CellPress

Trends in Cognitive Sciences

Opinion

Mental state decoders: game-changers or wishful thinking?

Andrew D. Vigotsky [1,*], Gian Domenico Iannetti [2,3], and A. Vania Apkarian [1]

Decoding mental and perceptual states using fMRI has become increasingly popular over the past two decades, with numerous highly-cited studies published in high-profile journals. Nevertheless, what have we learned from these decoders? In this opinion, we argue that fMRI-based decoders are not neurophysiologically informative and are not, and likely cannot be, applicable to real-world decision-making. The former point stems from the fact that decoding models cannot disentangle neural mechanisms from their epiphenomena. The latter point stems from both logical and ethical constraints. Constructing decoders requires precious time and resources that should instead be directed toward scientific endeavors more likely to yield meaningful scientific progress.

Highlights

Many fMRI papers in the mind-reading literature present decoders as a final product, arguing that decoders provide neurophysiological insight and have real-world utility.

However, mental state decoders are commonly built in a way that precludes straightforward physiological interpretations. This undermines the claim that decoders are interpretable or capture 'representations' of mental states.

In contrast to decoding models, encoding models of task fMRI are computationally straightforward and more interpretable.

Mind-reading research would benefit from shifting its focus from successful decoding *per se* to understanding how decoding is affected by different experimental parameters, which would demonstrate the information that decoders are sensitive to (e.g., the color of a banana versus its orientation).

The rise of decoding models in cognitive neuroscience

fMRI revolutionized neuroscience, facilitating the non-invasive study of brain-wide neuronal activity via blood oxygenation level-dependent (BOLD) signals [1]. Thousands of papers investigating the BOLD signal changes during various tasks have been published (reviewed in [2]). The predominant paradigm entails statistically modeling the BOLD signal (the dependent variable, DV) as a function of some aspect of the task, such as stimulus timing or intensity (the independent variable, IV) [3]. **Encoding models** (see Glossary) are consistent with fMRI experimental designs since researchers treat brain activity as something that is measured (i.e., a DV) rather than controlled (i.e., an IV). From encoding models, researchers infer the dependence of local BOLD signal changes on the task. By stark contrast, **decoding models** are a different analytical approach that has become popular over the past two decades. Decoding is discussed and studied broadly in neuroscience, engineering, and information theory. Here, we narrow our discussion to the use of decoding in human fMRI. Decoding models use BOLD signals to infer task parameters or mental states (Box 1). These models flip the standard BOLD analytical model: the brain activity becomes the IV (or model input), and the DV (or model output) is some aspect of the task (e.g., certain stimulus properties, a perceptual state, etc.) [4]. Given that encoding relies on an fMRI signal that is empirically measured, it has been termed 'forward inference', while decoding has been termed 'reverse inference' [5–8]. Decoders that reliably reconstruct some characteristic of the task or an individual's mental state from brain activity are typically claimed to approximate 'mind-reading' [9].

Mental state decoding was popularized in the human neuroimaging literature through two principal lofty claims. The first is that decoders facilitate the discovery of the neural underpinning of mental states. Considering the mind-decoding literature, this claim may be interpreted as though decoders (i) reflect the neural codes that give rise to mental states, or (ii) efficiently uncover 'information' to provide a basis for future exploration [9,10,12]. The second claim is that decoders are objective biomarkers of subjective experiences and, thus, have valuable real-world utility in the clinic, courts, and beyond. Here, we critically discuss whether decoders have fulfilled, or can fulfill, either of these two claims.

[1]Northwestern University, Chicago, IL, USA
[2]Italian Institute of Technology (IIT), Rome, Italy
[3]University College London (UCL), London, UK

*Correspondence: vigotsky@u.northwestern.edu (A. D. Vigotsky).

Box 1. Decoder construction and evaluation

Traditionally, constructing a decoder from a task fMRI experiment begins with fitting an encoding model. This is done by regressing BOLD time series of every voxel or a group of voxels onto the time series of the task. Thus, the encoding model produces a brain activity map for each task vector; each voxel or group of voxels is represented by a parameter estimate of its task-related brain activity. These brain activity maps, often termed 'beta maps' in the fMRI literature, serve as the basis for decoding.

Once these brain activity maps are constructed, they, or specific regions of interest from them, are used as predictors in decoding models. These models intend to answer the question: 'Given the observed brain activity, what was the task (or the percept)?' This prediction model is the 'decoder'. If these decoders are reliably predictive, then they should generalize to brain activity maps (observations) that were not used to train the decoder. Thus, decoders are tested using out-of-sample brain activity maps, with which their performance is quantified (e.g., using accuracy, area under the receiver operating characteristic curve, correlations, etc.). Dichotomous declarations of decoding 'success' typically depend on how out-of-sample predictions outperform a null model.

For more details on decoder construction and evaluation, we refer to previous literature [10] and software toolboxes [11].

Much has been written about encoding and decoding in human neuroimaging (e.g., [5,6,13–15]). However, the bulk of these papers focus on the technicalities of the decoding approaches and/or the obtained results, while skirting the higher-level questions of whether and how decoders tangibly advance neuroscience. In this opinion, we directly address the aforementioned higher-level questions and contend that fMRI-based decoders are not neurophysiologically informative and are not, and likely cannot be, applicable to real-world decision-making.

