Interpreting mental state decoding with deep learning models

Armin W. Thomas [★,◊,*], Christopher Ré [▪], and Russell A. Poldrack [★,◊]

★ Stanford Data Science, Stanford University, Stanford, CA, USA

◊ Department of Psychology, Stanford University, Stanford, CA, USA

▪ Department of Computer Science, Stanford University, Stanford, CA, USA

*Correspondence: athms@stanford.edu (A. W. Thomas)

Keywords: mental state decoding, deep learning, neuroimaging, explainable artificial intelligence, transfer learning, reproducibility, robustness

Highlights:

Deep learning (DL) models have transformed many areas of research and industry, with their unparalleled ability to autonomously learn versatile representations of complex data.

Due to this empirical success, neuroimaging researchers have started applying DL models to mental state decoding analyses, hoping that they can provide novel insights into the association between mental states (e.g., accepting/rejecting a gamble) and brain activity, beyond the capabilities of conventional machine learning approaches.

Yet, several challenges at the intersection of functional neuroimaging and DL research hinder the broad application of DL models in mental state decoding.

Here, we review recent advances in both fields to provide a set of solutions to these challenges and enable researchers to fully leverage the potential of DL models in mental state decoding.

In mental state decoding, researchers aim to identify the set of mental states (e.g., experiencing happiness or fear) that can be reliably identified from the activity patterns of a brain region (or network). Deep learning (DL) models are highly promising for mental state decoding, with their unmatched ability to learn versatile representations of complex data. Yet, their widespread application in mental state decoding is hindered by their lack of interpretability, difficulties in applying them to small datasets, and in ensuring their reproducibility and robustness. We recommend to approach these challenges by leveraging recent advances in explainable artificial intelligence and transfer learning, while also providing recommendations on how to improve the reproducibility and robustness of DL models in mental state decoding.

Glossary:

Mental state: An unobservable construct of psychological theory that refers to a particular mental operation or content and is often associated with specific observable behaviors.

Computer vision: An area of artificial intelligence research, which aims to enable computers to derive meaningful information from the visual world and to take actions based on that information.

DL: Deep learning (DL) describes a class of representation learning methods, which transform the input data in multiple sequential steps (or layers), each applying stacks of simple, but nonlinear, functions.

fMRI: Functional magnetic resonance imaging (fMRI) measures brain activity by detecting changes in activity associated with changes in local blood flow.

Natural language processing: An area of artificial intelligence research, which aims to enable computers to derive meaningful information from human language and to take actions based on that information.

Representation: As used in computer science, a transform of some data in terms of a different set of features. Note that this definition stands in contrast to the understanding of representations in cognitive neuroscience, where they indicate the set of mental states that is encoded in (or represented by) the patterns of neural activity of a brain region (or network).

XAI: Explainable artificial intelligence (XAI) represents a class of methods, which aim to make the behavior of DL models understandable to human observers, for example, by relating the features of the input data to the respective outputs of the model.

The promise of deep learning

Over the last decade, deep learning (**DL**; see Glossary and [1]) models have revolutionized many areas of research and industry with their ability to learn highly versatile **representations** of complex data. A defining feature of DL models is that they sequentially apply stacks of many simple, but nonlinear, transforms to their input data, allowing them to gain an increasingly abstracted view of the data. At each level of the transform, new representations of the data are built by the use of representations from preceding layers. The resulting high-level view of the data enables DL models to capture complex nonlinearities, associate a target signal with highly variable patterns in the data (e.g., when transcribing audio recordings), and effectively filter out aspects of the data that are irrelevant to the learning task at hand. A key driver for the empirical success of DL models is their ability to autonomously learn these different levels of abstraction from sufficiently large datasets, without the need for extensive data preprocessing or a prior understanding of the mapping between input data and target signal.

This empirical success has recently sparked interest in the application of DL models to the field of neuroimaging, focused on **mental state** decoding [2]. Here, researchers aim to understand

the mapping between a set of mental states (e.g., the experience of anger or sadness) and the underlying brain activity by training models to identify these states from measured brain activity [3]. At first sight, DL models seem ideally suited for these types of analyses, as the mapping between mental states and brain activity is often a priori unknown, can be highly variable within [4] and between individuals [5], and is subject to spatial and temporal non-linearities [6].

Yet, the application of DL models to mental state decoding analyses also poses several challenges for researchers who are interested in combining methods from both fields, namely, their general lack of interpretability, overall demand for large training datasets, and difficulties in ensuring the reproducibility and robustness of DL modeling results. Here, we outline these challenges and propose a set of solutions based on related empirical work and methodological advances in functional neuroimaging and DL research.

Opening up the black box

A key challenge for the application of DL models to functional neuroimaging data is the black box characteristic of DL models, whose highly non-linear nature deeply obscures the relationships between input data and a model's decoding decisions. Thus, even if a DL model accurately decodes a set of mental states from functional neuroimaging data, it is not clear which particular features of the data (or combinations thereof) support this decoding. To approach this challenge, functional neuroimaging researchers have begun turning towards research on explainable artificial intelligence (**XAI**; [7,8]), where techniques are being developed that aim to make the behavior of DL models understandable for human observers.

One line of research within this field seeks to explain the predictions of DL models by relating them to the features of the input data, thereby making the model interpretable for human observers [9]. While a plethora of such explanation approaches exist, we focus here on those that explain model predictions by attributing a relevance to each input feature for a model's prediction [10–17], due to the widespread application of these approaches in mental state decoding. We provide an overview of representative approaches to this type of XAI in Box 1. Of these approaches, sensitivity analyses, backward decompositions, and reference-based attributions are currently most prominent in the neuroimaging literature [18–31]. Sensitivity analyses attribute a relevance to each input feature according to how sensitive the model's prediction responds to the

feature's value. Backward decompositions, in contrast, attribute relevance by sequentially decomposing the model's prediction in a backward pass through the model into the contributions of lower-layer model units to the predictions, until the input space is reached and a contribution (i.e., relevance) can be defined for each input feature. Lastly, reference-based attribution methods attribute relevance by contrasting the model's response to an input of interest to its response to some reference input (e.g., a neutral input [13]).

Box 1. Representative XAI attribution approaches.

We assume that the analyzed model represents some function $f(\cdot)$, mapping an input $x \in \mathbb{R}^N$ to some output $f(x)$: $f(\cdot)$: $\mathbb{R}^N \rightarrow \mathbb{R}$. The presented explanation approaches $\eta(\cdot)$ seek to provide insights into this mapping by attributing a relevance r_n to each input feature $n \in 1,\ldots,N$ for output $f(x)$: $\eta(\cdot)$: $\mathbb{R} \rightarrow \mathbb{R}^N$ (Fig. I).

Occlusion analysis [16,138]**:** Occlusion analyses represent a form of perturbation analysis and quantify r_n by occluding x_n in the input data and measuring the resulting effect on $f(x)$: $r_n = f(x) - f(x \times o_n)$. Here, o_n indicates an occlusion vector (e.g., $o_n \in [0,1]^N$) and $\times$ the element-wise product.

Interpretable local surrogate model [15]**:** A local surrogate model is an interpretable model that is used to explain black-box model predictions by training it to approximate these predictions. In the LIME algorithm [15], r_n is quantified by approximating $f(x)$ for a specific x with an interpretable model $g(\cdot)$, e.g., a linear model, where $g(x) = \sum_n w_n x_n$, and which is trained by the use of a set of perturbed versions Z of x (e.g., through occlusion): $\min_g \sum_{z \in Z} \pi_x(z)\,(f(z) - g(z))^2$. Here, $\pi_x(\cdot)$ represents some similarity function weighting each $z \in Z$ by its similarity to x and r_n is given by linear model weight w_n: $r_n = w_n$.

Sensitivity analysis [10,12,139]**:** Sensitivity analysis defines r_n as the locally evaluated partial derivative of $f(x)$: $r_n = \frac{\partial f(x)}{\partial x_n}$ (or as its square $(\frac{\partial f(x)}{\partial x_n})^2$). Accordingly, relevance is assigned to those input features to which $f(x)$ responds most sensitively.

