jove

Through a Dog's Eyes: fMRI Decoding of Naturalistic Videos from the Dog Cortex

Erin M. Phillips[1], **Kirsten D. Gillette**[1], **Daniel D. Dilks**[1], **Gregory S. Berns**[1]

[1] Psychology Department, Emory University

Corresponding Author

Gregory S. Berns
gregory.berns@emory.edu

Citation

Phillips, E.M., Gillette, K.D., Dilks, D.D., Berns, G.S. Through a Dog's Eyes: fMRI Decoding of Naturalistic Videos from the Dog Cortex. *J. Vis. Exp.* (187), e64442, doi:10.3791/64442 (2022).

Date Published

September 13, 2022

DOI

10.3791/64442

URL

jove.com/video/64442

Abstract

Recent advancements using machine learning and functional magnetic resonance imaging (fMRI) to decode visual stimuli from the human and nonhuman cortex have resulted in new insights into the nature of perception. However, this approach has yet to be applied substantially to animals other than primates, raising questions about the nature of such representations across the animal kingdom. Here, we used awake fMRI in two domestic dogs and two humans, obtained while each watched specially created dog-appropriate naturalistic videos. We then trained a neural net (Ivis) to classify the video content from a total of 90 min of recorded brain activity from each. We tested both an object-based classifier, attempting to discriminate categories such as dog, human, and car, and an action-based classifier, attempting to discriminate categories such as eating, sniffing, and talking. Compared to the two human subjects, for whom both types of classifier performed well above chance, only action-based classifiers were successful in decoding video content from the dogs. These results demonstrate the first known application of machine learning to decode naturalistic videos from the brain of a carnivore and suggest that the dog's-eye view of the world may be quite different from our own.

Introduction

The brains of humans, like other primates, demonstrate the parcellation of the visual stream into dorsal and ventral pathways with distinct and well-known functions-the "what" and "where" of objects[1]. This what/where dichotomy has been a useful heuristic for decades, but its anatomical basis is now known to be much more complex, with many researchers favoring a parcellation based on recognition versus action ("what" vs. "how")[2,3,4,5]. Additionally, while our understanding of the organization of the primate visual system continues to be refined and debated, much remains unknown about how the brains of other mammalian species represent visual information. In part, this lacuna is a result of the historical focus on a handful of species in visual neuroscience. New approaches to brain imaging, however, are opening up the possibility of noninvasively studying the visual systems of a wider range of animals, which may yield

new insights into the organization of the mammalian nervous system.

Dogs (*Canis lupus familiaris*) present a rich opportunity to study the representation of visual stimuli in a species evolutionarily distant from primates, as they may be the only animal that can be trained to cooperatively participate in MRI scanning without the need for sedation or restraints[6,7,8]. Due to their co-evolution with humans over the last 15,000 years, dogs also inhabit our environments and are exposed to many of the stimuli that humans encounter on a daily basis, including video screens, which are the preferred way of presenting stimuli in an MRI scanner. Even so, dogs may process these common environmental stimuli in ways that are quite different from humans, which begs the question of how their visual cortex is organized. Basic differences-such as a lack of a fovea, or being a dichromat-may have significant downstream consequences not only for lower-level visual perception but also for higher-level visual representation. Several fMRI studies in dogs have demonstrated the existence of both face- and object-processing regions that appear to follow the general dorsal/ventral stream architecture seen in primates, although it remains unclear whether dogs have face-processing regions per se or whether these regions are selective for the morphology of the head (e.g., dog vs. human)[9,10,11,12,13]. Regardless, the brain of a dog, being smaller than most primates, would be predicted to be less modularized[14], so there may be more intermixing of types of information in the streams or even privileging of certain types of information, like actions. It has been suggested, for example, that movement might be a more salient feature in canine visual perception than texture or color[15]. Additionally, as dogs do not have hands, one of the primary means through which we interact with the world, their visual processing, particularly of objects, may be quite different than that of primates. In line with this, we recently found evidence that interaction with objects by mouth versus paw resulted in greater activation in object-selective regions in the dog brain[16].

Although dogs may be accustomed to video screens in their home environment, that does not mean they are used to looking at images in an experimental setting the same way a human would. The use of more naturalistic stimuli may help to resolve some of these questions. In the last decade, machine learning algorithms have achieved considerable success in decoding naturalistic visual stimuli from human brain activity. Early successes focused on adapting classical, blocked designs to use brain activity to both classify the types of stimuli an individual was seeing, as well as the brain networks that encoded these representations[17,18,19]. As more powerful algorithms were developed, especially neural networks, more complex stimuli could be decoded, including naturalistic videos[20,21]. These classifiers, which are typically trained on neural responses to these videos, generalize to novel stimuli, allowing them to identify what a particular subject was observing at the time of the fMRI response. For example, certain types of actions in movies can be accurately decoded from the human brain, like jumping and turning, while others (e.g., dragging) cannot[22]. Similarly, although many types of objects can be decoded from fMRI responses, general categories appear to be more difficult. Brain decoding is not limited to humans, providing a powerful tool to understand how information is organized in the brains of other species. Analogous fMRI experiments with nonhuman primates have found distinct representations in the temporal lobe for dimensions of animacy and faciness/bodiness, which parallels that in humans[23].

