/Sarafra/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, $\hat{p}(x | y) = \frac{1}{n\sigma_y} \sum_{i=1}^n K\left(\frac{x - x_i^y}{\sigma_y}\right)$ 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 $[10^{-2}, \dots, 10^0]$ (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 $[10^{-2}, \dots, 10^2]$ (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 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 $[2^{-1}, 2^{-2}, \dots, 2^{-8}]$ 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 $32 \times 32 \times 3$ 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 initialized 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 $\hat{p}_{s_i}(y | x)$ 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 $\hat{q}_{s_i}(y | x) = \frac{1}{n-1} \sum_{j \neq i} \hat{p}_{s_j}(y | x)$. 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, \mathbf{p}_{s_i} and \mathbf{q}_{s_i} . These vectors would have 8200 elements for the responses of an Experiment 1 subject (820 images \times 10 response categories) or 4800 elements for the responses of an Experiment 2 subject (480 images \times 10 response categories). The linear correlation between \mathbf{p}_{s_i} and \mathbf{q}_{s_i} measures how well can the subject's response pattern be predicted from the mean response pattern of her/his peers:

$$r(S, s_i) = \frac{\sum_{x,y} (\hat{p}_{s_i}(y | x) - \bar{\hat{p}}_{s_i}) (\hat{q}_{s_i}(y | x) - \bar{\hat{q}}_{s_i})}{\sqrt{\sum_{x,y} (\hat{p}_{s_i}(y | x) - \bar{\hat{p}}_{s_i})^2} \sqrt{\sum_{x,y} (\hat{q}_{s_i}(y | x) - \bar{\hat{q}}_{s_i})^2}}, \quad [\text{S.1}]$$

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 \mathbf{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.

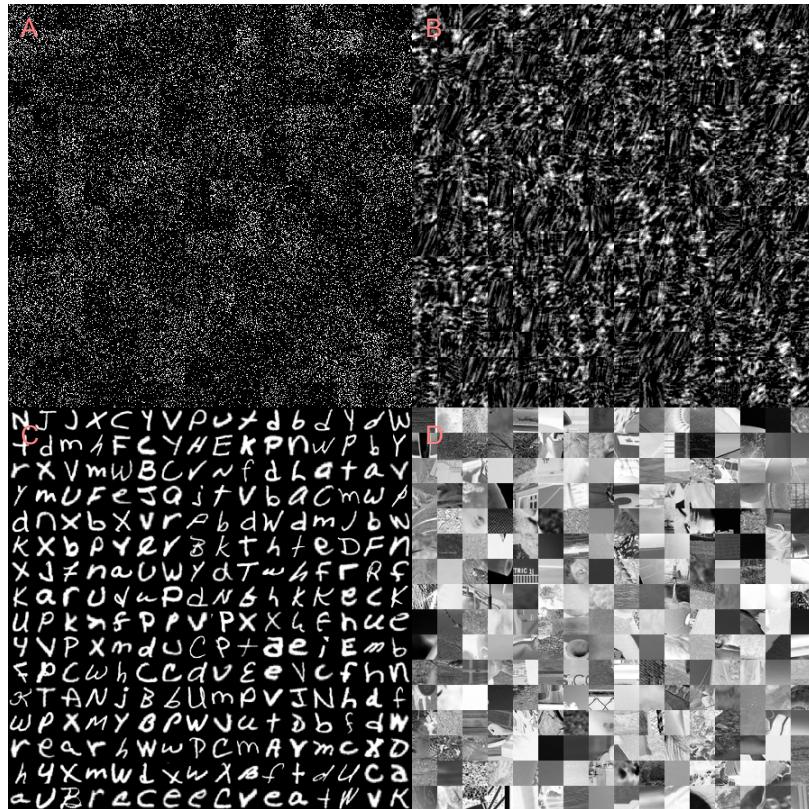


Fig. S1. A random sample of the negative examples (non-digits) used as a background class for the small VGG[−] model. (A) pixel-scrambled MNIST images. (B) Fourier-phase scrambled MNIST-images. (C) EMNIST letters (23), excluding the letters o,s,z,l,i,q and g. (D) patches cropped from natural images. 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.