Backward decomposition [11,14,16]**:** Backward decompositions make specific use of the graph structure of DL models by sequentially decomposing $f(x)$ in a backward pass through the model until the input space is reached. A prominent example is the layer-wise relevance propagation (LRP; [11]) technique: Let i and j be the indices of two model units in two successive layers l and $l+1$ and $r_j^{(l+1)}$ the relevance of unit j for $f(x)$. To redistribute relevance between successive layers, several rules have been proposed [140], which generally follow from: $r_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} r_j^{(l+1)}$, where a and w represent the input and weights of unit i in layer l. Importantly, LRP assumes that relevance is conserved between layers, such that $\sum_n r_n = \sum_i r_i^{(l)} = \sum_j r_j^{(l+1)} = f(x)$.

Reference-based attribution [13,14,17]: Reference-based attributions define r_n, given some x, by contrasting the model's response to x to its response to a reference input x^0. For example, integrated gradients (IG; [13]) defines r_n by integrating the gradient $\frac{\partial f(x)}{\partial x_n}$ along a linear trajectory in the input space connecting a neutral reference input x_n^0 to the current input x_n: $r_n = (x_n - x_n^0) \int_{\alpha=0}^{1} \frac{\delta f(x^0 + \alpha(x - x^0))}{\delta x_n} d\alpha$. Conceptually, IG identifies those input features that most impact the model's output when scaled from the reference value to their current value. Note that IG's attributions sum to the difference in model output for the current input x and the reference x^0: $\sum_n r_n = f(x) - f(x^0)$. Another prominent reference-based attribution method is SHAP (SHapley Additive exPlanations; [17]), an extension of Shapley values [141] to XAI, which uses other possible coalitions of input features as a reference.

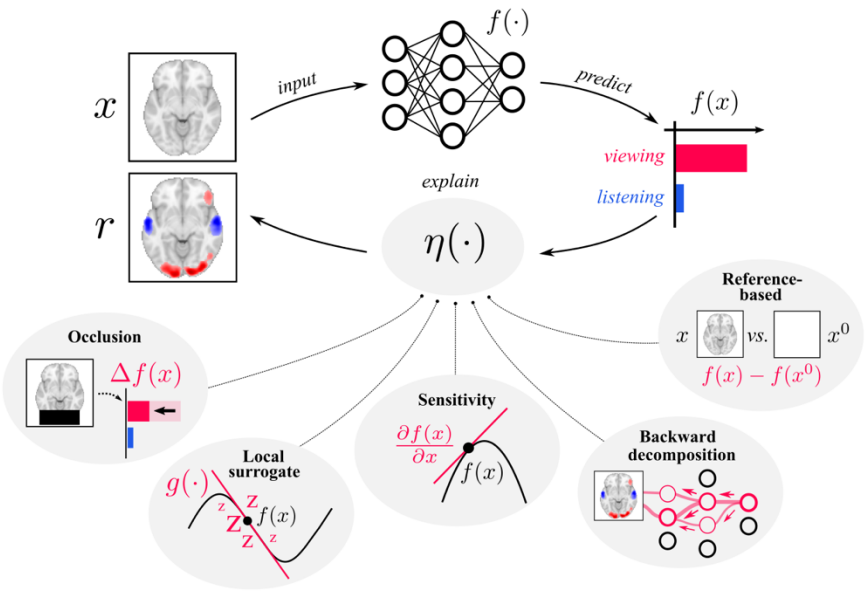

Figure I. Representative XAI attribution approaches.

At first sight, the explanations of different attribution methods are difficult for human observers to discern, making it challenging to compare the quality of their explanations. To approach this challenge, researchers have started developing methods to quantify the quality of such explanations. One prominent approach is to test the faithfulness of an explanation [32–34]. An explanation can generally be viewed as being faithful if it accurately captures the model's decision process [35] and thereby identifies those features of the input that are most relevant for the model's prediction. Accordingly, removing these features from the input (e.g., in an occlusion analysis; [16]) should lead to a meaningful decline in the model's predictive performance.

By the use of this test, researchers in computer vision have compared the faithfulness of explanations resulting from sensitivity analysis and backward decompositions [32]. This work has shown that backward decompositions generally perform better at identifying those features of the input that are most relevant for model predictions. Intuitively, this makes sense, as backward decompositions seek to directly quantify the contribution of each input feature to a specific model prediction. Sensitivity analysis, in contrast, does not evaluate the prediction itself but its local slope, thus identifying features that make the model more or less certain of its prediction, regardless of their actual contribution to the prediction.

Recent work in functional neuroimaging has performed a similar comparison of sensitivity analyses, backward decompositions, and reference-based attributions in a mental state decoding analysis with fMRI data [36]. Similar to findings in computer vision, this work shows that explanations from backward decompositions and reference-based attributions are generally more faithful than those of sensitivity analyses. Yet, it also demonstrates that the explanations of sensitivity analyses generally align better with the results of standard general linear model analyses of the fMRI data, when compared to those of backward decompositions or reference-based attributions. To make sense of this finding, it is important to remember that these types of XAI techniques seek to explain the mapping between brain activity and mental states learned by a model. Due to the generally strong spatial correlations of functional neuroimaging data, DL models can, in many cases, accurately decode a mental state by focusing solely on a subset of those voxels whose activity is associated with (and thereby predictive of) this state. In these cases, XAI methods with perfect faithfulness will produce explanations that do not identify all voxels of the input whose activity is in fact associated with the mental state, but solely those whose activity the model used as evidence for its decoding decision. Sensitivity analyses, in contrast, take a step back from the specific contribution of each input voxel to the decoding decision and instead ask how sensitively the model's decoding decision responds to a voxel's value, thereby identifying a broader set of voxels whose activity the model takes into account when forming its decoding decision.

Functional neuroimaging researchers have also used occlusion analyses to analyze mental state decoding models (in "virtual lesion analyses"; [23,37]). Yet, these applications have mostly been limited to linear models and to testing whether specific voxels (or brain regions), which received large weights in a linear model, are actually necessary for an accurate decoding. For functional neuroimaging data, occlusion analyses generally require a clear prior hypothesis on which features (or brain regions) of the input will be tested (e.g., based on other research), as randomly dropping out individual feature values will otherwise not account for the strong spatial correlation structure inherent to these data. To circumvent these issues, neuroimaging researchers can perform occlusion analyses on the level of functionally independent brain networks, as defined by a brain parcellation [38,39], instead of on the level of individual voxel values [40,41].

Taken together, we therefore make a two-fold recommendation for XAI techniques in mental state decoding (see Box 2): if researchers are interested in identifying those voxels of the input whose activity is most relevant for the model's decoding decision, we recommend the

application of backward decompositions or reference-based attributions, while we recommend sensitivity analyses when researchers are more interested in understanding the association between the underlying brain activity and studied mental states. Occlusion analyses also represent a viable alternative to these approaches, if researchers are interested in relating the activity of functionally independent brain networks to the decoded mental states rather than the activity of individual voxels.

Importantly, while XAI techniques represent a cornerstone to the application of DL models in mental state decoding, we advocate for caution in the interpretation of their explanations, as the mappings between brain activity and mental states learned by DL models can be highly complex and counterintuitive [22,42,43]. We therefore urge neuroimaging researchers to always interpret the results of an XAI analysis in the context of the results of standard analyses of the same data (e.g., with linear models; [44,45]) and related empirical findings (e.g., from NeuroSynth; [46]).

Box 2. Recommended XAI approaches for mental state decoding.