As a first step toward understanding dogs' representations of naturalistic visual stimuli, awake fMRI was used in two highly MRI-adept domestic dogs to measure cortical responses to dog-appropriate videos. In this study, naturalistic videos were used because of their potentially greater ecological validity to a dog and because of their demonstrated success with neural nets that map video content to dog movement[24]. Over three separate sessions, 90 min of fMRI data were obtained from each dog's responses to 256 unique video clips. For comparison, the same procedure was performed on two human volunteers. Then, using a neural network, we trained and tested classifiers to discriminate either "objects" (e.g., human, dog, car) or "actions" (e.g., talking, eating, sniffing) using varying numbers of classes. The goals of this study were two-fold: 1) determine whether naturalistic video stimuli could be decoded from the dog cortex; and 2) if so, provide a first look into whether the organization was similar to that of humans.

Protocol

The dog study was approved by the Emory University IACUC (PROTO201700572), and all owners gave written consent for their dog's participation in the study. Human study procedures were approved by the Emory University IRB, and all participants provided written consent before scanning (IRB00069592).

1. Participants

1. Select the participants (dogs and humans) with no previous exposure to the stimuli presented in the study.
   NOTE: Dog participants were two local pet dogs volunteered by their owners for participation in fMRI training and scanning consistent with that previously described[7]. Bhubo was a 4-year-old male Boxer-mix, and Daisy was an 11-year-old female Boston terrier-mix. Both dogs had previously participated in several fMRI studies (Bhubo: 8 studies, Daisy: 11 studies), some of which involved watching visual stimuli projected onto a screen while in the scanner. They were selected because of their demonstrated ability to stay in the scanner without moving for long periods of time with their owner out of view. Two humans (one male, 34 years old, and one female, 25 years old) also participated in the study. Neither dogs nor humans had previous exposure to the stimuli shown in this study.

2. Stimuli

1. Film the videos (1920 pixels x 1440 pixels, 60 frames per second [fps]) mounted on a handheld stabilizing gimbal.
   NOTE: In this study, the videos were filmed in Atlanta, Georgia, in 2019.
   1. Film naturalistic videos from a "dog's eye view", holding the gimbal at approximately knee height. Design the videos to capture everyday scenarios in the life of a dog.
      NOTE: These included scenes of walking, feeding, playing, humans interacting (with each other and with dogs), dogs interacting with each other, vehicles in motion, and non-dog animals (**Figure 1A**; **Supplementary Movie 1**). In some clips, the subjects in the video interacted directly with the camera, for example, petting, sniffing, or playing with it, while in others, the camera was ignored. Additional footage of deer was obtained from a locally placed camera trap (1920 pixels x 1080 pixels, 30 fps).
   2. Edit the videos into 256 unique 7 s "scenes". Each scene depicted a single event, such as humans

hugging, a dog running, or a deer walking. Assign each scene a unique number and label according to its content (see below).

2. Edit the scenes into five larger compilation videos of approximately 6 min each. Use compilation videos rather than one long film to present a wide variety of stimuli in sequence.

   NOTE: Presenting a wide variety of stimuli would be difficult to achieve if the videos were captured in one long "take". This is consistent with fMRI decoding studies in humans[20,22]. Additionally, presenting compilations of short clips allowed easier creation of a hold-out set on which the trained algorithm could be tested (see section 7, analyses, below), as it was possible to hold out the individual clips instead of one long movie. Four compilation videos had 51 unique scenes, and one had 52. There were no breaks or blank screens between the scenes.

3. Select the scenes semi-randomly to ensure that each video contains exemplars from all the major label categories-dogs, humans, vehicles, nonhuman animals, and interactions.

   NOTE: During the compiling process, all scenes were downsampled to 1920 pixels x 1080 pixels at 30 fps to match the resolution of the MRI projector.

3. Experimental design

1. Scan the participants in a 3T MRI scanner while watching the compilation videos projected onto a screen mounted at the rear of the MRI bore.
2. Play the videos without sound.
3. For dogs, achieve stable positioning of the head by prior training to place their head in a custom-made chin rest, molded to the lower jaw from mid-snout to behind the mandible.
   1. Affix the chin rest to a wood shelf that spans the coil but allows enough space for the paws underneath, resulting in each dog assuming a "sphinx" position (**Figure 1B**). No restraints were used. For further information on the training protocol, see previous awake fMRI dog studies[7].
4. Let the subjects participate in five runs per session, each run consisting of one compilation video watched from start to finish, presented in a random order. For dogs, take short breaks between each run. Deliver food rewards during these breaks to the dog.
5. Let each subject participate in three sessions over 2 weeks. This allows the subject to watch each of the five unique compilation videos three times, yielding an aggregate fMRI time of 90 min per individual.

4. Imaging

1. Scan the dog participants following a protocol consistent with that employed in previous awake fMRI dog studies[7,25].
   1. Obtain the functional scans using a single-shot echo-planar imaging sequence to acquire volumes of 22 sequential 2.5 mm slices with a 20% gap (TE = 28 ms, TR = 1,430 ms, flip angle = 70°, 64 x 64 matrix, 2.5 mm in-plane voxel size, FOV = 160 mm).
   2. For dogs, orient the slices dorsally to the brain with the phase-encoding direction right-to-left, as dogs sit in the MRI in a "sphinx" position, with the neck in line with the brain. Phase encoding right-to-left avoids wrap-around artifacts from the neck into the front of the head. In addition, the major susceptibility

artifact in scanning dogs comes from the frontal sinus, resulting in distortion of the frontal lobe.