Decoders have no intrinsic mechanistic value

After constructing and evaluating a decoder, it is common to interpret the decoding model itself. For example, some groups have advocated for analyzing the structure of the decoders and performing hypothesis tests on their constituent components (e.g., voxels or regions) [16], with the assumption that reliable predictive power yields more interpretable and mechanistically meaningful findings. In this case, fMRI decoders are said to uncover the 'neural codes' or 'representations' of mental states [9,12]. Despite the ambiguity of these terms [17,18], this claim appears to imply that the structure of the decoding model (its weights and spatial arrangement) somehow represents the neural processes that cause the inferred mental state, behavioral outcome, or task parameter. If decoders can uncover neural processes, then such insight would indeed be neurophysiologically interesting and allow us to infer general principles from decoding findings that allow for predictions in new contexts (e.g., the calculations performed by a specific brain region). Here, we show that this claim is incorrect and provide multiple lines of reasoning and evidence delineating the epistemic boundaries of decoding.

The decoder's dictum

Successful decoders rely on patterns of BOLD activity to distinguish between mental states. However, what do these patterns of neural activity really mean or capture? A common interpretation is the decoder's dictum, 'If information can be decoded from patterns of neural activity, then this provides strong evidence about what information those patterns represent' [19]. Despite its appeal, it has been cogently argued that the decoder's dictum is false and that successful decoding does not provide reasonable grounds for inference concerning the patterns used to decode [19]. The decoder's dictum is false for several reasons, but principally because we do not know which information is being decoded. Empirically, this uncertainty partly arises from the inability of fMRI to resolve neural activity at the level of local neuronal populations, such as the size of cortical columns [8,19]. Theoretically, there is no reason to believe that decoding models capture neural codes, partly because the models have no biological basis or theoretical constraints (see section 'Model specification precludes sensible inferences'). Here, we build on

Glossary

Counterfactual: scenario that did not occur and that may have led to a different outcome. Counterfactual reasoning can provide a framework for thinking about causality. For example, if a patient improves after taking a drug, one must consider the counterfactual: what would have happened if the patient did not take the drug? Unless we know the answer to this counterfactual question, we cannot determine the causal effect of that drug.
Decoding model: statistical model in task-based fMRI studies in which brain activity predicts some aspect of the task (e.g., which visual stimulus was presented to the participant).
Direct effect: extent to which the dependent variable is influenced by the independent variable after removing the effect of mediators.
Directed acyclic graph (DAG): graph that depicts a causal hypothesis of how different variables influence one another, without any closed loops.
Encoding model: statistical model in task-based fMRI studies in which brain activity is regressed onto a task vector, specified by the experimenter, such as the timing and duration of a visual stimulus presented to the participant.
Explanatory model: reason or statement concerning how phenomena occur. Explanatory research aims to understand phenomena, especially their causes, with the aid of explanatory models that strongly rely on theory to derive causal hypotheses.
Predictive model: forecast of future or unseen observations. Predictive research constructs predictive models that aim to accurately forecast (e.g., classify or estimate the value of) future observations based on observable features. This approach does not aim to uncover causal mechanisms.
Table 2 fallacy: tendency to interpret all regression parameters as total effects. This is a fallacy since it ignores causal hypotheses that imply that some regression parameters are total effects, while other regression parameters are direct effects.
Total effect: extent to which the dependent variable is influenced by the independent variable. This quantity includes the influence of potential mediators.

previous arguments against the decoder's dictum by providing stronger empirical support and a mathematical contextualization.

Decoders are too complex for causal inference

Statistical modeling has at least two distinct goals: explanation and prediction. **Explanatory models** evaluate causal hypotheses, while **predictive models** forecast new or future observations [20]. Notably, explanatory models do not necessarily rely on experimental manipulations to evaluate causal hypotheses. Rather, they strongly rely on theory to hypothesize and model a data-generating process, which makes assumptions about **counterfactuals** [21,22]. However, decoders are not developed as explanatory models since they lack strong, theory-driven hypotheses about counterfactuals [e.g., what would have happened had a region of interest (ROI) activated differently?]. This is perhaps understandable, because explanatory models may be intractable due to the sheer number of predictors (e.g., tens of thousands of voxels, tens or hundreds of ROIs, etc.) and the complexity of the brain. In other words, there are simply too many predictors to reason about and explicitly propose a causal structure. Instead of relying on explanatory modeling, decoders are more closely related to predictive modeling.

Decoders predict some aspect of the task based on tens, hundreds, or thousands of predictors derived from brain activation data. These decoders are not standardized across or even within studies. For instance, some studies even fit and contrast several decoders before examining the model that performs best (e.g., [23]) rather than studying a few constrained explanatory models derived from specific hypotheses. Since decoders aim to predict without committing to a theory-informed hypothesis, they are predictive rather than explanatory models. Predictive modeling is strictly concerned with accurate forecasting (e.g., predicting the task from brain activation) and not whether the model is aligned with theory or the experimental protocol. As a result, predictive models may yield good predictions, but the model parameters may be entirely inconsistent with theory and the true data-generating mechanism [20,24].