Our recommendations for XAI approaches in mental state decoding are two-fold: If researchers are interested in understanding the contribution of individual feature values to model decisions, we generally recommend backward decomposition or reference-based attribution methods (see Box 1 and [11,13,14,16,17]), while we recommend sensitivity analyses (see Box 1 and [10,12,139,142]) when researchers are more interested in understanding the association between the underlying brain activity and mental states. Below, we provide specific recommendations for respective XAI techniques:

Layer-wise relevance propagation (LRP) [11]: LRP represents a backward decomposition method (see Box 1). While several rules have been proposed to redistribute relevance r between units i and j of two successive layers l and $l+1$ [11, 140], the authors generally recommend a composite of these rules for computer vision models [140]. Specifically, combining the LRP-0 rule ($\mathrm{r}_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij}} \mathrm{r}_j^{(l+1)}$, where a and w represent the input and weights of unit i and $\Sigma_{0,i}$ runs over all inputs a_i plus the bias) for layers closer to the output, with the LRP-ϵ rule ($\mathrm{r}_i^{(l)} = \sum_j \frac{a_i w_{ij}}{\epsilon + \sum_{0,i} a_i w_{ij}} \mathrm{r}_j^{(l+1)}$, with $1e^{-4} \leq \epsilon < 1$) for middle layers, and the LRP-γ rule ($\mathrm{r}_i^{(l)} = \sum_j \frac{a_i (w_{ij} + \gamma w_{ij}^+)}{\sum_{0,i} a_i (w_{ij} + \gamma w_{ij}^+)} \mathrm{r}_j^{(l+1)}$, where γ controls positive contributions and is generally $0 < \gamma$) for layers closer to the input. A TensorFlow implementation of LRP is provided by iNNvestigate [143], while Zennit [144] provides a PyTorch implementation.

Integrated gradients (IG) [13]: IG represents a reference-based attribution method that is applicable to any differentiable model (see Box 1). An important hyperparameter choice for IG is the choice of a reference input x^0, which should be chosen to be neutral. The authors generally recommend an all-zero reference, the addition of noise to the input or a reference involving instances from other decoding classes (e.g., their average), while an average over the attributions of multiple references is also

possible [145]. A tutorial on how to use IG in TensorFlow can be found at tensorflow.org/tutorials/interpretability/integrated_gradients, while Captum [146] provides a PyTorch implementation. A comparable alternative to IG is the DeepLift algorithm [14], which generally runs faster than IG and is therefore often preferred for larger datasets.

Sensitivity analysis [16]: Similar to IG, sensitivity analyses are applicable to any differentiable model (see Box 1). Note that various adaptations of the standard sensitivity analysis have been developed, for example, by omitting negative gradients in rectified linear unit activation functions [142], multiplying gradients and input [147] or by adding noise to the inputs [12]. A TensorFlow implementation of sensitivity analysis (as well as many of its adaptations) is provided by iNNvestigate [143], while Captum [146] provides respective PyTorch implementations.

Leveraging public data

A second major challenge for DL models in functional neuroimaging research is the high dimensionality and low sample size of conventional functional neuroimaging datasets. A typical functional Magnetic Resonance Imaging (**fMRI**) dataset contains a few hundred volumes for each of tens to hundreds of individuals, while each volume contains several hundred thousand voxels (i.e., dimensions). Current state-of-the-art DL models, in contrast, can easily contain many hundred million parameters [47,48], while recent language models have pushed this boundary even further with many billion parameters [49]. In most cases, DL models thus contain many more trainable parameters than there are samples in their training data. While this vast overparameterization represents a key element to the empirical success of DL models, by enabling them to find near-perfect solutions for most standard learning tasks [50] and to generalize well between datasets [49,51], it also represents one of the biggest challenges for their application in fields where data are scarce, as the performance of DL models is strongly dependent on the amount of available training data [51,52].

To approach this challenge, various methods have been developed that aim to improve the performance of DL models in smaller datasets [53–55]. One prominent method, with strong empirical success, is transfer learning [55]. The goal of transfer learning is to leverage the knowledge about a mapping between input data and a target variable that can be learned from one dataset (i.e., the source domain) to subsequently improve the learning of a similar mapping in another dataset of a related domain (i.e., the target domain). Knowledge is typically transferred in

the form of the parameters that a model has learned in the source domain and that are then used to initialize the model (or parts of the model) when beginning learning in the target domain. Transfer learning has been especially successful in computer vision and natural language processing, where large publicly available datasets exist (e.g., [56,57] and commoncrawl.org). Here, DL models are first pre-trained on these large datasets (e.g., to classify objects in images or to predict the next word in a sentence) and subsequently fine-tuned on smaller datasets of a related target domain (e.g., to classify brain tumors in medical imaging [58] or to analyze sentiment in text [48]). Computationally, pre-training can aid subsequent optimizations by placing the model's parameters near a local minimum of the loss function [59] and by acting as a regularizer [60]**.** Pre-trained models generally exhibit faster learning and higher predictive accuracy, while also requiring less training data when compared to models that are trained from scratch [49,51,61]. However, the benefits of pre-training can diminish with increasing size of the target dataset [51] and as the overall differences between source and target learning task and/or domain increase [62].

Over recent years, functional neuroimaging research has experienced a similar increase in the availability of public datasets, which are provided by large neuroimaging initiatives as well as individual researchers [63]. In addition, several efforts have been made to standardize the organization [64,65] and preprocessing [66] of functional neuroimaging data. These developments have paved the way for the field of functional neuroimaging to enter a big data era, allowing for transfer learning.

Recent empirical evidence indicates that transfer learning between individuals [24,67–73], experiment tasks [74–77], and datasets [78–81] is possible and that pre-training generally improves the decoding performance of DL models in conventional fMRI datasets [68,69,74,77,78,80]. Most of this work has utilized traditional supervised learning techniques during pre-training by assigning a mental state to each sample in the data and training a decoding model to identify these states from the data. While this is a fruitful approach to decoding analyses within individual datasets, it is often difficult to extend to analyses across many datasets. In spite of several attempts [82,83], functional neuroimaging research has yet to widely adopt standardized definitions of mental states. Without this type of standardization, it is often unclear whether two experiments from two separate laboratories elicit the same or different sets of mental states. Imagine the following experiments: In the first, participants read aloud a sequence of sentences and are then asked to repeat the last word of each sentence. In the second, participants first hear a sequence of letters and digits and are

297 then asked to report back the letters and digits in alphabetical and numerical order respectively
298 (the letter–number sequencing task; [84]). While both experiments label the associated mental
299 state as "working memory", one could argue that the experiments in fact elicit two distinct mental
300 states, as one solely requires temporarily storing information while the other also requires actively
301 manipulating this information.

302 To enable successful learning across datasets with these types of imprecise mental state
303 labels, we recommend three learning approaches (see Box 3):

304 First, one can consider each dataset as a separate learning task and train a single model to
305 jointly solve all tasks [85]. Recent empirical work has already demonstrated the versatility of this
306 kind of multi-task learning approach for mental state decoding by training a single model to learn
307 a common data representation from many datasets and using dataset-specific decoding models to
308 identify mental states from the learned common representation [80].

309 A second approach comes from weakly-supervised learning, where techniques have been
310 developed that enable model training with noisy or incomplete data labels [86]. Data programming,
311 a weakly-supervised learning technique, is particularly promising for training DL models across
312 neuroimaging datasets with imprecise labels for mental states (see Box 3 and [87]). Here, simple
313 functions are used to generate new labels for the training data. These functions automatically label
314 subsets of the data by implementing simple domain heuristics of subject matter experts (e.g., label
315 a YouTube text comment as Spam if it contains a URL or the words "check this out"). The
316 generated labels are then used to train models in a supervised manner. Recent empirical work has
317 demonstrated that this type of weak supervision can be successfully used for the classification of
318 unlabeled medical imaging data (e.g., radiography or computer tomography data; [88]) by
319 designing labeling functions that extract labels from the accompanying medical text reports. A
320 similar approach could be fruitful to generate standardized labels of mental states (e.g., according
321 to the Cognitive Atlas; [83]) by applying automatic labeling functions to the accompanying
322 publication texts (e.g., label an fMRI scan as "visual perception" if the publication text contains
323 the words "viewed" or "viewing" in the Methods section).