2. For humans, obtain axial slices with phase-encoding in the anterior-posterior direction.
    1. To allow for comparison with the dog scans (same TR/TE), use multiband slice acquisition (CMRR, University of Minnesota) for the humans with a multiband acceleration factor of 2 (GRAPPA = 2, TE = 28 ms, TR = 1,430 ms, flip angle = 55°, 88 x 88 matrix, 2.5 mm in-plane voxels, forty four 2.5 mm slices with a 20% gap).
3. For the dogs, also acquire a T2-weighted structural image of the whole brain for each participant using a turbo spin-echo sequence with 1.5 mm isotropic voxels. For the human participants, use a T1-weighted MPRAGE sequence with 1 mm isotropic voxels.
    NOTE: Over the course of three sessions, approximately 4,000 functional volumes were obtained for each participant.

5. Stimulus labels

1. In order to train a model to classify the content presented in the videos, label the scenes first. To do this, divide the 7 s scenes that make up each compilation video into 1.4 s clips. Label short clips rather than individual frames as there are elements of video that cannot be captured by still frames, some of which may be particularly salient to dogs and, therefore, useful in decoding, such as movement.
    NOTE: A clip length of 1.4 s was chosen because this was long enough to capture these dynamic elements and closely matched the TR of 1.43 s, which allows for performing the classification on a volume-by-volume basis.
2. Randomly distribute these 1.4 s clips (n = 1,280) to lab members to manually label each clip using a pre-programmed check-box style submission form.
    NOTE: There were 94 labels chosen to encompass as many key features of the videos as possible, including subjects (e.g., dog, human, cat), number of subjects (1, 2, 3+), objects (e.g., car, bike, toy), actions (e.g., eating, sniffing, talking), interactions (e.g., human-human, human-dog), and setting (indoors, outdoors), among others. This produced a 94-dimensional label vector for each clip (**Supplementary Table 1**).
3. As a consistency check, select a random subset for relabeling by a second lab member. Here, labels were found to be highly consistent across individuals (>95%). For those labels that were not consistent, allow the two lab members to rewatch the clip in question and come to a consensus on the label.
4. For each run, use timestamped log files to determine the onset of the video stimulus relative to the first scan volume.
5. To account for the delay between stimulus presentation and the BOLD response, convolve the labels with a double gamma hemodynamic response function (HRF) and interpolate to the TR of the functional images (1,430 ms) using the Python functions numpy.convolve() and interp().
    NOTE: The end result was a matrix of convolved labels by the total number of scan volumes for each participant (94 labels x 3,932, 3,920, 3,939, and 3,925 volumes for Daisy, Bhubo, Human 1, and Human 2, respectively).
6. Group these labels wherever necessary to create macrolabels for further analysis. For example, combine

all instances of walking (dog walking, human walking, donkey walking) to create a "walking" label.

7. To further remove redundancy in the label set, calculate the variance inflation factor (VIF) for each label, excluding the macrolabels, which are obviously highly correlated.

   NOTE: VIF is a measure of multicollinearity in predictor variables, calculated by regressing each predictor against every other. Higher VIFs indicate more highly correlated predictors. This study employed a VIF threshold of 2, reducing the 94 labels to 52 unique, largely uncorrelated labels (**Supplementary Table 1**).

6. fMRI preprocessing

1. Preprocessing involves motion correction, censoring, and normalization using the AFNI suite (NIH) and its associated functions[26,27]. Use a two-pass, six-parameter rigid-body motion correction to align the volumes to a target volume that is representative of the participant's average head position across runs.
2. Perform censoring to remove volumes with more than 1 mm displacement between scans, as well as those with outlier voxel signal intensities greater than 0.1%. For both dogs, more than 80% of volumes were retained after censoring, and for humans, more than 90% were retained.
3. To improve the signal-to-noise ratio of individual voxels, perform mild spatial smoothing using 3dmerge and a 4 mm Gaussian kernel at full-width half-maximum.
4. To control for the effect of low-level visual features, such as motion or speed, that may differ according to stimulus, calculate the optical flow between consecutive frames of video clips[22,28]. Calculate the optical flow using the Farneback algorithm in OpenCV after downsampling to 10 frames per second[29].
5. To estimate the motion energy in each frame, calculate the sum of squares of the optic flow of each pixel and take the square root of the result, effectively calculating the Euclidean average optic flow from one frame to the next[28,30]. This generates time courses of motion energy for each compilation video.
6. Resample these to match the temporal resolution of the fMRI data, convolved with a double gamma hemodynamic response function (HRF) as above and concatenated to align with stimulus presentation for each subject.
7. Use this time course, along with the motion parameters generated from the motion correction described above, as the only regressors to a general linear model (GLM) estimated for each voxel with AFNI's 3dDeconvolve. Use the residuals of this model as inputs to the machine learning algorithm described below.

7. Analyses

1. Decode those regions of the brain that contribute significantly to the classification of visual stimuli, training a model for each individual participant that can then be used to classify video content based on participants' brain data. Use the Ivis machine learning algorithm, a nonlinear method based on Siamese neural networks (SNNs) that has shown success on high dimensional biological data[31].

   NOTE: SNNs contain two identical sub-networks that are used to learn the similarity of inputs in either supervised or unsupervised modes. Although neural networks have grown in popularity for brain decoding because of their

generally greater power over linear methods like support vector machines (SVMs), we used an SNN here because of its robustness to class imbalance and the need for fewer exemplars. Compared to support vector machines (SVM) and random forest (RF) classifiers trained on the same data, we found Ivis to be more successful in classifying brain data across multiple label combinations, as determined by various metrics, including mean F1 score, precision, recall, and test accuracy (see below).