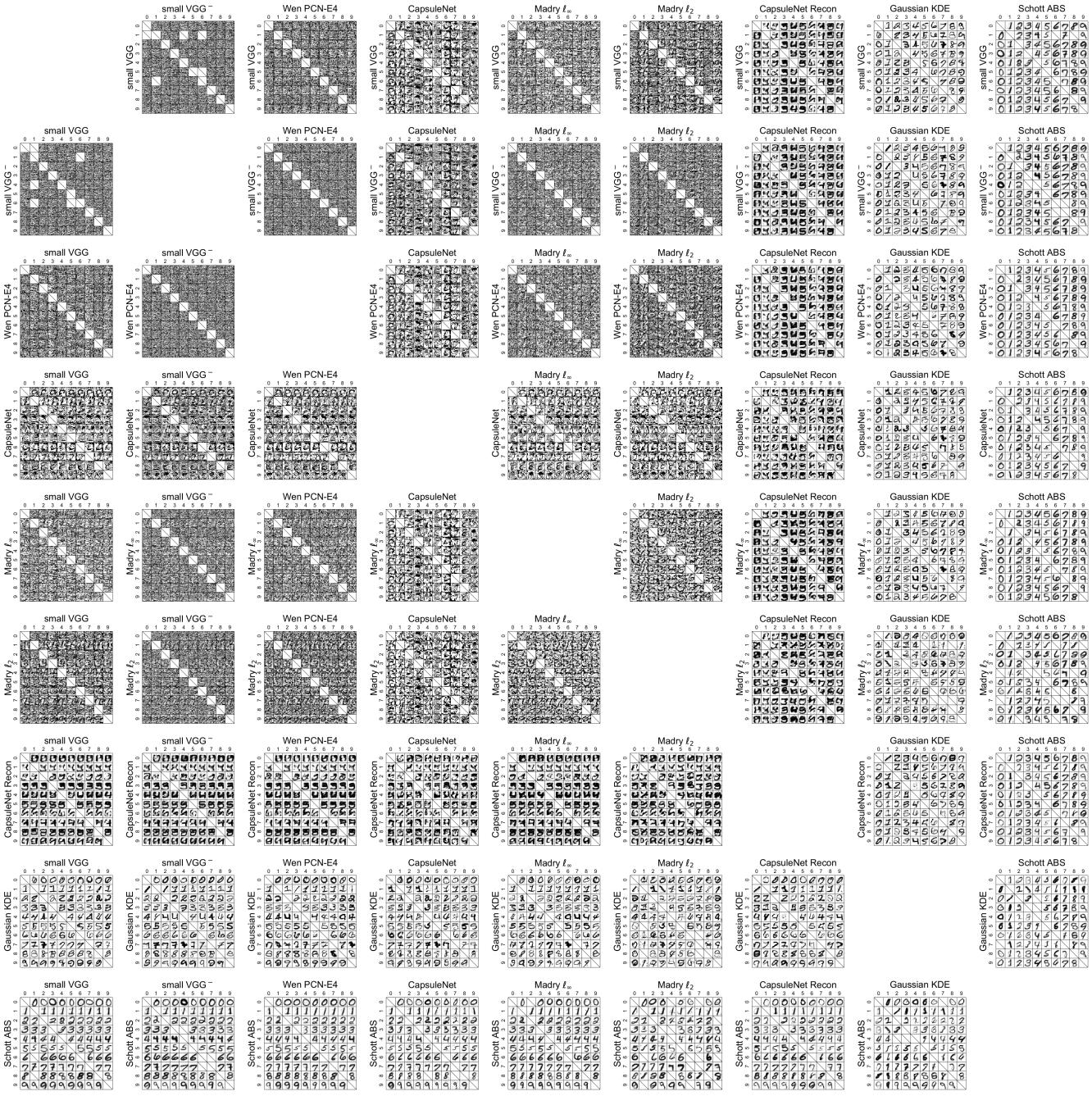


Fig. S2. The entire set of controversial stimuli we synthesized for Experiment 1 (MNIST), organized by model pairs. Each panel indicates a targeted model pair and the rows and columns within each panel indicate the targeted class pairs. For example, consider the bottom-left image in the bottom-left panel. 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.

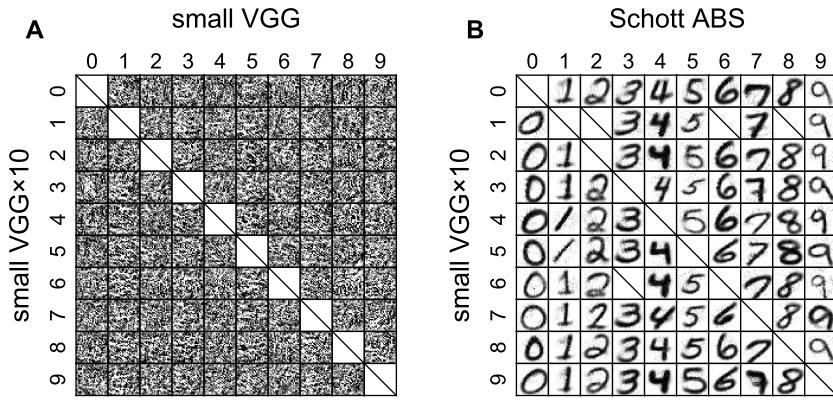


Fig. S3. A set of separate digit-specific discriminative DNNs do not match the set of digit-specific VAEs (the Schott ABS model) in human consistency. One potential alternative explanation of the advantage of the Schott ABS model is that the ABS model has one DNN per MNIST class, whereas the other candidate 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 \times 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 \times 10 model that do not resemble digits (panel A). Furthermore, the human recognition of the controversial stimuli for the small VGG \times 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.