324 Yet, even standardized labels for mental states can be imprecise with respect to the
325 underlying distribution of brain activity. Imagine a simple experiment in which individuals view
326 images of faces and houses. A decoding model might perform well in identifying that a face or

house is seen, while missing out on other important characteristics of the brain activity associated with the more fine-grained characteristics of the stimuli, such as an individual's age and gender.

Here, self-supervised (or unsupervised) learning techniques provide a means to learning that does not consider any labeling of the data and instead enables models to autonomously learn meaningful representations of the data (see Box 3 and [89]). Two prominent examples of self-supervised learning, with strong recent empirical success [48,49], are contrastive and generative learning [90]. Both learn a representation of the data by training an encoder model to project the data into a higher-level representation. In contrastive learning [91], the encoder model is trained by the use of an additional discriminator model, which aims to determine the similarity of a pair of data samples based on their projection through the encoder model. Generative learning [92], in contrast, trains the encoder model by the use of an additional decoder model, which seeks to reconstruct the input (or parts of the input) from the higher-level representation of the encoder model (a prominent example of generative learning models are autoencoders; [93]). Researchers have already demonstrated that self-supervised learning techniques can be successfully used to pre-train DL models across many and diverse fMRI datasets, leading to models that generalize well to other fMRI datasets in mental state decoding analyses [94].

Box 3. Approaches to pre-training across many neuroimaging datasets.

Transfer learning aims to improve the performance of model $f(\cdot)$ in a target learning task T_T in a target domain D_T by leveraging knowledge that can be learned by *pre-training* $f(\cdot)$ in a related source learning task T_S and source domain D_S [55]. A domain D is defined by feature space X with samples $x \in \mathbb{R}^N$ whose N feature values are characterized by some probability distribution $P(X)$. Knowledge is generally transferred in the form of the weights W that $f(\cdot)$ has learned during pre-training. A key challenge for pre-training in mental state decoding is that the labels assigned to individual mental states can be imprecise, such that two datasets might assign the same label to a mental state while the underlying mental states are in fact different from one another. We recommend three learning approaches (Fig. I) to enable $f(\cdot)$ to learn in a source domain that is characterized by a set A of datasets a^j , where $A = \{a^1, \ldots, a^j\}$ and $a^j = \{(x_1^j, y_1^j), \ldots, (x_i^j, y_i^j)\}$, with imprecise mental state labels $y_i^j \in a^j$.

Multi-task learning ([85]; Fig. I A**):** In multi-task learning, each dataset a^j is considered as a separate supervised learning task and one model $f(\cdot)$ is trained to jointly solve all tasks. A prominent approach to multi-task learning [80] is to train $f(\cdot)$ in conjunction with dataset-specific decoding models $d^j(\cdot)$, such that $f(\cdot)$ learns a common representation of the data $f(\cdot): \mathbb{R}^N \rightarrow \mathbb{R}^L$, which is then used by the individual decoding models to identify their set of mental states: $d^j(f(x_i^j)) = y_i^j$

Weakly-supervised learning ([86]; Fig. I B): A prominent example of weak supervision is data programming [87], where noisy target values $\hat{y}_i^j$ are generated for the samples $x_i^j \in A$ by the use of user-specified labeling functions $l(\cdot)$. These labeling

functions implement domain heuristics of subject matter experts (e.g., label a chest radiograph as "abnormal" if the corresponding medical text report contains a word with the prefix ''pneumo''; [88]). The generated target values are then used to train $f(\cdot)$ in a supervised way, such that $f(x_i^j) = \hat{y}_i^j$, while the labeling process itself is treated as a generative model to account for noise and conflicts in the generated labels.

Self-supervised learning ([89,90]; Fig. I C): Self-supervised learning does not consider any labeling of the data. Instead, a new learning task is devised, which requires $f(\cdot)$ to independently learn a useful representation of the data in the source domain. Two prominent self-supervised learning strategies are contrastive and generative learning. Both treat $f(\cdot)$ as an encoder model, which is trained to project the samples $x_i^j \in A$ into a higher-level representation: $f(\cdot)\colon \mathbb{R}^N \to \mathbb{R}^L$. In contrastive learning [91], $f(\cdot)$ is trained by the use of an additional discriminator model $d(\cdot)\colon \mathbb{R}^L \to \mathbb{R}$, which learns to determine the similarity of a pair of data samples based on the encoder's projection. During training, augmentation functions $\Gamma(\cdot)\colon \mathbb{R}^N \to \mathbb{R}^N$ are used to create augmented versions $\{\hat{x}_i^1, \ldots, \hat{x}_i^z\}$ of data samples x_i^j (e.g., by adding noise) and the discriminator's task is to identify pairs $\{\hat{x}_i^k, \hat{x}_i^l\}$ that result from the same sample x_i^j. In generative learning [92], $f(\cdot)$ is trained by the use of an additional decoder model $d(\cdot)\colon \mathbb{R}^L \to \mathbb{R}^N$, which aims to reconstruct the original data sample from the encoder's projection: $d(f(x_i^j)) = x_i^j$.

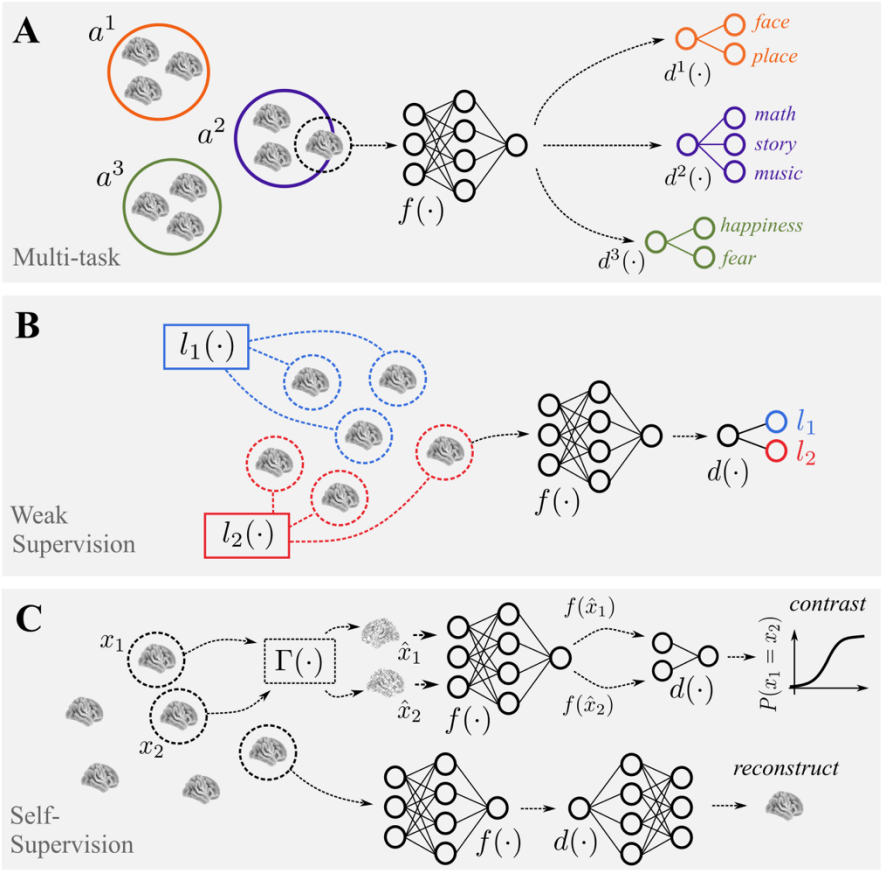

Figure I. Recommended approaches to pre-training DL models across multiple neuroimaging datasets.

Ensuring reproducibility

Recent work in functional neuroimaging has exposed the high flexibility of its standard analysis workflows, leading to substantial variability in results and scientific conclusions [95]. In light of these issues, several efforts have been made to improve the standardization and reproducibility of functional neuroimaging analyses [64,66]. DL research is currently facing similar concerns, with model performances that are often hard to reproduce [96–98]. Functional neuroimaging researchers who are interested in applying DL models to mental state decoding analyses are thus faced with additional challenges for the reproducibility of their work, which arise at the intersection of both fields.