Unfortunately, the predictive models used to decode cannot be interpreted causally. In the epidemiology literature, the blind causal interpretation of covariates is called the **Table 2 fallacy** [25], since model results are typically presented as Table 2 in epidemiology papers. The Table 2 fallacy was also recently discussed in the context of human neuroimaging [26]. This fallacy was originally described using an example concerning how age, HIV, and smoking influence the risk of stroke [25]. From a multivariable logistic regression, the effect of HIV on stroke risk has a fundamentally different interpretation compared with the effect of smoking on stroke risk since the HIV effect may partly mediate that of smoking. This can occur if, for example, immunodepression due to smoking increases HIV risk. Specifically, the parameter for HIV would be considered a **total effect**, representing the entire contribution of HIV to the risk of stroke. By contrast, the parameter for smoking would be considered a **direct effect**, that is, the effect of smoking after removing its indirect effect, attributable to the smoking→HIV→stroke pathway. We present below a similar analogy in the neuroimaging field.

Mesulam's description of sensory-fugal gradients of information flow can be viewed as a **directed acyclic graph (DAG)** (Figure 1) [27]. Since DAGs depict causal hypotheses, they provide a framework to aid the interpretation of model parameters. The DAG structure constrains the interpretation of multivariable regression coefficients in neuroimaging, such as the parameters in a multivoxel or multi-ROI prediction model. For example, suppose that a decoder predicts motor responses to an auditory stimulus using 'downstream unimodal' and 'paralimbic' regions (Figure 1). Since paralimbic activation contains information from downstream unimodal areas, a model including both regions would estimate a direct effect of downstream unimodal activation

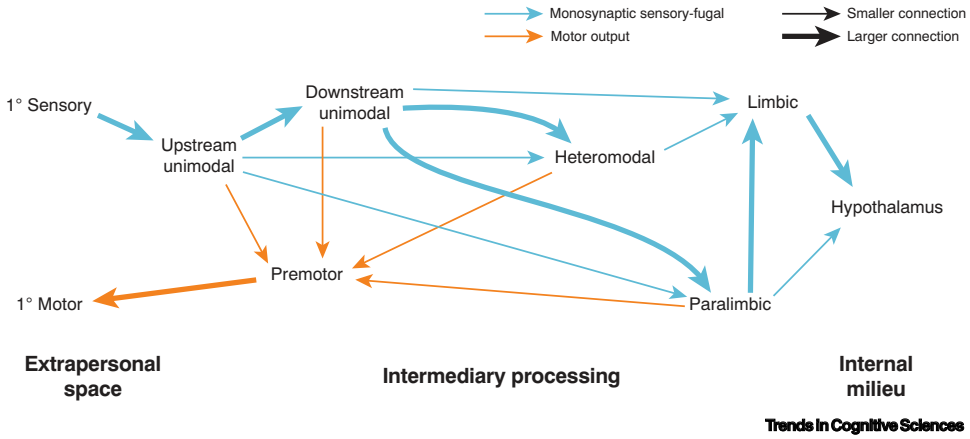

Figure 1. Auditory and visual sensory-fugal gradients of connectivity can be viewed as a causal model of information flow in the brain. Sensory information flows along several parallel and interacting pathways. Since the arrows represent the putative sequence of information flow, this can be interpreted as a causal model, facilitating the rigorous evaluation of causal assumptions and, thus, interpretations. For instance, if one were to build a decoder with the depicted regions, one can constrain one's analysis and inferences based on the assumed information flow. By doing so, one may choose to exclude limbic regions and hypothalamus from one's model, while appreciating that the premotor parameter has a different interpretation compared with the paralimbic parameter. Constrained theory-based hypotheses make neural assumptions more explicit to aid inferences but may be unwieldy for many problems. Adapted from [27].

rather than a total effect. Thus, decoder parameters from different regions have distinct interpretations that depend on the presumed causal structure, precluding straightforward interpretations.

Decoders that are not constrained by hypotheses are likely to produce parameters that are causally vacuous. For example, consider an unconstrained decoder that uses 'limbic' activation to decode a task. Although limbic activation may not contribute to the output (1° Motor), it may contain information from regions that do contribute, such as paralimbic, downstream unimodal, and heteromodal areas, leading investigators to the wrong conclusion (Figure 1). Similarly, decoders can exploit noise to improve predictions. For instance, suppose there are two regions: region A contains a task-relevant signal and noise, while region B contains no task-relevant signal but the same noise as region A. A decoder could use the measured data from region B to extract the signal of interest from region A (e.g., by subtracting the measured data from region B from that of region A). This would not suggest that region B has any signal of interest, let alone an inhibitory effect; rather, it simply implies that adjusting for the noise present in region B can improve the predictions afforded by region A [28,29]. These interpretive issues would compound as more predictors are included in the model. Rather than drawing inferences concerning the information contained within each predictor, decoders draw inferences on the level of the entire set of predictors.

These causal inferential barriers of mental state decoders also stem from the extremely high dimensionality of fMRI data. Unless investigators have a massive sample (subjects, repetitions, etc.) or select a small subset of voxels or ROIs, decoders will often be derived from data with more predictors than samples. Such data are degenerate (i.e., they can be fully captured in a lower dimension space), precluding an identifiable causal structure since the role of each variable cannot be distinguished [30]. Therefore, the inverse problem intrinsic to mental state decoding is ill-posed, unless several assumptions concerning the causal neural structure are made.