2. For each participant, convert the whole-brain residuals to a format appropriate for input into the Ivis neural network. Concatenate and mask the five runs in each of their three sessions, retaining only brain voxels.

3. Flatten the spatial dimension, resulting in a two-dimensional matrix of voxels by time.

4. Concatenate the convolved labels of the videos shown in each run, thus corresponding to the fMRI runs.

5. Censor both the fMRI data and corresponding labels according to the volumes flagged in preprocessing.

6. Select the target labels to be decoded-hereafter referred to as "classes"-and retain only those volumes containing these classes. For simplicity, treat classes as mutually exclusive and do not include volumes belonging to multiple classes for decoding, leaving only pure exemplars.

7. Split the data into training and test sets. Use a five-fold split, randomly selecting 20% of the scenes to act as the test set.

   NOTE: This meant that, if a given scene was selected for the test set, all the clips and functional volumes obtained during this scene were held out from the training set. Had the split been performed independent of the scene, volumes from the same scene would have appeared in both the training set and the test set, and the classifier would only have had to match them to that particular scene to be successful in classifying them. However, to correctly classify held-out volumes from new scenes, the classifier had to match them to a more general, scene-independent class. This was a more robust test of the generalizability of the classifier's success compared to holding out individual clips.

8. Balance the training set by undersampling the number of volumes in each class to match that of the smallest class using the scikit-learn package imbalanced-learn.

9. For each participant, train and test the Ivis algorithm on 100 iterations, each time using a unique test-train split (Ivis parameters: k = 5, model = "maaten", n_epochs_without_progress = 30, supervision_weight = 1). These parameter values were largely selected on the basis of dataset size and complexity as recommended by the algorithm's authors in its documentation[32]. "Number of epochs without progress" and "supervision weight" (0 for unsupervised, 1 for supervised) underwent additional parameter tuning to optimize the model.

10. To reduce the number of features used to train the classifier from the whole brain to only the most informative voxels, use a random forest classifier (RFC) using scikit-learn to rank each voxel according to its feature importance.

    NOTE: Although the RFC did not perform above chance by itself, it did serve the useful purpose of screening out noninformative voxels, which would have contributed only noise to the Ivis algorithm. This is similar to using F-tests for feature selection before passing to the classifier[33]. Only the top 5% of voxels from the training set were used in training and testing. The

preferred number of voxels was selected as 5% as a conservative threshold in an effort to reduce the number of noninformative voxels prior to training the neural net. Qualitatively similar results were also obtained for both humans and dogs when using a larger proportion of voxels. Though human brains are larger than dog brains, human models were also successful when trained on an absolute number of voxels equal to those included in dog models, far smaller than 5% of voxels (~250 voxels; all mean LRAP scores >99th percentile). For consistency, we, therefore, present the results using the top 5% of voxels for both species.

11. Normalize the average 5% most informative voxels across all 100 runs, transform to each participant's structural space and then to group atlas space (atlases: humans[34] and dogs[35]), and sum it across participants for each species. Overlay feature importance on the atlases and color them according to importance score using ITK-SNAP[36].

Representative Results

The most common metrics to assess model performance in machine learning analyses include precision, accuracy, recall, and F1 score. Accuracy is the overall percentage of model predictions that are correct, given the true data. Precision is the percentage of the model's positive predictions that are actually positive (i.e., the true positive rate), while recall is the percentage of true positives in the original data that the model is able to successfully predict. F1 score is the weighted average of precision and recall and acts as an alternate measure of accuracy that is more robust to class imbalance. However, the Ivis differs from other commonly used machine learning algorithms in that its output is not binary. Given a particular input of brain voxels, each output element represents the probabilities corresponding to each of the classes. Computing accuracy, precision, recall, and F1 for these outputs required binarizing them in a "winner takes all" fashion, where the class with the highest probability was considered the one predicted for that volume. This approach eliminated important information about the ranking of these probabilities that was relevant to assessing the quality of the model. Thus, while we still computed these traditional metrics, we used the Label Ranking Average Precision (LRAP) score as the primary metric to compute the accuracy of the model on the test set. This metric essentially measures to what extent the classifier assigned higher probabilities to true labels[37].

To different degrees, the neural net classifier was successful for both humans and dogs. For humans, the algorithm was able to classify both objects and actions, with three-class models for both achieving a mean accuracy of 70%. The LRAP score was used as the primary metric to compute the accuracy of the model on the test set; this metric measures the extent to which the classifier assigned higher probabilities to true labels[37]. For both humans, the median LRAP scores were greater than the 99th percentile of a randomly permuted set of labels for all models tested (**Table 1**; **Figure 2**). For dogs, only the action model had a median LRAP percentile rank significantly greater than chance in both participants (**Table 1**; $p = 0.13$ for objects and $p < 0.001$ for actions; mean three-class action model LRAP score for dogs = 78th percentile). These results were true for all subjects individually, as well as when grouped by species.