A key driver for methodological progress in DL research is the hunt for state-of-the-art performances in benchmarks (see paperswithcode.com/sota), that is, by whether a new methodology outperforms existing ones in pre-defined test datasets. While this approach has helped the field of DL to evolve fast and quickly develop accurate models, it has also established a research culture that often sacrifices scientific rigor for maximal performance metrics [99,100], not unlike the "p-hacking" phenomenon in null hypothesis testing [101].

A central argument for predefined test datasets is that all models should be compared on the same grounds (i.e., the same sets of training and testing samples). Yet, these types of point estimates are often insufficient to determine whether a model actually outperforms others in new data. Recent empirical work has demonstrated, for example, that the convergence of DL models and thereby their final performance in a test dataset is dependent on many non-deterministic factors of the training, such as random weight initializations and random shufflings or augmentations of the data during training [98,102,103], as well as the specific choices for hyper-parameters, such as the specification of model layers and optimization algorithm [104]. In some cases, researchers can thus achieve state-of-the-art performance simply by investing large computational budgets into tuning these types of factors for a specific test dataset [102]. Consequently, many reported DL benchmarks are built on top of massive computational budgets and are often difficult to reproduce by other researchers [98,103,105]. Recent empirical findings further suggest that the comparisons performed on several of these benchmarks lack the statistical power required to accurately determine the reported improvements in model performance [106], a problem similarly evident in neuroimaging research [107].

For these reasons, researchers have started advocating for more comprehensive and standardized reporting of the training history of DL models [108], more extensive evaluation procedures [109,110] as well as an increased scientific rigor in DL research [99]. To avoid similar pitfalls in mental state decoding, we have derived a set of recommendations from recent DL research, which aim to improve the reproducibility of DL model performances (see Box 4).

Most DL training pipelines are too complex to allow for a comprehensive evaluation of all possible coalitions of the training's non-deterministic factors. However, evaluating only a specific instance of these choices (e.g., by fixing the random seed) does not give a reliable estimate of a model's expected performance in new data. Instead, the variance in model performance associated with these factors can be better captured by randomizing as many of them as possible, for instance, by choosing different random seeds for each of multiple training runs [103,108,111].

In addition, multiple random splits of the data into training, validation, and test datasets are needed when evaluating model performances, to account for the variance in model performance associated with different data splits (e.g., by the use of cross-validation; [97,111,112]). A single, predefined test dataset contains limited information about the whole underlying data distribution and is thus limited in its ability to provide an accurate estimate of the model's expected performance. Yet, recent work has also shown that cross-validation analyses on small functional neuroimaging datasets often underestimate the error in estimates of a model's expected performance [112]. When using small datasets, cross-validation analyses should therefore be treated with caution.

Further, to ensure that the chosen combination of statistical comparison method and test dataset size provide sufficient statistical power to accurately determine the studied difference in model performance, simple simulation studies can be used by first identifying and estimating the required quantities of the statistical testing procedure (e.g., McNemar's test for paired data requires the models' probabilities of making a correct prediction as well as their agreement rate) and subsequently using these estimates to simulate model comparisons at different test dataset sizes [106]. In addition to ensuring that the chosen performance evaluation procedure does not lack statistical power, recent work in neuroimaging also suggests controlling for multiple sequential model comparisons, as multiple sequential hypothesis tests (e.g., performance comparisons) on the same dataset can inflate false positive rates [113].

Improving robustness

In addition to the presented reproducibility challenges, a wealth of recent empirical work has shown that highly-tuned DL models often lack basic robustness towards slight distributional shifts [109,114] or corruptions [115] of the data, such that minor changes to their input, often not recognizable for human observers, can have drastic effects on model performances [116,117]. DL models trained on functional neuroimaging data seem especially susceptible to these kinds of robustness issues, due to the many systematic sources of noise inherent to these data, which can be specific to the imaging acquisition site and studied individual [118] as well as the general variability of the associations of brain activity and mental states between experimental studies and individuals [119–121]. For this reason, training models on large, homogenous datasets (e.g., comprising data acquired at the same imaging site from a homogenous group of individuals performing the same experiment task) can result in models that do not generalize well to data from other imaging sites or subject populations [122–124].

To strengthen robustness towards slight distributional shifts or corruptions of the data, DL researchers generally suggest applying random augmentations to the data during training, such as randomly cropping, rotating or flipping images [125] or occluding parts of the input [126]. Recent empirical work in functional neuroimaging has shown, however, that many of these standard augmentation techniques do not generalize well to functional neuroimaging data [127]. Instead, neuroimaging researchers advocate for the use of more powerful data synthesis strategies, for example, by the use of generative models trained to capture the characteristics of a training dataset well and which can then be used to synthesize artificial training data [128–130].

DL model performances often also vary highly across the different, often unrecognized, subpopulations of a dataset (a phenomenon known as “hidden stratification”; [131,132]). A DL model trained to decode natural images from functional brain activity might perform well on average, while consistently misclassifying specific image sub-categories. To identify hidden stratification, we generally recommend both manual and automated evaluation approaches, for example, by inspecting falsely classified data instances [132] or applying automated clustering algorithms to the hidden representations of trained DL models to identify possible subpopulations in the data [131]. Similarly, DL models trained on large datasets often learn biases in favor of over-represented sub-populations (e.g., based on individuals’ gender; [133]). To identify these types of

biases in mental state decoding, we recommend evaluating the performance of trained models on the various sub-populations of the data. Once hidden stratification or bias is detected, dedicated learning techniques can be used to improve model performances on specific subpopulations, such as importance weighting [122] or regularization [134].

Lastly, DL models can be susceptible to learning spurious shortcuts that allow them to perform well in a given training dataset but which do not generalize well to other scenarios [135]. For instance, researchers found that a pneumonia detection model trained with medical imaging data can learn to perform well on average solely by learning to identify hospital-specific artifacts in the medical images in addition to learning the hospitals' pneumonia prevalence rates [136]. Similarly, biomarker models trained on functional neuroimaging data can learn to identify patients by their generally increased head motion (as suggested by [137]). To detect these types of confounds, we recommend that neuroimaging researchers evaluate the performance of mental state decoding models on out-of-distribution data (e.g., public neuroimaging data from other laboratories and individuals as provided by OpenNeuro [65]), and that researchers inspect instances of the data whenever out-of-distribution error rates are high relative to in-distribution errors (e.g., with the application of XAI techniques; see Box 1). If confounds are identified in a model's decoding decisions, adaptations of the classical cross-validation procedure, tailored to functional neuroimaging data, can be utilized to obtain an unbiased estimate of decoding performance [137].

Box 4. Recommendations to improve the reproducibility and robustness of DL models in mental state decoding.

The performances of DL models in benchmarks are often difficult to reproduce by other researchers or in new data, as the convergence of DL models (and thereby their final performance) is strongly dependent on many non-deterministic aspects of the training [98,102,108,111]. Further, the resulting highly tuned benchmark performances are often not robust towards the diversity of real-world data [109,110,114]. To avoid these kinds of pitfalls, we provide a set of recommendations to improve the reproducibility and robustness of DL model performances in mental state decoding analyses:

- ❖ Use multiple training runs to estimate a model's expected performance, while randomizing as many non-deterministic aspects of your training pipeline as possible (including random seeds, random weight initializations, and random shufflings of the training data) and using multiple random splits of the data into training, validation, and test datasets (e.g., by the use of bootstrapping or cross-validation) (for methodological details, see [111]).
- ❖ If model comparisons are performed, ensure that the chosen combination of statistical comparison procedure and test dataset size has enough statistical power to accurately determine the studied differences in model performance (e.g., by the use of simple simulation studies; [106]).