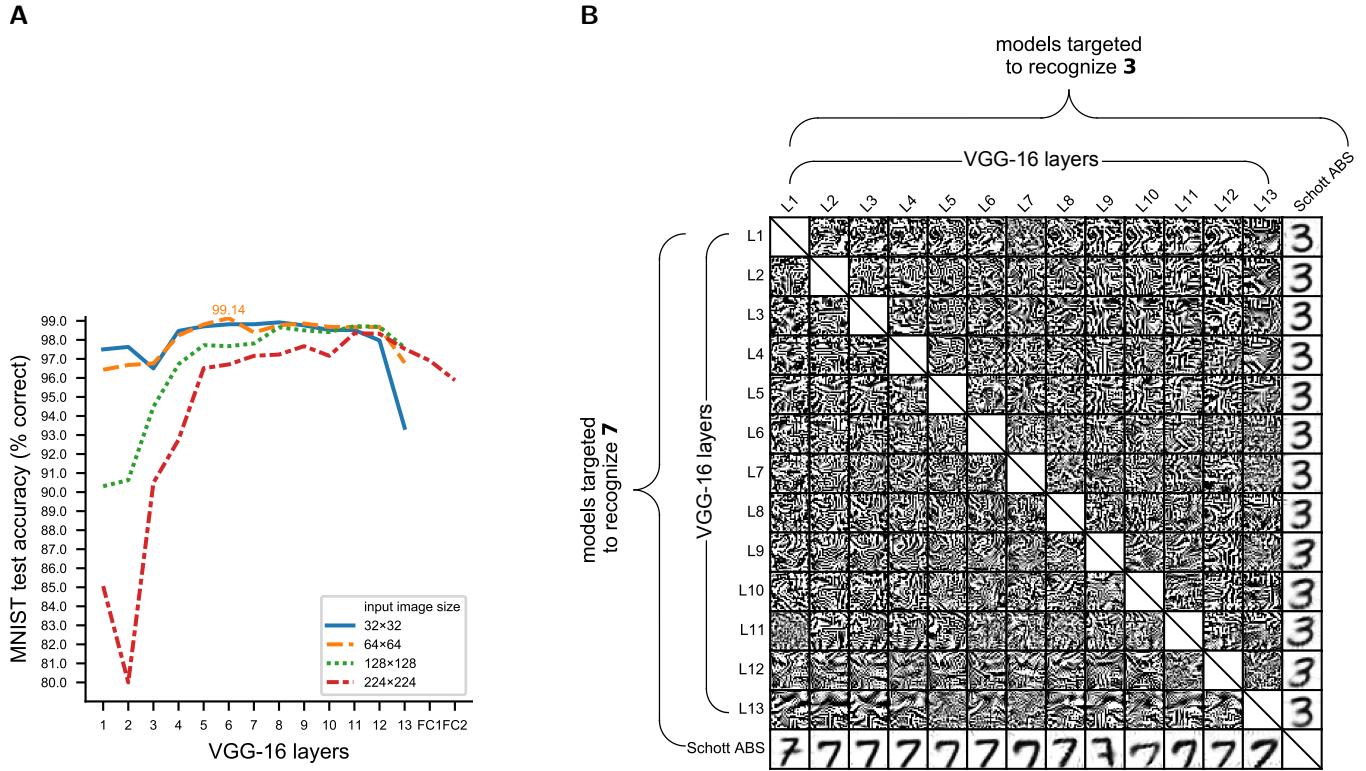


Fig. S4. MNIST linear classifiers reading out hidden-layer features of an ImageNet-trained VGG-16 do not exhibit human-consistent classification. Discriminative DNNs trained on object recognition in natural images (e.g., the ImageNet dataset) were found to have hidden-layer features that can serve as 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.

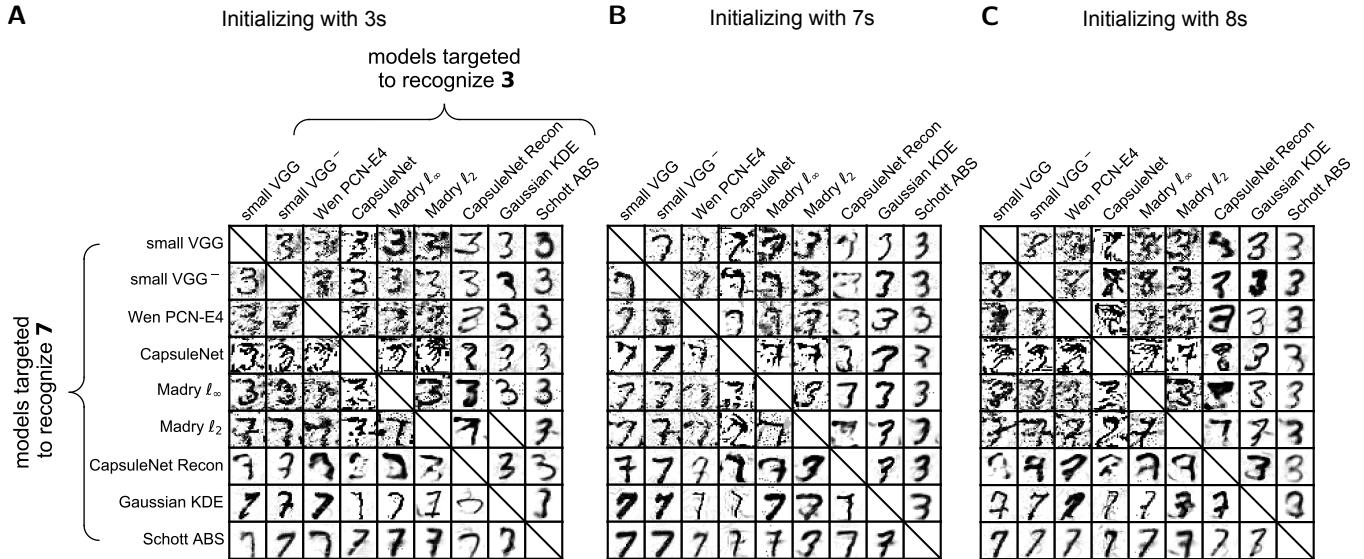


Fig. S5. Initializing controversial stimuli with digit images rather than with random-noise images does not change the observed hierarchy of human-consistency among the tested MNIST models. Each controversial stimulus was initialized with an MNIST test example randomly selected from either the 3 (A), 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.