Given the classifier's success, we trained and tested with additional classes to determine the limits of the model. This included computing dissimilarity matrices for the entire 52 potential classes of interest using the Python package scipy's hierarchical clustering algorithm, which clustered classes

based on the similarity of an individual's brain response to each, as defined by pairwise correlation. Of the additional models tested, the model with the highest median LRAP percentile ranking in both dogs had five classes: the original "talking", "eating", and "sniffing", plus two new classes, "petting" and "playing" (**Figure 2**). This model had a median LRAP percentile rank significantly greater than that predicted by chance for all participants (**Table 1**; $p < 0.001$ for both dogs and humans; mean five-class action model LRAP score for dogs = 81st percentile).

When back-mapped to their respective brain atlases, the feature importance scores of voxels revealed a number of clusters of informative voxels in the occipital, parietal, and temporal cortices of both dogs and humans (**Figure 3**). In humans, the object-based and action-based models revealed a more focal pattern than in the dogs and in regions typically associated with object recognition, although with slight differences in the spatial location of object-based voxels and action-based voxels.

We checked that these species differences were not a result of the task-correlated motion of the dogs moving more to some types of videos than others (e.g., videos other than dogs, say, cars). We calculated the Euclidean norm of the six motion parameters and fit a linear mixed-effects model using the R package lme4, with class as a fixed effect and run number as a random effect for each dog. For each of the final models, we found no significant effect of class type on motion for either Daisy ($F(2, 2252) = 0.83$, $p = 0.44$ for object-based and $F(4, 1235) = 1.87$, $p = 0.11$ for action-based) or Bhubo ($F(2, 2231) = 1.71$, $p = 0.18$ for object-based and $F(4, 1221) = 0.94$, $p = 0.45$ for action-based).

Figure 1: Naturalistic videos and presentation in MRI bore. (**A**) Example frames from video clips shown to the participants. (**B**) Bhubo, a 4-year-old Boxer-mix, watching videos while undergoing awake fMRI. Please click here to view a larger version of this figure.

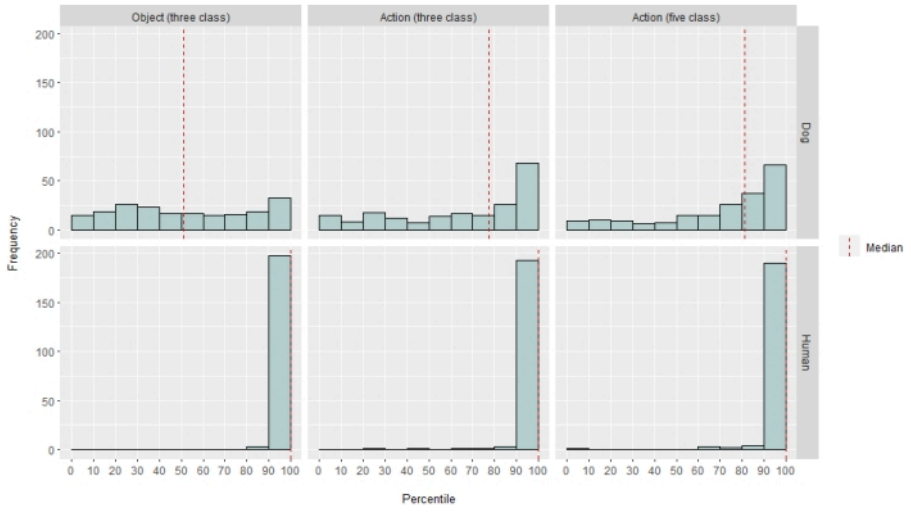

Figure 2: Model performance in dogs and humans. The distribution of LRAP scores, presented as percentile rankings of their null distributions, over 100 iterations of training and testing the Ivis machine learning algorithm for a three-class object-based model, a three-class action-based model, and a five-class action-based model, where models attempted to classify BOLD responses to naturalistic video stimuli obtained *via* awake fMRI in dogs and humans. Scores are aggregated by species. An LRAP score with a very high percentile ranking indicates that the model would be very unlikely to achieve that LRAP score by chance. A model performing no better than chance would have a median LRAP score percentile ranking of ~50. Dashed lines represent the median LRAP score percentile ranking for each species across all 100 runs. Please click here to view a larger version of this figure.

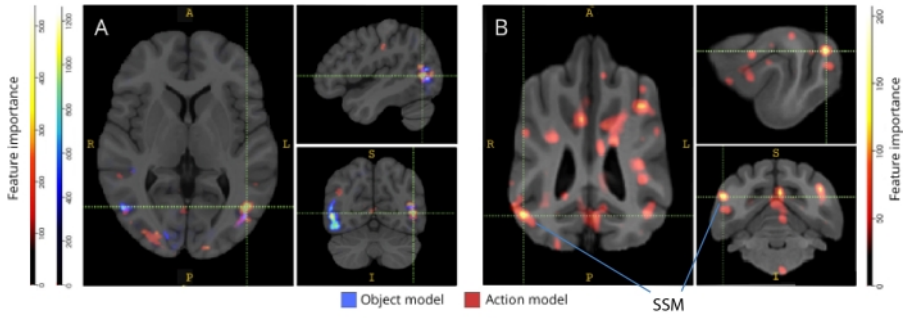

Figure 3: Regions important for the discrimination of three-class object and five-class action models. (**A**) Human and (**B**) dog participants. Voxels were ranked according to their feature importance using a random forest classifier, averaged across all iterations of the models. The top 5% of voxels (i.e., those used to train models) are presented here, aggregated by species and transformed to group space for visualization purposes (atlases: humans[34] and dogs[35]). Labels show dog brain regions with high feature importance scores, based on those identified by Johnson et al.[35]. Abbreviation: SSM = the suprasylvian gyrus. Please click here to view a larger version of this figure.