- ❖ Evaluate model performances on out-of-distribution data (e.g., by using neuroimaging data from different laboratories and individuals; [148]) and, whenever possible, test for hidden stratification, bias, and confounds [122,132,137] (e.g., by inspecting model performances for the different sub-populations of the data and by inspecting falsely-classified data instances with XAI techniques).
- ❖ Finally, publicly share the resulting models, used data, analysis code, and computing environment (e.g., by the use of containerization with Docker or Singularity) in a dedicated repository (e.g., GitHub or Open Science Framework; [149]).

Concluding remarks

DL models have experienced great success in research and industry and have had major impacts on society [1]. This success has triggered interest in their application to the field of mental state decoding, where researchers aim to characterize the set of mental states that are associated with the activity patterns of different brain regions and can thereby be accurately decoded (i.e., identified) from the activity of these regions. DL models hold a great promise to revolutionize mental state decoding with their unmatched ability to learn versatile representations of complex data. Yet, fully leveraging the potential of DL models in mental state decoding is currently hindered by three main challenges, which result from a general lack of interpretability of DL models as well as difficulties in applying them to small datasets and ensuring their reproducibility and robustness.

Here, we have provided a detailed discussion of these three challenges and proposed a set of solutions that are informed by recent advances in functional neuroimaging and DL research. In sum, we recommend that researchers utilize XAI techniques to identify the mapping between mental states and brain activity that a DL model has learned (Box 1-2), improve the performance of DL models in conventional neuroimaging datasets by pre-training these models on public neuroimaging data (Box 3), and follow specific recommendations to improve the reproducibility and robustness of DL model performances in mental state decoding (Box 4). We hope that researchers will take inspiration from our discussion and explore the many open research questions that remain on the path to determining whether DL models can live up to their promise for mental state decoding (see Outstanding Questions).

Outstanding Questions

- The mappings learned by a DL model between input data and target signals can be highly complex and counterintuitive. Given this complexity, what are the limits of current XAI techniques, which often simplify the model's decision process to allow for interpretability, in providing insights into a model's learned mapping between brain activity and mental states?
- Can data programming be used to effectively generate standardized labels of mental states for public neuroimaging datasets (e.g., according to the Cognitive Atlas) and if so, how do models trained with these generated labels compare to models trained with self-supervision?
- Which kinds of simple data augmentation techniques (akin to adding noise or occluding parts of an input) can help improve the robustness of DL models trained with functional neuroimaging data?
- How can functional neuroimaging researchers provide easy access to (and use of) their pre-trained DL models (e.g., to enable others to easily adapt these models to their collected datasets)?
- Can benchmarks be a useful tool for functional neuroimaging research to accelerate the development of accurate and versatile DL models, when taking the appropriate measures to ensure reproducibility and robustness?

Acknowledgments

Armin W. Thomas is supported by Stanford Data Science through the Ram and Vijay Shriram Data Science Fellowship. Russell A. Poldrack is supported by the National Science Foundation under Grant No. OAC-1760950. Christopher Ré gratefully acknowledges the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Salesforce, Total, the HAI-AWS Cloud Credits for Research program, Stanford Data Science, and members of the Stanford DAWN project: Facebook, Google, and VMWare. The Mobilize Center is a Biomedical Technology Resource Center, funded by the NIH National Institute of Biomedical Imaging and Bioengineering through Grant P41EB027060. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those

of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

References

1 Goodfellow, I. *et al.* (2016) *Deep Learning*, MIT Press.

2 Livezey, J.A. and Glaser, J.I. (2021) Deep learning approaches for neural decoding across architectures and recording modalities. *Brief. Bioinform.* 22, 1577–1591

3 Norman, K.A. *et al.* (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430

4 Poldrack, R.A. *et al.* (2015) Long-term neural and physiological phenotyping of a single human. *Nat. Commun.* 6, 8885

5 Tavor, I. *et al.* (2016) Task-free MRI predicts individual differences in brain activity during task performance. *Science* 352, 216–220

6 Cole, M.W. *et al.* (2014) Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron* 83, 238–251

7 Samek, W. *et al.* (2021) Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* 109, 247–278

8 Doshi-Velez, F. and Kim, B. (2017) Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv170208608 Cs Stat* at <http://arxiv.org/abs/1702.08608>

9 Montavon, G. *et al.* (2018) Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73, 1–15

10 Simonyan, K. *et al.* (2014) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv13126034 Cs* at <http://arxiv.org/abs/1312.6034>

11 Bach, S. *et al.* (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* 10, e0130140

12 Smilkov, D. *et al.* (2017) SmoothGrad: removing noise by adding noise. *ArXiv170603825 Cs Stat* at <http://arxiv.org/abs/1706.03825>

13 Sundararajan, M. *et al.* (2017) Axiomatic Attribution for Deep Networks. in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328

14 Shrikumar, A. *et al.* (2017) Learning Important Features Through Propagating Activation Differences. in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–3153

15 Ribeiro, M.T. *et al.* (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144

16 Zeiler, M.D. and Fergus, R. (2014) Visualizing and Understanding Convolutional Networks. in *Computer Vision – ECCV 2014*, pp. 818–833

17 Lundberg, S.M. and Lee, S.-I. (2017) A unified approach to interpreting model predictions. in *Advances in Neural Information Processing System* 30, pp. 4768–4777

18 Thomas, A.W. *et al.* (2019) Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. *Front. Neurosci.* 13, 1321

19 Wang, X. *et al.* (2020) Decoding and mapping task states of the human brain via deep

605 learning. *Hum. Brain Mapp.* 41, 1505–1519
606 20 Dinsdale, N.K. *et al.* (2021) Learning patterns of the ageing brain in MRI using deep
607 convolutional networks. *NeuroImage* 224, 117401
608 21 Oh, K. *et al.* (2019) Classification and Visualization of Alzheimer's Disease using
609 Volumetric Convolutional Neural Network and Transfer Learning. *Sci. Rep.* 9, 18150
610 22 Thomas, A.W. *et al.* (2021) Evaluating deep transfer learning for whole-brain cognitive
611 decoding. *ArXiv211101562 Cs Q-Bio* at <http://arxiv.org/abs/2111.01562>
612 23 Kohoutová, L. *et al.* (2020) Toward a unified framework for interpreting machine-learning
613 models in neuroimaging. *Nat. Protoc.* 15, 1399–1435
614 24 Zhang, Y. *et al.* (2022) Deep learning models of cognitive processes constrained by human
615 brain connectomes. *Medical Image Analysis* 80, 102507
616 25 Hu, J. *et al.* (2021) Deep Learning-Based Classification and Voxel-Based Visualization of
617 Frontotemporal Dementia and Alzheimer's Disease. *Front. Neurosci.* 14, 626154
618 26 Zhang, T. *et al.* (2020) Separated Channel Attention Convolutional Neural Network (SC-
619 CNN-Attention) to Identify ADHD in Multi-Site Rs-fMRI Dataset. *Entropy* 22, 893
620 27 Lin, K.-Y. *et al.* (2021) Classification and Visualization of Chemotherapy-Induced
621 Cognitive Impairment in Volumetric Convolutional Neural Networks. *J. Pers. Med.* 11,
622 1025
623 28 Choi, H. *et al.* (2022) Subgroups of Eating Behavior Traits Independent of Obesity Defined
624 Using Functional Connectivity and Feature Representation Learning. *bioRxiv.* DOI:
625 10.1101/2019.12.11.123456
626 29 Supekar, K. *et al.* (2022) Deep learning identifies robust gender differences in functional
627 brain organization and their dissociable links to clinical symptoms in autism. *Br. J.*
628 *Psychiatry* 220, 202–209
629 30 Gupta, S. *et al.* (2019) Decoding Brain Functional Connectivity Implicated in AD and MCI.
630 in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp.
631 781–789
632 31 McClure, P. *et al.* (2020) Improving the Interpretability of fMRI Decoding using Deep
633 Neural Networks and Adversarial Robustness. *ArXiv200411114 Cs Q-Bio Stat* at
634 <http://arxiv.org/abs/2004.11114>
635 32 Samek, W. *et al.* (2017) Evaluating the Visualization of What a Deep Neural Network Has
636 Learned. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 2660–2673
637 33 Kindermans, P.-J. *et al.* (2019) The (Un)reliability of Saliency Methods. In *Explainable AI:*
638 *Interpreting, Explaining and Visualizing Deep Learning* (Samek, W. et al., eds), pp. 267–
639 280, Springer International Publishing
640 34 Adebayo, J. *et al.* (2018) Sanity checks for saliency maps. in *Advances in Neural*
641 *Information Processing Systems* 31, pp. 9525–9536
642 35 Jacovi, A. and Goldberg, Y. (2020) Towards Faithfully Interpretable NLP Systems: How
643 Should We Define and Evaluate Faithfulness? in *Proceedings of the 58th Annual Meeting*
644 *of the Association for Computational Linguistics*, pp. 4198–4205
645 36 Thomas, A.W. *et al.* (2022) Comparing interpretation methods in mental state decoding
646 analyses with deep learning models. *ArXiv* 220515581 *Cs Q-Bio* at
647 <https://arxiv.org/abs/2205.15581>
648 37 Hanson, S.J. *et al.* (2004) Combinatorial codes in ventral temporal lobe for object
649 recognition: Haxby (2001) revisited: is there a "face" area? *NeuroImage* 23, 156–166
650 38 Dadi, K. *et al.* (2020) Fine-grain atlases of functional modes for fMRI analysis.