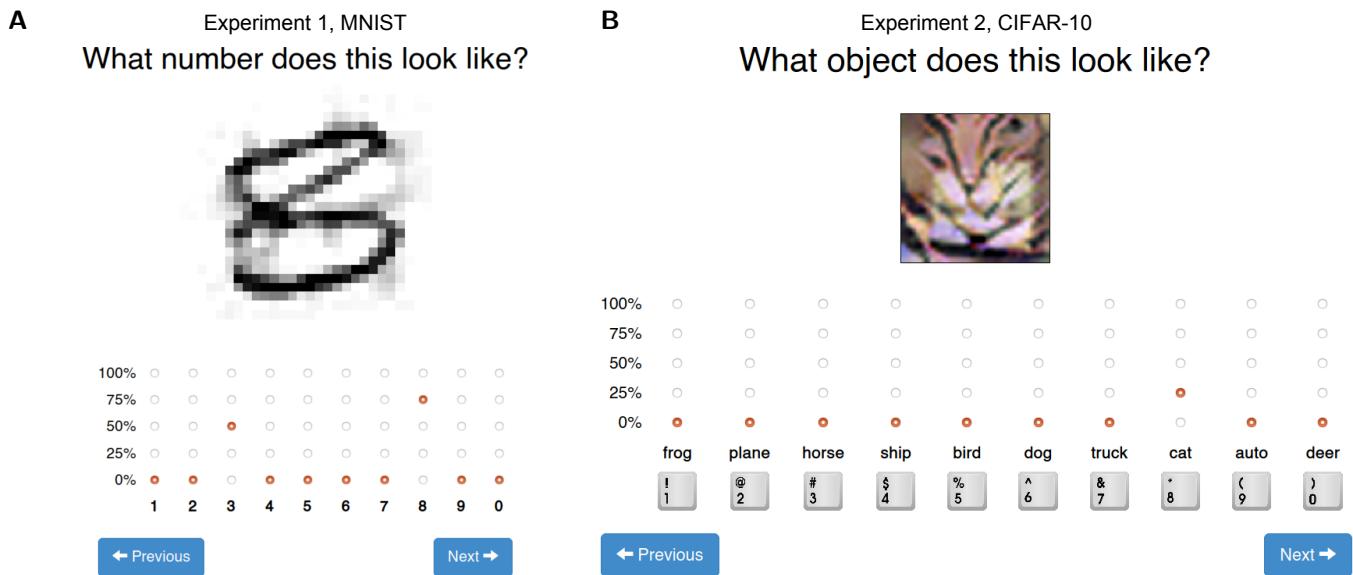


Fig. S6. Trials in the human online experiments. The subjects had to rate the presence of each class from 0% to 100%. These ratings do not need to sum to 100%. Image presentation order was randomized for each subject. The 'Previous' button enabled subjects to go back one trial to correct their responses. (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.

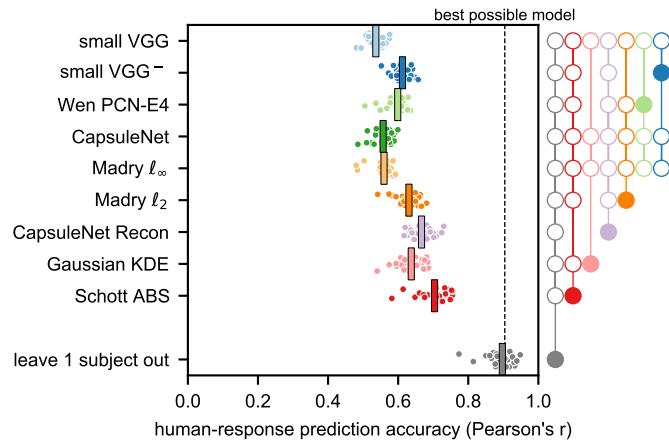
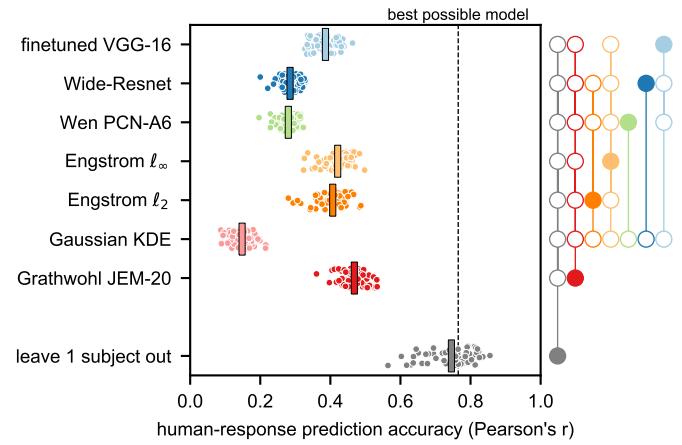
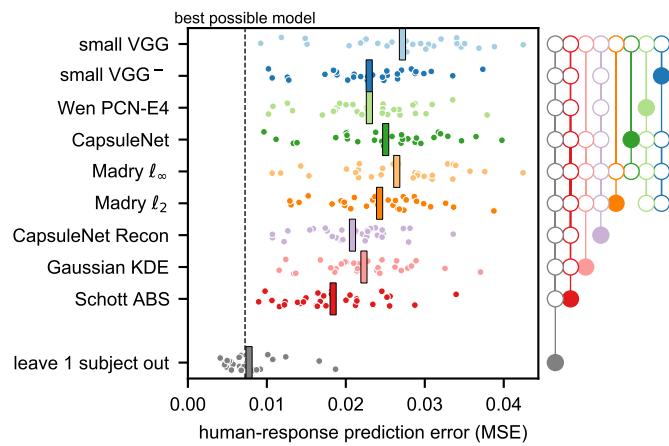
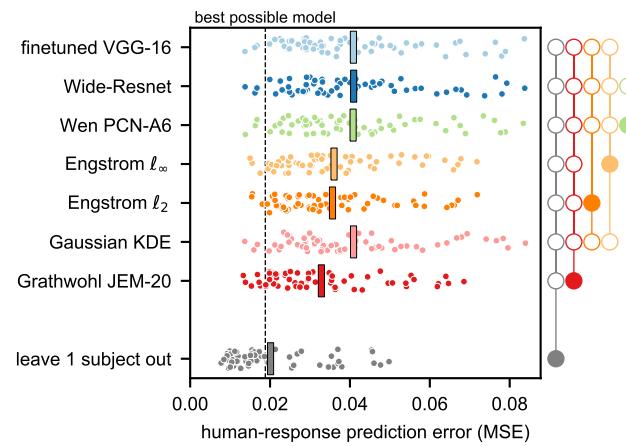
A MNIST, linear correlation coefficients with model recalibration**B** CIFAR-10, linear correlation coefficients with model recalibration**C** MNIST, Mean Squared Error, with model recalibration**D** CIFAR-10, Mean Squared Error, with model recalibration