Causality is at the heart of scientific research, at least when it attempts to elucidate how phenomena arise. Since mental state decoding is not performed under a causal framework, either

experimental or analytical, but instead tries to maximize prediction accuracy, it follows that decoding is a tool for prediction rather than explanation.

Model specification precludes sensible inferences

Admittedly, researchers may not be interested in interpreting decoders as causal models. Indeed, much of the decoding literature skirts this issue, opting to rely on ambiguous language, such as 'neural codes', 'representations', 'signature', or 'information' [17,31,32] instead of forgoing inference. No matter how one interprets these terms, a more important fact makes decoders mechanistically uninteresting: decoders are arbitrary and not unique.

Decoder arbitrariness can be clearly illustrated when considering the multiplicity of model-fitting procedures. Consider, for example, that decoding models often incorporate dimensionality reduction or regularization. Each procedure yields a different loss function and, thus, model parameters. For instance, logistic regression with L^1 regularization will produce a different model compared with logistic regression with L^2 regularization. The non-uniqueness of decoders is not just a theoretical concern; there are many examples of such non-uniqueness in practice. For instance, in the cognition literature, one effort to improve decoding performance yielded only marginal improvements relative to more 'standard' decoders that already consistently outperformed chance [33]. In a more extreme example, consider the mental state decoders depicted in Figure 2, which shows three distinct 'pain decoders' with similar average performance. If one of these decoders truly represented the neural code underlying pain, then it should outperform the others. Instead, there is no clear winner. Similar examples can be found in high-impact publications in the pain literature. In one notable case, investigators trained 5916 models and focused on a single 'best' model despite many tested candidate models performing similarly [23]. It remains entirely mysterious how the chosen model could be deemed 'best' based on point estimates of predictive performance. In another highly debated example, researchers combined LASSO with principal components logistic regression to derive a decoder labeled the 'Neurologic Pain Signature' (NPS) [34]. However, the authors themself stated that the 'predictions and accuracy levels of NPS were nearly identical with [support vector regression] in all cases' [34], meaning that NPS is not uniquely identified.

One may argue that decoders can be transformed within their high-dimensional space to yield new decoders with similar 'information'. This is an empirical hypothesis that can be tested [35,36]. Even if seemingly distinct decoders were redundant, this would exacerbate the decoder's dictum: many patterns can decode neural activity. What, therefore, is the meaning of any individual pattern? Such an interpretation would necessitate a principled rationale; for instance, that the cost function used to derive a decoder is theoretically interesting.

The lack of theoretically informed decoders has led to nonsensical findings. For example, the winners of the 2006 Pittsburgh Brain Activity Interpretation Competition, a team of data scientists, fit a classifier that heavily relied on voxels in the ventricles and other regions highly affected by motion artifacts and physiological noise [37]. It has also been noted that fMRI-based decoding of motion in the visual cortex produces findings discordant with what would be expected; that is, that motion can be better decoded using voxels in V1 than in V5 [19]. Thus, decoding can produce surprising results, which are better explained by uninteresting side effects than by neurophysiological signals of interest.

Ubiquitous information undermines the meaningfulness of decoding

Human fMRI and animal studies suggest that task-correlated neural signals are widespread across the brain [36,38–49]. Clearly, the presence of widespread signals renders decoding an easier task. By aggregating enough information (e.g., thousands or tens of thousands of voxels),