NeuroImage 221, 117126
39 Schaefer, A. *et al.* (2018) Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb. Cortex* 28, 3095–3114
40 Chang, L.J. *et al.* (2015) A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biol.* 13, e1002180
41 Koban, L. *et al.* (2019) Different brain networks mediate the effects of social and conditioned expectations on pain. *Nat. Commun.* 10, 4096
42 Richards, B.A. *et al.* (2019) A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770
43 Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215
44 Friston, K.J. *et al.* (1994) Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* 2, 189–210
45 Grosenick, L. *et al.* (2013) Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* 72, 304–321
46 Yarkoni, T. *et al.* (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670
47 Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* at <http://arxiv.org/abs/1409.1556>
48 Devlin, J. *et al.* (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv181004805 Cs* at <http://arxiv.org/abs/1810.04805>
49 Brown, T. *et al.* (2020) Language Models are Few-Shot Learners. in *Advances in Neural Information Processing Systems* 33, pp. 1877–1901
50 Allen-Zhu, Z. *et al.* (2019) A Convergence Theory for Deep Learning via Over-Parameterization. in *Proceedings of the 36th International Conference on Machine Learning*, pp. 242–252
51 Raffel, C. *et al.* (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1–67
52 Sun, C. *et al.* (2017) Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852
53 Krogh, A. and Hertz, J.A. (1992) A Simple Weight Decay Can Improve Generalization. in *Advances in Neural Information Processing Systems* 4, pp. 950–957
54 Srivastava, N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958
55 Pan, S.J. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359
56 Deng, J. *et al.* (2009) ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255
57 Gao, L. *et al.* (2020) The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv210100027 Cs* at <https://arxiv.org/abs/2101.00027>
58 Deepak, S. and Ameer, P.M. (2019) Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* 111, 103345
59 Bengio, Y. *et al.* (2006) Greedy Layer-Wise Training of Deep Networks. in *Advances in Neural Information Processing Systems* 19
60 Erhan, D. *et al.* (2010) Why Does Unsupervised Pre-training Help Deep Learning? in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and*

Statistics, pp. 201–208
61 Kolesnikov, A. *et al.* (2020) Big Transfer (BiT): General Visual Representation Learning. in Computer Vision – ECCV 2020 (Vedaldi, A., Bischof, H., Brox, T., Frahm, JM., eds), pp. 491–507. Springer, Cham
62 He, K. *et al.* (2019) Rethinking ImageNet Pre-Training. in *2019 IEEE/CVF International Conference on Computer Vision,* pp. 4917–4926
63 Horien, C. *et al.* (2021) A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat. Hum. Behav.* 5, 185–193
64 Gorgolewski, K.J. *et al.* (2016) The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 160044
65 Markiewicz, C.J. *et al.* (2021) The OpenNeuro resource for sharing of neuroscience data. *eLife* 10, e71774
66 Esteban, O. *et al.* (2019) fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116
67 Hebling Vieira, B. *et al.* (2021) A deep learning based approach identifies regions more relevant than resting-state networks to the prediction of general intelligence from resting-state fMRI. *Hum. Brain Mapp.* 42, 5873–5887
68 Mahmood, U. *et al.* (2019) Transfer Learning of fMRI Dynamics. *ArXiv191106813 Cs Eess Stat* at <http://arxiv.org/abs/1911.06813>
69 Koyamada, S. *et al.* (2015) Deep learning of fMRI big data: a novel approach to subject-transfer decoding. *ArXiv150200093 Cs Q-Bio Stat* at <http://arxiv.org/abs/1502.00093>
70 Zheng, W.-L. and Lu, B.-L. (2016) Personalizing EEG-based affective models with transfer learning. in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2732–2738
71 Bazeille, T. *et al.* (2021) An empirical evaluation of functional alignment using inter-subject decoding. *NeuroImage* 245, 118683
72 Li, H. *et al.* (2018) A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. *Front. Neurosci.* 12, 491
73 He, T. *et al.* (2022) Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* 25, 795–804
74 Zhang, Y. *et al.* (2021) Functional annotation of human cognitive states using deep graph convolution. *NeuroImage* 231, 117847
75 Wang, X. *et al.* (2020) Decoding and mapping task states of the human brain via deep learning. *Hum. Brain Mapp.* 41, 1505–1519
76 Nguyen, S. *et al.* (2020) Attend and Decode: 4D fMRI Task State Decoding Using Attention Models. in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, pp. 267–279
77 Thomas, A.W. *et al.* (2019) Deep Transfer Learning for Whole-Brain FMRI Analyses. in *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, pp. 59–67, Springer, Cham
78 Zhang, H. *et al.* (2018) Transfer Learning on fMRI Datasets. in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 595–603
79 Yousefnezhad, M. *et al.* (2020) Shared Space Transfer Learning for analyzing multi-site fMRI data. in *Advances in Neural Information Processing Systems* 33, pp. 15990-16000
80 Mensch, A. *et al.* (2021) Extracting representations of cognition across neuroimaging studies improves brain decoding. *PLOS Comput. Biol.* 17, e1008795

81 Zhou, S. *et al.* (2019) Improving Whole-Brain Neural Decoding of fMRI with Domain Adaptation. in *Machine Learning in Medical Imaging* (Suk, HI. et al., eds.), pp. 265–273, Springer, Cham

82 Turner, J.A. and Laird, A.R. (2012) The cognitive paradigm ontology: design and application. *Neuroinformatics* 10, 57–66

83 Poldrack, R.A. *et al.* (2011) The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Front. Neuroinformatics* 5,

84 Wechsler, D. (2008) Wechsler Adult Intelligence Scale--Fourth Edition. *Archives of Clinical Neuropsychology*.

85 Caruana, R. (1997) Multitask Learning. *Mach. Learn.* 28, 41–75

86 Zhou, Z.-H. (2018) A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* 5, 44–53

87 Ratner, A. *et al.* (2016) Data Programming: Creating Large Training Sets, Quickly. in *Advances in Neural Information Processing Systems* 29, pp. 3567–3575

88 Dunnmon, J.A. *et al.* (2020) Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *Patterns* 1, 100019

89 Bengio, Y. *et al.* (2013) Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828

90 Liu, X. *et al.* (2021) Self-supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* DOI: 10.1109/TKDE.2021.3090866

91 Chen, T. *et al.* (2020) A Simple Framework for Contrastive Learning of Visual Representations. in *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607

92 Hinton, G.E. *et al.* (2006) A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554

93 Hinton, G.E. and Zemel, R. (1993) Autoencoders, Minimum Description Length and Helmholtz Free Energy. in *Advances in Neural Information Processing Systems* 6