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 $\frac{1}{|X||C|} \sum_{x,y} (\hat{p}_{S_i}(y | x) - \hat{p}_M(y | x))^2$ 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 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).

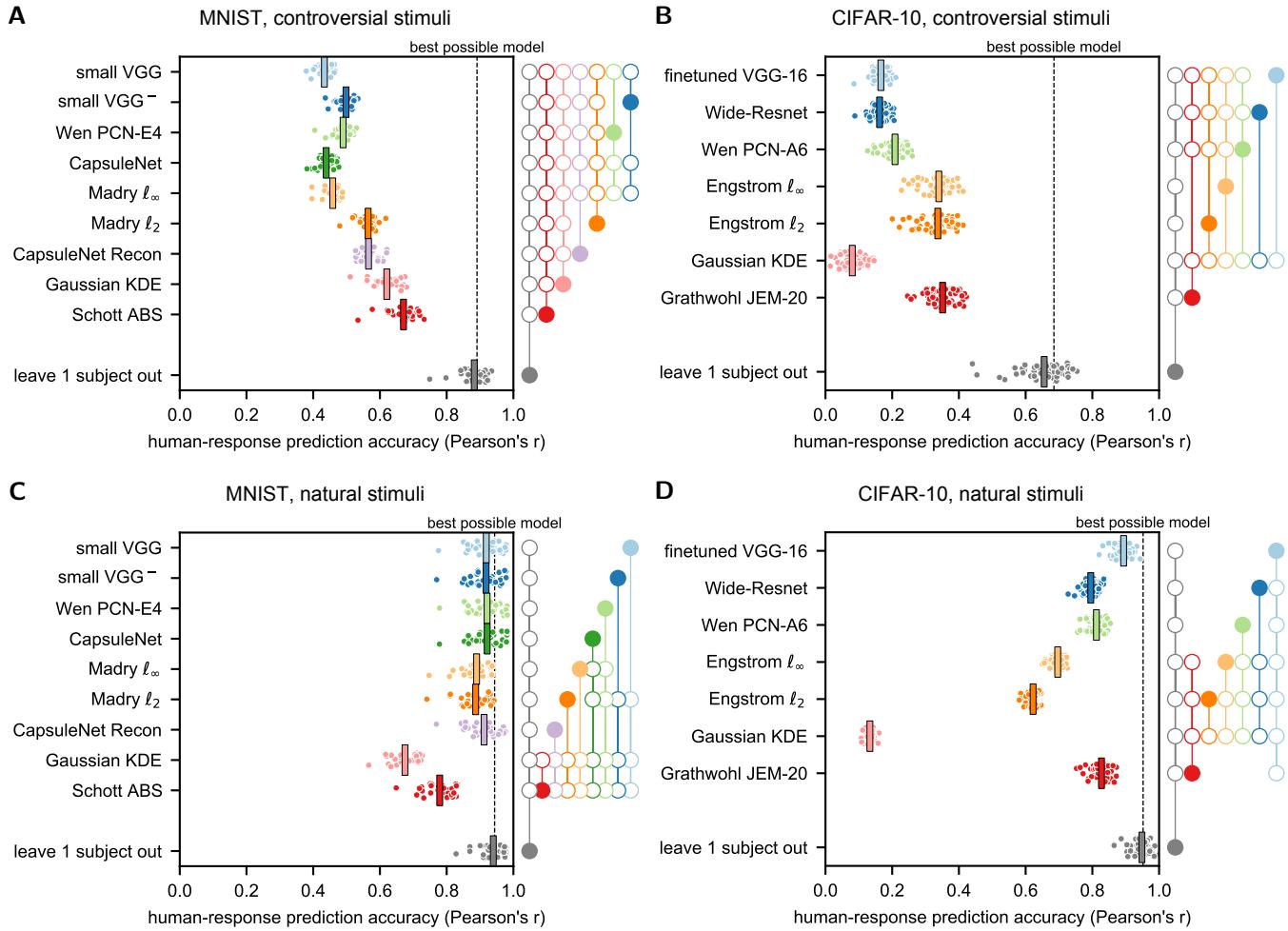


Fig. S8. The performance of the candidate models in predicting the human responses to controversial stimuli (panel A, Experiment 1; panel B, Experiment 2) and to natural stimuli (panel C, Experiment 1; panel D, Experiment 2). Unlike Fig. 4 (main text), where the model-human response correlation coefficients 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).