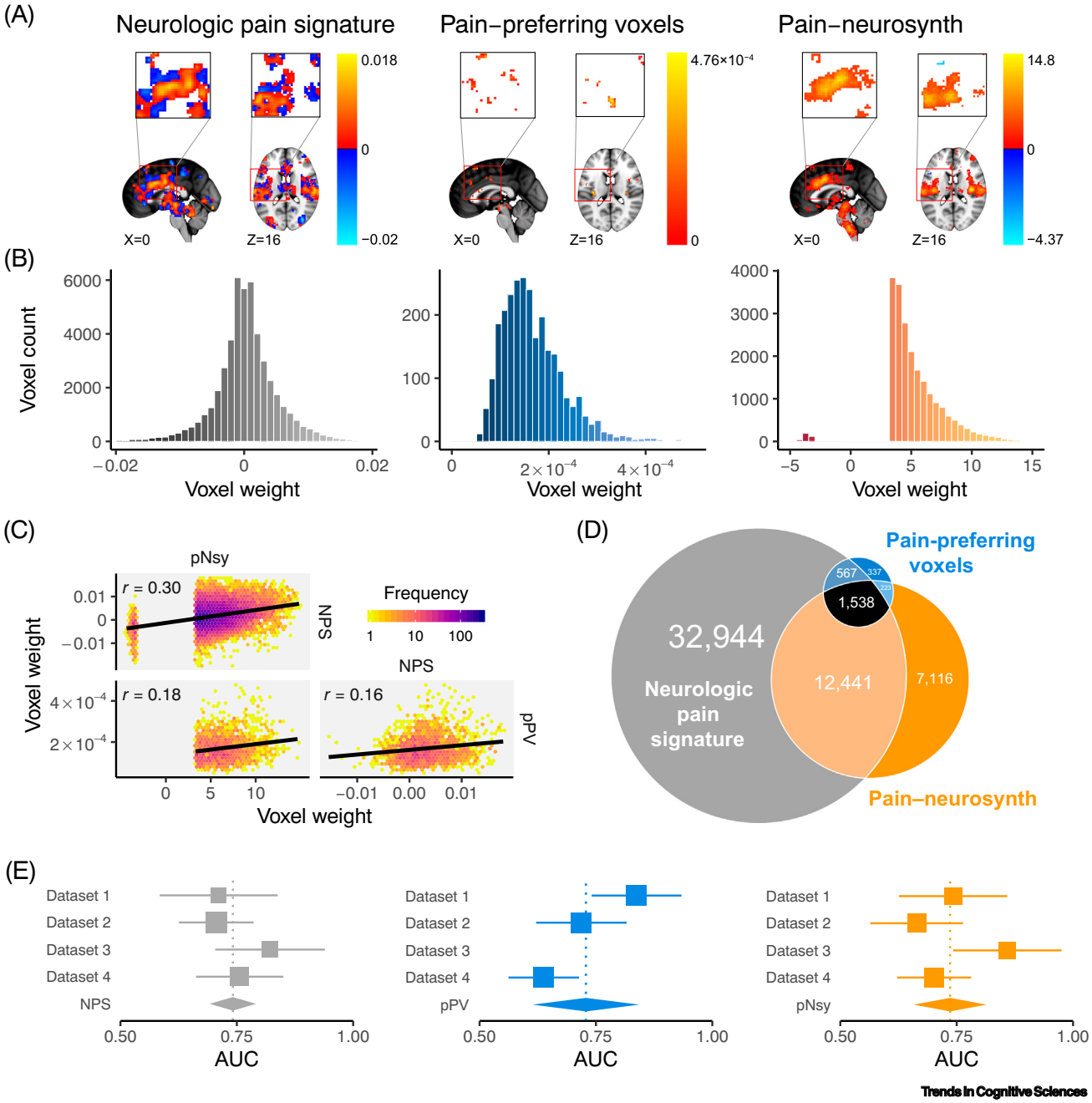

Figure 2. Three spatially discordant 'pain decoders' perform similarly. (A) Each decoder (abbreviated NPS, pPV, and pNsy) has a distinct pattern of voxel weights. (B) The weight distributions of the decoders are also distinct. NPS weights are distributed around zero; pPV weights are strictly positive; pNsy has only a few negative weights. (C) Pairwise relationships between the weights of common voxels within each of the three decoders. Lines depict total least squares regression fits. All three correlations are weak. (D) Euler diagram depicting the relative size of each decoder and the spatial overlap between them. (E) Meta-analysis of the discrimination performance of each decoder [area under the receiver operating characteristic curve (AUC) chance = 0.5] for decoding noxious from innocuous stimuli. Meta-analyses only included datasets that were independent of decoder derivation: since pPV was trained on Dataset 3, Dataset 3 was excluded from the meta-analysis of pPV. On average, all decoders perform similarly (AUC ≈ 0.73), but each estimate has appreciable variance. Square sizes indicate the meta-analytic weight, and lines indicate their 95% confidence intervals (CIs). Diamonds are the meta-analytic estimates, the width of which is the 95% CI of the meta-analytic estimate. Vertical broken lines pass through each meta-analytic point estimate. Adapted from [8].

one can discriminate between tasks, even in voxels with univariate *t*-statistics close to zero [36]. It follows that the ability to decode from arbitrary groups of voxels across the entire brain makes decoders even more challenging to interpret physiologically. That is, not only are the weights of decoders arbitrary due to the issues discussed in the previous section, but the voxels included in the decoders may also be nonspecific.

Have decoders yielded meaningful progress?

After 20 years of mind-reading [34,50,51], the question of whether fMRI-based decoders have yielded any novel, meaningful neurophysiological knowledge is inescapable. We address this question by critically assessing some thought-provoking decoding findings. For instance, using a probabilistic generative model, investigators uncovered semantic information from different listening tasks across the entirety of the human cortex, resulting in a semantic-specific parcellation of the neocortex [52]. Do such findings, albeit intriguing, represent true progress in understanding the brain mechanisms behind language processing? After all, if this widespread semantic 'representation' is causal, we should expect distinct semantic deficits with different cortical lesions. Still, clear semantic deficits do not arise from lesions reflecting the identified decoding patterns. Similarly, one may contend that seminal fMRI decoding findings, that different visual stimuli can be inferred using signals from the fusiform face area and parahippocampal gyrus [50], have not progressed our understanding of visual processing, for two reasons: (i) these areas were already known to respond to higher-order properties of visual stimuli [53,54]; and (ii) in contrast to the decoder's dictum, the specific information contained within the patterns used to decode was not elucidated. Finally, the ever-growing literature on decoding different pain states from fMRI signals (e.g., [8,23,34,36,55–57]) has not brought us any closer to understanding how pain experience is generated.

Oddly, despite all the issues described above, the field of fMRI decoding is thriving and yields some of the highest-impact papers in neuroimaging. Yet, the bold claims that these results have resulted in substantial progress in understanding how the brain works are plain wrong. It is reassuring to see that some authors adopt a more prudent stance. For example, one review states, 'In general, reward studies that use [multivoxel pattern analysis] approaches have largely confirmed previous results from univariate fMRI studies' [58]. Indeed, many decoding studies rely on encoding studies to fallaciously attempt to interpret their decoders, a flawed logic process akin to reverse inference [13].

Is decoding superior to encoding?

