ENCODING MODELS PREDICT BRAIN ACTIVITY FROM A STIMULUS SUCH AS A WORD, PHRASE, OR IMAGE.

1. Scientists teach an artificial neural network to associate brain activity with certain stimuli.

2. Scientists feed an auditory signal input (such as “I saw a big dog”) into the neural network, the first layer computes and feeds it into the next hidden layers.

3. Hidden layers select specific stimulus features that correlate with brain activity. An example of a feature might be a word, such as “dog.” Related features such as “dog” and “cat” might elicit more similar brain activity than “dog” and “apple.”

4. The neural network returns these features in the final layer. These features can be mapped onto the brain activity data with a computational model.

5. To compare neural networks and brains, scientists can test whether the two pick out similar features of the stimulus or behave similarly during tasks, such as when predicting a missing word.

DECODING MODELS USE NEURAL NETWORKS TO DECODE THE PHRASE FROM BRAIN ACTIVITY.

1. Scientists have decoded natural speech from fMRI data using the transformer neural network, GPT. The neural network keeps a list of potential phrases that participants could have heard such as “I saw a big” and “I saw a dog.”

2. This neural network is trained to predict the next word in the sequence.

3. Scientists seek help from an encoding model to predict brain responses for the proposed next word.

4. The model spits out the most likely phrases and loops to repeat the process with a new list of potential phrases.

5. The encoding model predicts brain responses based on the presented auditory stimuli, while the decoding model attempts to infer the corresponding auditory stimuli using the observed brain responses.

6. Scientists compare the BOLD predictions made by the neural network to the actual BOLD signals.