	Model Type	Training Accuracy	Test Accuracy	F1 Score	Precision	Recall	LRAP score median percentile
Human 1	Object (3 class)	0.98	0.69	0.48	0.52	0.49	**>99**
	Action (3 class)	0.98	0.72	0.51	0.54	0.54	**>99**
	Action (5 class)	0.97	0.51	0.28	0.37	0.27	**>99**
Human 2	Object (3 class)	0.98	0.68	0.45	0.5	0.47	**>99**
	Action (3 class)	0.98	0.69	0.46	0.5	0.48	**>99**
	Action (5 class)	0.97	0.53	0.3	0.4	0.27	**>99**
Bhubo	Object (3 class)	0.99	0.61	0.38	0.41	0.39	57
	Action (3 class)	0.98	0.63	0.38	0.4	0.4	**87**
	Action (5 class)	0.99	0.45	0.16	0.29	0.13	**88**
Daisy	Object (3 class)	1	0.61	0.38	0.43	0.39	43
	Action (3 class)	0.97	0.62	0.35	0.38	0.35	**60**
	Action (5 class)	0.99	0.44	0.16	0.27	0.13	**76**

Table 1: Aggregated metrics of the Ivis machine learning algorithm over 100 iterations of training and testing on BOLD responses to naturalistic video stimuli obtained *via* awake fMRI in dogs and humans. The object models had three target classes ("dog", "human", "car"), and the action models had either three or five classes (three class: "talking",

"eating", "sniffing"; five class: "talking", "eating", "sniffing", "petting", "playing"). Values significantly greater than chance are shown in bold.

Supplementary Table 1: Class labels. Please click here to download this File.

Supplementary Movie 1: Sample video clip. Please click here to download this File.

Discussion

The results of this study demonstrate that naturalistic videos induce representations in dogs' brains that are stable enough over multiple imaging sessions that they can be decoded with fMRI-similar to results obtained in both humans and monkeys[20,23]. While previous fMRI studies of the canine visual system have presented stripped-down stimuli, such as a face or object against a neutral background, the results here demonstrate that naturalistic videos, with multiple people and objects interacting with each other, induce activation patterns in the dog cortex that can be decoded with a reliability approaching that seen in the human cortex. This approach opens up new avenues of investigation for how the dog's visual system is organized.

Although the field of canine fMRI has grown rapidly, to date, these experiments have relied on relatively impoverished stimuli, such as pictures of people or objects against neutral backgrounds[10,12,13]. Additionally, while these experiments have begun to identify brain regions analogous to the primate fusiform face area (FFA), involved in face processing, and the lateral occipital cortex (LOC), for object processing, there remains disagreement over the nature of these representations, such as whether dogs have face areas per se responding to similar salient features as primates or whether they have separate representations for dogs and humans or faces and heads, for example[9,13]. Dogs, of course, are not primates, and we do not know how they interpret these artificial stimuli divorced from their usual multisensory contexts with sounds and smells. Some evidence suggests that dogs do not treat images of objects as representations of real things[12]. Although it is not possible to create a true multisensory experience in the scanner, the use of naturalistic videos may mitigate some of the artificialness by providing dynamic stimuli that more closely match the real world, at least to a dog. For the same reasons, the use of naturalistic stimuli in human fMRI research has gained popularity, demonstrating, for example, that sequences of events in a movie are represented in the cortex across multiple time scales and that movies are effective at inducing reliable emotion activation[38]. As such, while naturalistic videos do remain relatively impoverished stimuli, their success in human neuroscience begs the question of whether similar results can be obtained in dogs.

Our results show that a neural net classifier was successful in decoding some types of naturalistic content from dog brains. This success is an impressive feat given the complexity of the stimuli. Importantly, because the classifier was tested on unseen video clips, the decoding model picked up broad categories that were identifiable across clips rather than properties specific to individual scenes. We should note there are multiple metrics for quantifying the performance of a machine learning classifier (**Table 1**). As naturalistic videos, by their nature, will not have equal occurrences of all classes, we took a prudent approach by constructing a null distribution from the random permutation of labels

and assessing the significance referenced to that. Then, we found that the success of the dog models was statistically significant, achieving 75th-90th percentile scores, but only when the videos were coded based on the actions present, such as playing or talking.

The test sets, unlike the training sets, were not balanced across classes. Comprising only 20% of the data, undersampling to the smallest class size would have resulted in very small sample sizes for each class, such that any statistics calculated would have been unreliable. To avoid the possibility of inflated accuracy from this imbalance, the null distribution of the LRAP was computed by randomly permuting the order of the classes 1,000 times for each model iteration. This null distribution acted as a reference for how well the model was likely to perform by chance. Then, the true LRAP was then converted to a percentile ranking in this null distribution. A very high percentile ranking, for example, 95%, would indicate that a score that high arose only 5% of the time in 1,000 random permutations. Such a model could, therefore, be deemed to be performing well above chance. To determine if these percentile rankings are significantly greater than that expected by chance-that is, the 50th percentile-statistically, the median LRAP percentile ranking across all 100 iterations for each model was calculated and a one-sample Wilcoxon signed rank test was performed.