Encoding models are the inferential bread-and-butter of task fMRI, and they draw inferences differently compared with decoding models. Indeed, encoding and decoding models have been previously compared and contrasted [59]. Here, we add a few points to this discussion, especially concerning previous arguments that decoding analyses are more sensitive than encoding analyses.

In encoding, the time series of each voxel is typically regressed onto that of the task vector [8]. This can be done univariately (i.e., one voxel at a time; the DV is a random variable) or multivariately (i.e., many voxels at a time; the DV is a random vector) [60]. Although univariate and multivariate approaches both produce stimulus-related brain activation (or 'beta') maps, multivariate analyses facilitate inferences concerning how different voxels or regions activate together, whereas univariate analyses are blind to the correlation structure between different voxels or ROIs [60]. By contrast, many, although not all [61,62], fMRI decoding studies rely on spatial patterns derived from an encoding model.

It is often claimed that decoding allows one to study these spatial structures in a more sensitive way compared with univariate encoding models. However, encoding can also facilitate the study of 'patterns', arguably in a more principled way than decoding (i.e., by examining encoding

vectors), and can naturally do so using multivariate methods [60]. In contrast to mainstream positions in the decoding literature [12,63], there is evidence that these multivariate encoding models are more powerful compared with classification-based models such as decoders [64]. In addition, there are also conceptual benefits to multivariate encoding analyses. For instance, the null hypothesis tested by multivariate encoders differs from that tested by decoders. Specifically, the null hypothesis of decoders is that there is no effect in any subject, while encoders test the null hypothesis that the average encoding vector is the same between different conditions [65]. It follows that rejecting the encoding null hypothesis is of greater interest than rejecting the decoding null hypothesis [65]; we are typically more interested in understanding brain activity than in classifying individuals. Due to these benefits, we and others contend that researchers should strongly consider replacing decoding with multivariate encoding methods, such as multivariate analysis of variance [65,66]. Multivariate encoders are more aligned with the question of how conditions differ, are more interpretable, and are straightforward to implement.

Importantly, linear multivariate encoding models are not totally detached from linear decoding models. In fact, one can calculate encoding weights from decoding weights [28]. Since linear multivariate encoding models can capture the same information as linear decoders [28,59] while being more sensitive and interpretable, encoding models should remain the inferential bread-and-butter of task fMRI.

Mental state decoders cannot be used for real-world decision-making

Many studies in the decoding literature emphasize their potential for real-world application in many disciplines. For example, 11 years ago, a group of researchers claimed to have described an approach to identify pain, noting 'If our findings are extended to clinical populations, brain-based signatures could be useful in confirming pain in situations in which patients are unable to communicate pain effectively or when self-reports are otherwise suspect' [34]. Such unwarranted overoptimistic views have been echoed by others [67,68], neglecting cautionary notes [13,69]. Proposed real-world applications of decoders are manifold. Does the patient truly feel pain? Did the defendant commit the crime and, if so, did they commit it intentionally? Peering into someone's mind would absolve much uncertainty surrounding these issues. This is one implication of mental state decoders: By declaring whether a patient is in pain, a doctor can decide whether to prescribe them opioids and the insurance company can decide whether to pay the patient's bills. A jury can find a defendant guilty by declaring that a defendant is lying based on their brain scan. Although the implications of such applications are massive, the real-world utility of mental state decoders is inherently limited.

Decoder derivations are often detached real-world scenarios

Mental state decoders have been claimed to complement self-reports or, when self-reports are not available, such as in noncommunicative populations, replace self-reports [34,70,71]. In the context of these goals, it is important to consider that prediction models perform best in settings closest to those in which they were trained. Therefore, decoding works better when the condition to be decoded is similar to the condition used to generate the brain activity pattern used for decoding. However, many decoders are derived from the brain responses elicited by simple stimuli, such as noxious laser or heat stimuli to elicit pain [34,55], with the goal of generalizing these decoders to other 'mental states', conditions, or populations.

There is a clear issue when such decoders attempt to identify complex aspects of clinical conditions, such as spontaneous pain fluctuations in patients with chronic pain. These issues are more clearly elucidated by several questions concerning the clinical importance of decoders derived from contrived stimuli. Why should a decoder derived using evoked pain provide insight into ongoing clinical pain? We should not necessarily expect decoders trained on evoked pain to

generalize to spontaneous clinical pain, since evoked and spontaneous pains are associated with distinct brain activation patterns [72,73]. Why should a decoder built using a healthy population work in a clinical population? We should not expect it to, especially given that pathoanatomy and/or pathophysiology are presumed to drive the clinical condition. Why should a decoder trained and tested in a communicative sample be valid in a noncommunicative sample? Generalizability across populations is only assumed and neither justifiable nor falsifiable. If noncommunicative individuals differ neurophysiologically from those on whom a decoder was trained, such generalizability assumptions are especially dubious.

Arguments that these decoding tasks are justifiable must explicitly address these questions and justify the assumptions being made. Without justifications, these are simply assertions that should not be relied upon for real-world decision making. A corollary of this is that efforts to develop decoders intended to translate to the real world should be done in good faith and developed and tested in accordance with that goal, rather than relying on questionable assumptions.