94 Thomas, A.W. *et al. (2022)* Self-Supervised Learning Of Brain Dynamics From Broad Neuroimaging Data. *ArXiv220611417 Cs Q-Bio* at <https://arxiv.org/abs/2206.11417>

95 Botvinik-Nezer, R. *et al.* (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 84–88

96 Bouthillier, X. *et al.* (2019) Unreproducible Research is Reproducible. in *Proceedings of the 36th International Conference on Machine Learning*, pp. 725–734

97 Gorman, K. and Bedrick, S. (2019) We Need to Talk about Standard Splits. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2786–2791

98 Henderson, P. *et al.* (2018) Deep Reinforcement Learning That Matters. in *Proceedings of the AAAI Conference on Artificial Intelligence*, 32

99 Lipton, Z.C. and Steinhardt, J. (2018) Troubling Trends in Machine Learning Scholarship. *ArXiv180703341 Cs Stat* at <http://arxiv.org/abs/1807.03341>

100 Ethayarajh, K. and Jurafsky, D. (2020) Utility is in the Eye of the User: A Critique of NLP Leaderboards. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4846–4853

101 Simmons, J.P. *et al.* (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* 22, 1359–1366

102 Lucic, M. *et al.* (2018) Are GANs Created Equal? A Large-Scale Study. in *Advances in Neural Information Processing Systems* 31

103 Reimers, N. and Gurevych, I. (2017) Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348

104 Melis, G. *et al.* (2017) On the State of the Art of Evaluation in Neural Language Models. *ArXiv170705589 Cs* at <http://arxiv.org/abs/1707.05589>

105 Raff, E. (2019) A Step Toward Quantifying Independently Reproducible Machine Learning Research. in *Advances in Neural Information Processing Systems* 32

106 Card, D. *et al.* (2020) With Little Power Comes Great Responsibility. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9263–9274

107 Button, K.S. *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376

108 Dodge, J. *et al.* (2019) Show Your Work: Improved Reporting of Experimental Results. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2185–2194

109 Goel, K. *et al.* (2021) Robustness Gym: Unifying the NLP Evaluation Landscape. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pp. 42–55

110 Kiela, D. *et al.* (2021) Dynabench: Rethinking Benchmarking in NLP. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124

111 Bouthillier, X. *et al.* (2021) Accounting for Variance in Machine Learning Benchmarks. in *Proceedings of Machine Learning and Systems*, 3, pp. 747–769

112 Varoquaux, G. (2018) Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180, 68–77

113 Thompson, W.H. *et al.* (2020) Dataset decay and the problem of sequential analyses on open datasets. *eLife* 9, e53498

114 Koh, P.W. *et al.* (2021) WILDS: A Benchmark of in-the-Wild Distribution Shifts. in *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664

115 Belinkov, Y. and Bisk, Y. (2018) Synthetic and Natural Noise Both Break Neural Machine Translation. *ArXiv171102173 Cs* at <http://arxiv.org/abs/1711.02173>

116 Szegedy, C. *et al.* (2014) Intriguing properties of neural networks. *ArXiv13126199 Cs* at <http://arxiv.org/abs/1312.6199>

117 Moosavi-Dezfooli, S.-M. *et al.* (2016) DeepFool: a simple and accurate method to fool deep neural networks. *ArXiv151104599 Cs* at <http://arxiv.org/abs/1511.04599>

118 Liu, T.T. (2016) Noise contributions to the fMRI signal: An overview. *NeuroImage* 143, 141–151

119 Kragel, P.A. *et al.* (2018) Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* 21, 283–289

120 Dubois, J. and Adolphs, R. (2016) Building a Science of Individual Differences from fMRI. *Trends Cogn. Sci.* 20, 425–443

121 Van Oudenhove, L. *et al.* (2020) Common and distinct neural representations of aversive somatic and visceral stimulation in healthy individuals. *Nat. Commun.* 11, 5939

122 Dockès, J. *et al.* (2021) Preventing dataset shift from breaking machine-learning

biomarkers. *GigaScience* 10, giab055
123 Traut, N. *et al.* (2022) Insights from an autism imaging biomarker challenge: Promises and threats to biomarker discovery. *NeuroImage* 255, 119171
124 Varoquaux, G. and Cheplygina, V. (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digit. Med.* 5, 1–8
125 He, K. *et al.* (2016) Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778
126 DeVries, T. and Taylor, G.W. (2017) Improved Regularization of Convolutional Neural Networks with Cutout. *ArXiv170804552 Cs* at <http://arxiv.org/abs/1708.04552>
127 Jönemo, J. *et al.* (2021) Evaluation of augmentation methods in classifying autism spectrum disorders from fMRI data with 3D convolutional neural networks. *ArXiv211010489 Cs Eess Q-Bio* at <http://arxiv.org/abs/2110.10489>
128 Tajini, B. *et al.* (2021) Functional Magnetic Resonance Imaging Data Augmentation Through Conditional ICA. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 491–500
129 Zhuang, P. *et al.* (2019) FMRI Data Augmentation Via Synthesis. in *2019 IEEE 16th International Symposium on Biomedical Imaging*, pp. 1783–1787
130 Qiang, N. *et al.* (2021) Modeling and augmenting of fMRI data using deep recurrent variational auto-encoder. *J. Neural Eng.* 18, 0460b6
131 Sohoni, N. *et al.* (2020) No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. in *Advances in Neural Information Processing Systems* 33, pp. 19339–19352
132 Oakden-Rayner, L. *et al.* (2020) Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. in *Proceedings of the ACM Conference on Health, Inference, and Learning - CHIL '20*, pp. 151–159
133 Bommasani, R. *et al.* (2021) On the Opportunities and Risks of Foundation Models. *ArXiv210807258 Cs* at <http://arxiv.org/abs/2108.07258>
134 Sagawa, S. *et al.* (2020) Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ArXiv191108731 Cs Stat* at <http://arxiv.org/abs/1911.08731>
135 Geirhos, R. *et al.* (2020) Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 665–673
136 Zech, J.R. *et al.* (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* 15, e1002683
137 Chyzhyk, D. *et al.* (2022) How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* 11, giac014
138 Fong, R.C. and Vedaldi, A. (2017) Interpretable Explanations of Black Boxes by Meaningful Perturbation. in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437
139 Zurada, J.M. *et al.* (1994) Sensitivity analysis for minimization of input data dimension for feedforward neural network. in *Proceedings of IEEE International Symposium on Circuits and Systems - ISCAS '94*, 6, pp. 447–450
140 Montavon, G. *et al.* (2019) Layer-Wise Relevance Propagation: An Overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Samek, W. et al., eds.), pp. 193–209, Springer
141 Shapley, L.S. (1952) *A Value for N-Person Games*, RAND Corporation.

142 Springenberg, J.T. *et al.* (2015) Striving for Simplicity: The All Convolutional Net. *ArXiv14126806 Cs* at <http://arxiv.org/abs/1412.6806>
143 Alber, M. *et al.* (2019) iNNvestigate Neural Networks! *J. Mach. Learn. Res.* 20, 1–8
144 Anders, C.J. *et al.* (2021) Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. *ArXiv210613200 CS* at <https://arxiv.org/abs/2106.13200>
145 Sturmfels, P. *et al.* (2020) Visualizing the Impact of Feature Attribution Baselines. *Distill* 5, e22
146 Kokhlikyan, N. *et al.* (2020) Captum: A unified and generic model interpretability library for PyTorch. *ArXiv200907896 Cs Stat* at <http://arxiv.org/abs/2009.07896>
147 Shrikumar, A. *et al.* (2017) Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *ArXiv160501713 Cs* at <http://arxiv.org/abs/1605.01713>
148 Rosenberg, M.D. and Finn, E.S. (2022) How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* DOI: 10.1038/s41593-022-01110-9
149 Foster, E.D. and Deardorff, A. (2017) Open Science Framework (OSF). *J. Med. Libr. Assoc. JMLA* 105, 203–206