Although the primary goal was to develop a decoder of naturalistic visual stimuli for dogs, comparisons to humans are unavoidable. Here, we note two major differences: for each type of classifier, the human models performed better than the dog models; and the human models performed well for both object- and action-based models, while the dog models performed for action-based only. The superior performance of the human models could be due to several factors. Human brains are roughly 10 times larger than dog brains, so there are more voxels from which to choose to build a classifier. To put the models on equal footing, one should use the same number of voxels, but this could be in either an absolute or relative sense. Although the final model was based on the top 5% of informative voxels in each brain (a relative measure), similar results were obtained using a fixed number of voxels. Thus, it seems more likely that performance differences are related to how humans and dogs perceive video stimuli. As noted above, while dogs and humans are both multisensory in their perception, the stimuli may be more impoverished to a dog than a human. Size cues, for example, may be lost, with everything appearing to be a toy version of the real world. There is some evidence that dogs categorize objects based on size and texture before shape, which is almost opposite to humans[39]. Additionally, scent, not considered here, is likely an important source of information for object discrimination in dogs, particularly in the identification of conspecifics or humans[40,41,42]. However, even in the absence of size or scent cues, in the unusual environment of the MRI scanner, the fact that the classifier worked at all says that there was still information relevant to the dogs that could be recovered from their brains. With only two dogs and two humans, the species differences could also be due to individual differences. The two dogs, however, represented the best of the MRI-trained dogs and excelled at holding still while watching videos. While a larger sample size would certainly allow more reliable distinctions to be drawn between species, the small number of dogs that are capable of doing awake fMRI and who will watch videos for periods long enough will always limit generalizability to all dogs. While it is possible that specialized breeds, like sighthounds, might have more finely tuned visual brain responses, we believe that

individual temperament and training are more likely to be the major determinants of what is recoverable from a dog's brain.

These species differences raise the question of what aspect of the videos the dogs were paying attention to. One approach to answering this question relies on simpler video stimuli. Then, by using isolated images of, say, humans, dogs, and cars, both individually and together against neutral backgrounds, we might be able to reverse engineer the salient dimensions to a dog. However, this is both methodologically inefficient and further impoverishes the stimuli from the real world. The question of attention can be solved by the decoding approach alone, in effect, using the model performance to determine what is being attended to[43]. Along these lines, the results here suggest that, while the humans attended to both the actors and the actions, the dogs were more focused on the actions themselves. This might be due to differences in low-level motion features, such as the movement frequency when individuals are playing versus eating, or it might be due to a categorical representation of these activities at a higher level. The distribution of informative voxels throughout the dog's cortex suggests that these representations are not just low-level features that would otherwise be confined to visual regions. Further study using a wider variety of video stimuli may illuminate the role of motion in category discrimination by dogs.

In summary, this study has demonstrated the feasibility of recovering naturalistic visual information from the dog cortex using fMRI in the same way that is done for the human cortex. This demonstration shows that, even without sound or smells, salient dimensions of complex scenes are encoded by dogs watching videos and that these dimensions can be recovered from their brains. Secondly, based on the small number of dogs that can do this type of task, the information may be more widely distributed in the cortex than typically seen in humans, and the types of actions seem to be more easily recovered than the identity of the actors or objects. These results open up a new way of examining how dogs perceive the environments they share with humans, including video screens, and suggest rich avenues for future exploration of how they and other non-primate animals "see" the world.

Disclosures

None.

Acknowledgments

We thank Kate Revill, Raveena Chhibber, and Jon King for their helpful insights in the development of this analysis, Mark Spivak for his assistance recruiting and training dogs for MRI, and Phyllis Guo for her help in video creation and labeling. We also thank our dedicated dog owners, Rebecca Beasley (Daisy) and Ashwin Sakhardande (Bhubo). The human studies were supported by a grant from the National Eye Institute (Grant R01 EY029724 to D.D.D.).

References

1. Mishkin, M., Ungerleider, L. G., Macko, K. A. Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences.* **6**, 414-417 (1983).
2. de Haan, E. H. F., Cowey, A. On the usefulness of 'what' and 'where' pathways in vision. *Trends in Cognitive Sciences.* **15** (10), 460-466 (2011).
3. Freud, E., Plaut, D. C., Behrmann, M. 'What' is happening in the dorsal visual pathway. *Trends in Cognitive Sciences.* **20** (10), 773-784 (2016).