Decoders cannot replace or supplement self-reports

Decoders intended to unveil subjective states (*cf.* stimulus or task properties) introduce critical questions concerning mental privacy and measurement. Decoders intending to decode subjective states are typically trained using self-reports describing the quality or intensity of a percept. In other words, the decoders try to predict self-reports. For example, in a recent study, investigators constructed a decoder intended to capture 'craving' ratings in response to visual cues in drug users and nonusers [74]. The authors stated '… given the role of self-reported craving in predicting outcomes, this brain-based pattern may function as both a diagnostic and predictive biomarker with potential utility in predicting clinically relevant individual differences and future outcomes' [74]. In other words, the authors suggest that their decoder responses provide information beyond craving self-reports alone, but can this be true? Of course not, since the decoder is just a noisy proxy of the self-report.

When building a decoder, self-reports are considered the gold standard. A crucial issue arises when applying the decoder to a communicative individual, as in a 'lie detector' situation. What if the individual's self-report conflicts with the prediction of the decoder? Why should one believe a measure trained using a self-report and intended to predict a self-report (proxy), over a self-report itself (gold standard)? Clearly, as we already described elsewhere [69], decoding has no additional value when self-reports are available.

Future directions

> We need less research, better research, and research done for the right reasons.
>
> [Douglas Altman, British Medical Journal [75]]

As papers on decoders continue to be published, it is imperative to ask whether the work behind them represents the best use of precious time and resources. Decoders neither advance our knowledge of brain physiology nor pave a path to sound real-world implementations, providing another burning example of 'research waste' described in biomedicine [75,76]. Studies in which decoders are the primary deliverable are arguably exercises that prioritize visibility over genuine scientific progress. We contend that, unless decoders offer neurophysiological insights or tangible real-world benefits, time and funding are better spent on more fruitful scientific efforts.

Still, mental state decoders can be part of a broader and more solid question-centered research paradigm. In this paradigm, decoders are a means to obtain a different objective rather than an

objective in themselves. Instead of identifying the presence of information, we should seek to understand the nature of the identified information. Experimenters can attempt to achieve this by exploiting decoders to study how information content shifts as a function of the task or participants' responses. This view yields an important distinction between two classes of decoding studies: (i) studies in which the authors are satisfied with showing that the decoder works, and (ii) studies in which the primary deliverable is knowledge derived from applying the decoder. Studies within the former group are still the majority; they assume the decoder represents a fundamental scientific contribution, tangibly advancing a field [12,34,67,71]. However, as we argued, this assumption is untenable. By contrast, studies using decoders as tools to probe neural activity are less problematic: the inferences come from experimental manipulations rather than from the decoder itself (e.g., [58]).

Our arguments do not preclude encoding or decoding from being used as a 'first step' in a scientific effort to understand the nature of brain activity. For instance, the ability to decode brain-wide audition-related signals decreased with increasing levels of sedation while auditory cortex activity remained stable [36]. By assuming that brain activity is necessary for perception and because auditory cortex activity remained stable, it was concluded that at least some of the brain-wide information must be necessary for conscious perception. This conclusion relied on using a decoder to understand the nature of the information across different experimental conditions, which could not have been drawn by simply showing that the decoder worked. Although fMRI cannot directly manipulate brain activity to establish causality, it can provide hypotheses that future experiments can test by perturbing brain circuits.

Concluding remarks

We argued that decoders themselves are not mechanistically interesting and do not show promise for real-world decision-making. The mechanistic vapidity of decoders arises from their non-uniqueness, non-interpretability, and incompatibility with the data-generating process. The poor applicability of decoders stems from both being philosophically ungrounded and a model fitting incompatible with the intended application. Some of these issues can be remedied by shifting the research focus from decoding itself to the experimental context in which the decoder predictions are made. There are several examples of research questions that use decoding as a means to gain novel physiological knowledge (see Outstanding questions).

Outstanding questions

How can the interpretability of mental state decoders be improved? Investigators often use plug-and-play models to develop decoders that are not based on physiological knowledge. Models constrained by *a priori* physiological hypotheses, or decoding paired with experiments to investigate the nature of the decoded information, will facilitate decoder interpretability.

Can decoding studies be improved to yield novel mechanistic insights above and beyond encoding? Many of the insights obtained by decoders are also possible with encoding. If decoding is to be used fruitfully in fMRI, then decoders must demonstrate unique value over more interpretable and parsimonious encoders.

What test could be used to demonstrate the specificity of a decoder? A decoder may successfully predict a mental state, but the same decoder can also reliably discriminate between other mental states. For instance, if a decoder built to discriminate painful from warm sensations can also discriminate between auditory and visual sensations, then that decoder cannot be called specific.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1324585, National Institute of Drug Abuse at National Institutes of Health (1P50DA044121), and the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number F31NS126012. The content is solely the authors' responsibility and does not represent the official views of the National Institutes of Health.

Declaration of interests

None declared by authors.