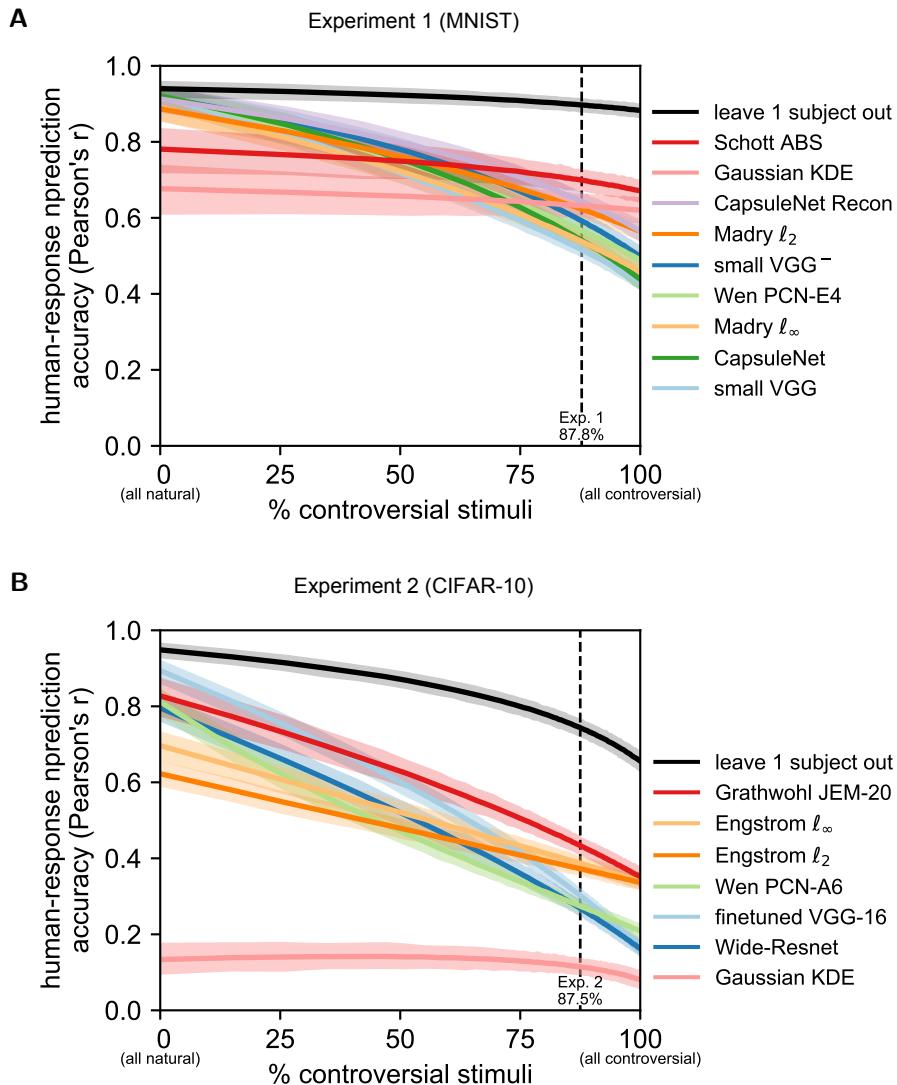


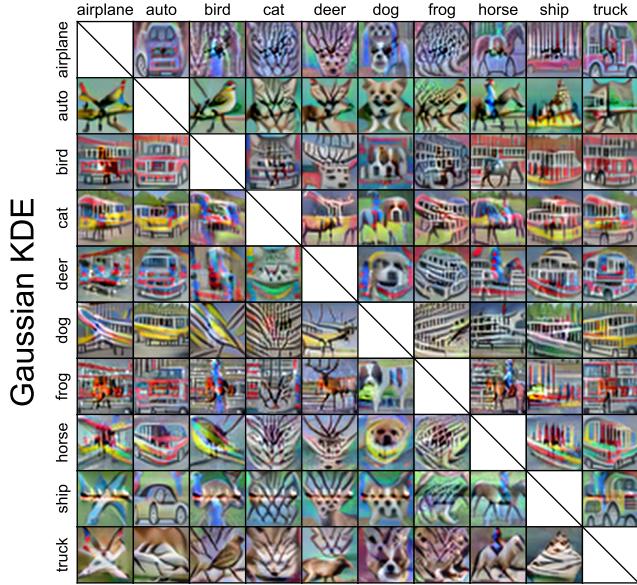
Fig. S9. The effect of the ratio of controversial to natural stimuli on the models' human-response prediction accuracy. The human-response prediction accuracy reported in Fig. 4 is measured across the entire stimulus set presented to the human subjects. Since this stimulus set included both natural and 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 bands 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.