4. Goodale, M. A., Milner, A. D. Separate visual pathways for perception and action. *Trends in Neurosciences.* **15** (1), 20-25 (1992).
5. Schenk, T., McIntosh, R. D. Do we have independent visual streams for perception and action? Do we have independent visual streams for perception and action? *Cognitive Neuroscience.* **1** (1), 52-78 (2010).
6. Andics, A., Gácsi, M., Faragó, T., Kis, A., Miklós, Á. Report voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Current Biology.* **24** (5), 574-578 (2014).
7. Berns, G. S., Brooks, A. M., Spivak, M. Functional MRI in awake unrestrained dogs. *PLoS One.* **7** (5), e38027 (2012).
8. Karl, S. et al. Training pet dogs for eye-tracking and awake fMRI. *Behaviour Research Methods.* **52**, 838-856 (2019).
9. Bunford, N. et al. Comparative brain imaging reveals analogous and divergent patterns of species and face sensitivity in humans and dogs. *Journal of Neuroscience.* **40** (43), 8396-8408 (2020).
10. Cuaya, L. V., Hernández-Pérez, R., Concha, L. Our faces in the dog's brain: Functional imaging reveals temporal cortex activation during perception of human faces. *PLoS One.* **11** (3), e0149431 (2016).
11. Dilks, D. D. et al. Awake fMRI reveals a specialized region in dog temporal cortex for face processing. *PeerJ.* **2015** (8), e1115 (2015).
12. Prichard, A. et al. 2D or not 2D? An fMRI study of how dogs visually process objects. *Animal Cognition.* **24** (5), 1143-1151 (2021).
13. Thompkins, A. M. et al. Separate brain areas for processing human and dog faces as revealed by awake fMRI in dogs (*Canis familiaris*). *Learning & Behavior.* **46** (4), 561-573 (2018).
14. Zhang, K., Sejnowski, T. J. A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences of the United States of America.* **97** (10), 5621-5626 (2000).
15. Bradshaw, J., Rooney, N. *Dog Social Behavior and Communication. The Domestic Dog: Its Evolution, Behavior and Interactions with People, .*edited by Serpell, J., 133-160, Cambridge University Press. Cambridge, UK (2017).
16. Prichard, A. et al. The mouth matters most: A functional magnetic resonance imaging study of how dogs perceive inanimate objects. *The Journal of Comparative Neurology.* **529** (11), 2987-2994 (2021).
17. Haxby, J. V., Connolly, A. C., Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience.* **37**, 435-456 (2014).
18. Kamitani, Y., Tong, F. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience.* **8** (5), 679-685 (2005).
19. Kay, K. N., Naselaris, T., Prenger, R. J., Gallant, J. L. Identifying natural images from human brain activity. *Nature.* **452** (7185), 352-355 (2008).
20. Nishimoto, S. et al. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology.* **21** (19), 1641-1646 (2011).
21. van der Meer, J. N., Breakspear, M., Chang, L. J., Sonkusare, S., Cocchi, L. Movie viewing elicits rich and

reliable brain state dynamics. *Nature Communications.* **11** (1), 5004 (2020).

22. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature.* **532** (7600), 453-458 (2016).
23. Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron.* **60** (6), 1126-1141 (2008).
24. Ehsani, K., Bagherinezhad, H., Redmon, J., Mottaghi, R., Farhadi, A. Who let the dogs out? Modeling dog behavior from visual data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* **2018**, 4051-4060 (2018).
25. Berns, G. S., Brooks, A., Spivak, M. Replicability and heterogeneity of awake unrestrained canine fMRI responses. *PLoS One.* **9** (5), e98421 (2013).
26. Cox, R. W. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research.* **29** (3), 162-173 (1996).
27. Prichard, A., Chhibber, R., Athanassiades, K., Spivak, M, Berns, G. S. Fast neural learning in dogs: A multimodal sensory fMRI study. *Scientific Reports.* **8**, 14614 (2018).
28. Russ, B. E., Kaneko, T., Saleem, K. S., Berman, R. A., Leopold, D. A. Distinct fMRI responses to self-induced versus stimulus motion during free viewing in the macaque. *The Journal of Neuroscience.* **36** (37), 9580-9589 (2016).
29. Farnebäck, G. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Image Analysis. Scandinavian Conference on Image Analysis. Lecture Notes in Computer Science,* .edited by Bigun, J., Gustavsson, T. (eds), volume 2749, 363-370, Springer, Berlin, Heidelberg (2003).
30. Elias, D. O., Land, B. R., Mason, A. C., Hoy, R. R. Measuring and quantifying dynamic visual signals in jumping spiders. *Journal of Comparative Physiology A.* **192**, 799-800 (2006).
31. Szubert, B., Cole, J. E., Monaco, C., Drozdov, I. Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific Reports.* **9**, 8914 (2019).
32. Tian, H., Tao, P. IVIS dimensionality reduction framework for biomacromolecular simulations. *Journal of Chemical Information and Modeling.* **60** (10), 4569-4581 (2020).
33. Hebart, M. N., Gorgen, K., Haynes, J. The decoding toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics.* **8**, 88 (2015).
34. Mazziotta, J. et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical Transactions of the Royal Society B: Biological Sciences.* **356** (1412), 1293-1322 (2001).
35. Johnson, P. J. et al. Stereotactic cortical atlas of the domestic canine brain. *Scientific Reports.* **10**, 4781 (2020).
36. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage.* **31** (3), 1116-1128 (2006).
37. Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., Brinker, K. Multilabel classification via calibrated label ranking. *Machine Learning.* **73** (2), 133-153 (2008).

38. Sonkusare, S., Breakspear, M., Guo, C. Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences.* **23** (8), 699-714 (2019).

39. van der Zee, E., Zulch, H., Mills, D. Word generalization by a dog (*Canis familiaris*): Is shape important? *PLoS One.* **7** (11), e49382 (2012).

40. Bekoff, M. Observations of scent-marking and discriminating self from others by a domestic dog (*Canis familiaris*): Tales of displaced yellow snow. *Behavioural Processes.* **55** (2), 75-79 (2001).

41. Berns, G. S., Brooks, A. M., Spivak, M. Scent of the familiar: An fMRI study of canine brain responses to familiar and unfamiliar human and dog odors. *Behavioural Processes.* **110**, 37-46 (2015).

42. Schoon, G. A. A., de Bruin, J. C. The ability of dogs to recognize and cross-match human odours. *Forensic Science International.* **69** (2), 111-118 (1994).

43. Kamitani, Y., Tong, F. Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology.* **16** (11), 1096-1102 (2006).