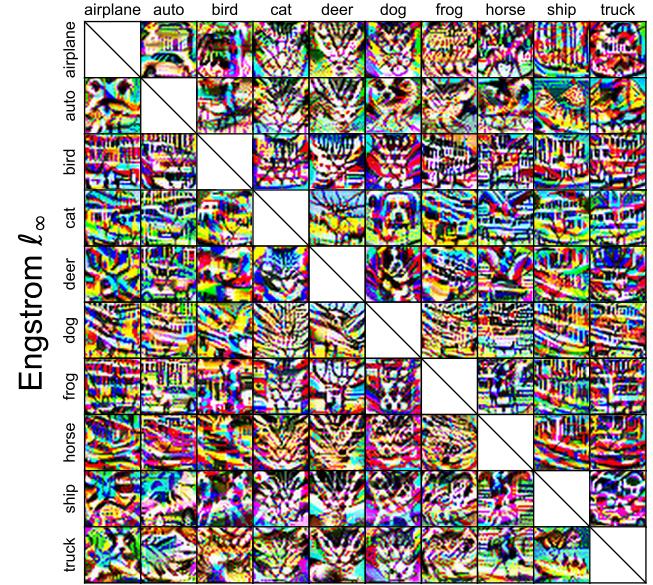
In Experiment 1 (panel A), the purely generative ABS model beats the discriminative MNIST models only when the controversial stimuli consist of the vast majority of the test stimuli. In contrast, in Experiment 2 (panel B) the hybrid discriminative-generative JEM-20 model beats the other CIFAR-10 models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

A

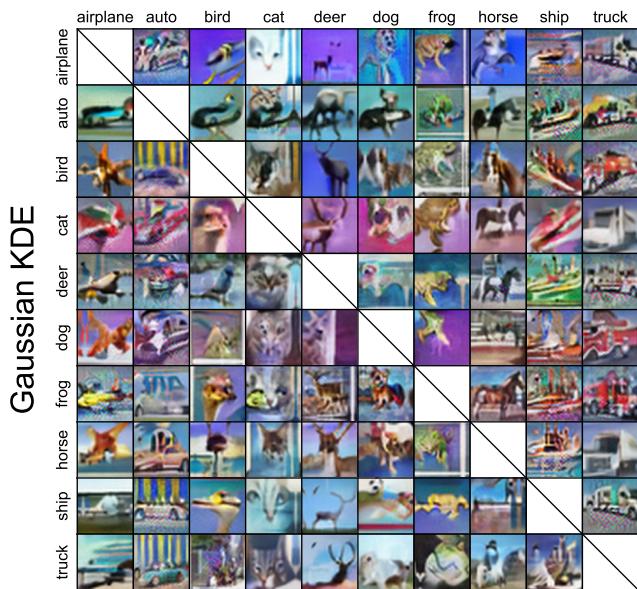
Engstrom ℓ_2 

B

Engstrom ℓ_2 

C

Grathwohl JEM-20



D

Grathwohl JEM-20

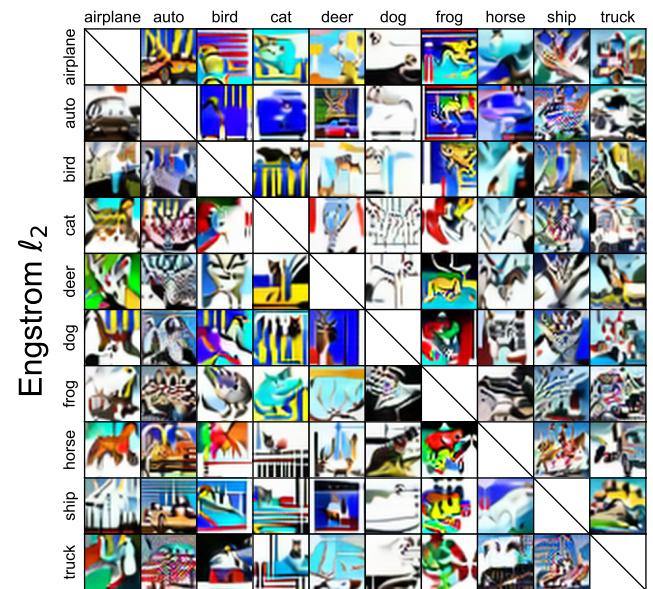


Fig. S10. Controversial stimuli for CIFAR-10 models, here organized to show the results of targeting each possible CIFAR class pair for four different model pairs. The rows and columns within each panel indicate the targeted classes. For example, the top-right image in panel C was optimized to be detected 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.

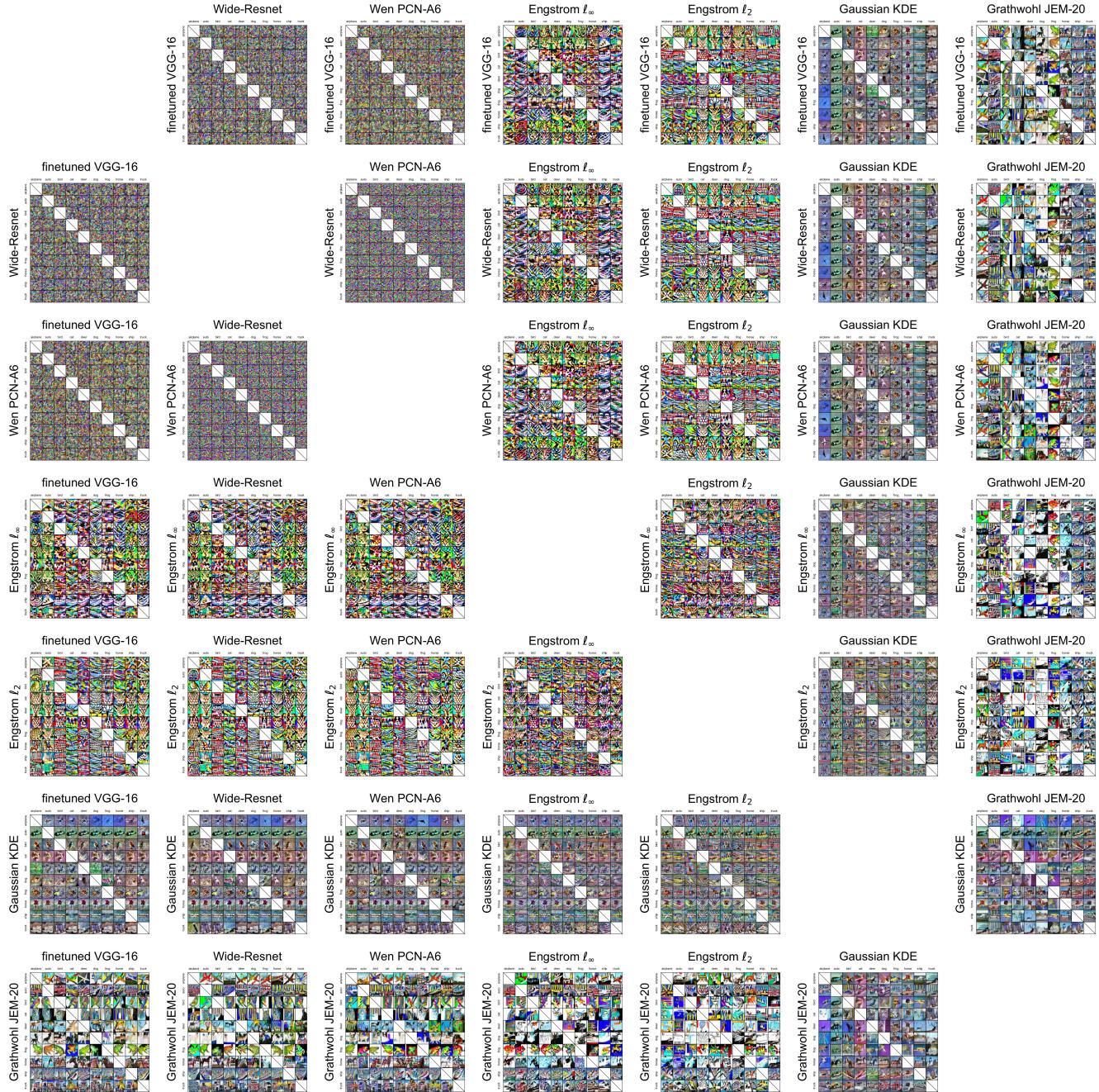


Fig. S11. The entire set of controversial stimuli we synthesized for Experiment 2 (CIFAR-10), organized by model pairs. Each panel indicates a targeted model pair, and the rows and columns within each panel indicate the targeted class pairs. For example, consider the bottom-left image in the bottom-left 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.

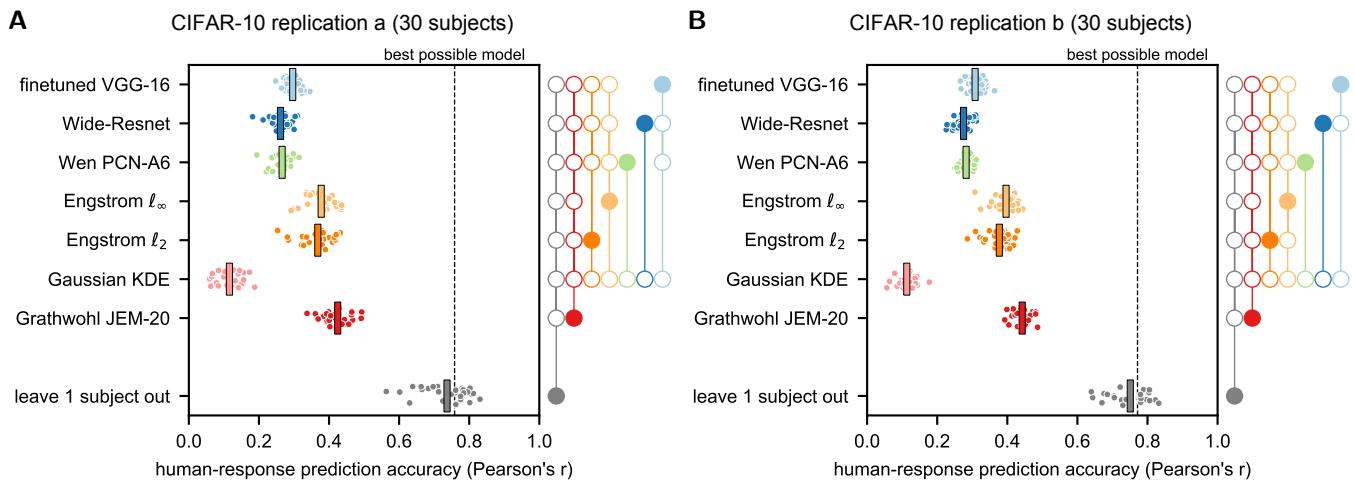


Fig. S12. A side by side comparison of two replications of Experiment 2. Each replication was conducted with an independently synthesized set of 420 controversial stimuli and an independently selected set of 60 CIFAR-10 test examples. 30 subjects participated in each replication and no subject participated 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

model family	model	MNIST test error	human-response prediction accuracy		
			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. Candidate CIFAR-10 models

model family	model	CIFAR-10 test error	human-response prediction accuracy		
			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, 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, 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*.