References

1. Poldrack, R.A. and Farah, M.J. (2015) Progress and challenges in probing the human brain. *Nature* 526, 371–379
2. Yarkoni, T. *et al.* (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670
3. Jezzard, P. *et al.* (2001) *Functional MRI : An Introduction to Methods*, Oxford University Press
4. Pereira, F. *et al.* (2009) Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209
5. Poldrack, R.A. (2011) Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* 72, 692–697
6. Poldrack, R.A. (2006) Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63
7. Varoquaux, G. and Thirion, B. (2014) How machine learning is shaping cognitive neuroimaging. *Gigascience* 3, 28
8. Jabakhanji, R. *et al.* (2022) Limits of decoding mental states with fMRI. *Cortex* 149, 101–122
9. Norman, K.A. *et al.* (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430
10. Haxby, J.V. *et al.* (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37, 435–456

11. Hanke, M. *et al.* (2009) PyMVPA: a unifying approach to the analysis of neuroscientific data. *Front. Neuroinform.* 3, 3
12. Kragel, P.A. *et al.* (2018) Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron* 99, 257–273
13. Hu, L. and Iannetti, G.D. (2016) Painful issues in pain prediction. *Trends Neurosci.* 39, 212–220
14. Henson, R. (2006) Forward inference using functional neuroimaging: dissociations versus associations. *Trends Cogn. Sci.* 10, 64–69
15. Haynes, J.D. and Rees, G. (2006) Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534
16. Kohoutova, L. *et al.* (2020) Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* 15, 1399–1435
17. Baker, B. *et al.* (2022) Three aspects of representation in neuroscience. *Trends Cogn. Sci.* 26, 942–958
18. Pessoa, L. (2022) *The Entangled Brain : How Perception, Cognition, and Emotion Are Woven Together*, MIT Press
19. Ritchie, J.B. *et al.* (2019) Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *Br. J. Philos. Sci.* 70, 581–607
20. Shmueli, G. (2010) To explain or to predict? *Stat. Sci.* 25, 289–310
21. Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701
22. Imbens, G. and Rubin, D.B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press
23. Lee, J.J. *et al.* (2021) A neuroimaging biomarker for sustained experimental and clinical pain. *Nat. Med.* 27, 174–182
24. Hagerty, M.R. and Srinivasan, V. (1991) Comparing the predictive powers of alternative multiple regression models. *Psychometrika* 56, 77–85
25. Westreich, D. and Greenland, S. (2013) The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* 177, 292–298
26. Chen, G. *et al.* (2024) Through the lens of causal inference: decisions and pitfalls of covariate selection. *bioRxiv*, Published online January 12, 2024. https://doi.org/10.1101/2024.01.11.575211
27. Mesulam, M.M. (1998) From sensation to cognition. *Brain* 121, 1013–1052
28. Haufe, S. *et al.* (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110
29. Haynes, J.D. (2015) A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* 87, 257–270
30. Wang, Y. and Wang, L. (2020) Causal inference in degenerate systems: an impossibility result. *Proc. Mach. Learn. Res.* 108, 3383–3392
31. Nizami, L. (2019) Information theory is abused in neuroscience. *Cybern. Hum. Knowing* 26, 47–97
32. Pessoa, L. *et al.* (2022) Refocusing neuroscience: moving away from mental categories and towards complex behaviours. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 377, 20200534
33. Mensch, A. *et al.* (2021) Extracting representations of cognition across neuroimaging studies improves brain decoding. *PLoS Comput. Biol.* 17, e1008795
34. Wager, T.D. *et al.* (2013) An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* 368, 1388–1397
35. Carter, R.M. *et al.* (2012) A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science* 337, 109–111
36. Vigotsky, A.D. *et al.* (2023) Widespread, perception-related information in the human brain scales with levels of consciousness. *bioRxiv*, Published online January 23, 2023. https://doi.org/10.1101/2022.09.19.508437
37. Hebart, M.N. and Baker, C.I. (2018) Deconstructing multivariate decoding for the study of brain function. *Neuroimage* 180, 4–18
38. Gonzalez-Castillo, J. *et al.* (2012) Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci. U. S. A.* 109, 5487–5492
39. Kumar, S. *et al.* (2020) Searching through functional space reveals distributed visual, auditory, and semantic coding in the human brain. *PLoS Comput. Biol.* 16, e1008457
40. Cox, C.R. and Rogers, T.T. (2021) Finding distributed needles in neural haystacks. *J. Neurosci.* 41, 1019–1032
41. Rish, I. and Cecchi, G.A. (2017) Holographic brain: distributed versus local activation patterns in fMRI. *IBM J. Res. Dev.* 61, 3: 1–3:9
42. Rish, I. *et al.* (2012) Sparse regression analysis of task-relevant information distribution in the brain. In *Medical Imaging 2012: Image Processing* (Haynor, D.R. and Ourselin, S., eds), p. 831412, SPIE
43. Mohr, H. *et al.* (2015) Sparse regularization techniques provide novel insights into outcome integration processes. *Neuroimage* 104, 163–176
44. Vickery, T.J. *et al.* (2011) Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron* 72, 166–177
45. Ren, C. and Komiyama, T. (2021) Characterizing cortex-wide dynamics with wide-field calcium Imaging. *J. Neurosci.* 41, 4160–4168
46. Pinto, L. *et al.* (2019) Task-dependent changes in the large-scale dynamics and necessity of cortical regions. *Neuron* 104, 810–824
47. Stringer, C. *et al.* (2019) Spontaneous behaviors drive multidimensional, brainwide activity. *Science* 364, 255
48. Gilad, A. and Helmchen, F. (2020) Spatiotemporal refinement of signal flow through association cortex during learning. *Nat. Commun.* 11, 1744
49. Lab, I.B. *et al.* (2023) A brain-wide map of neural activity during complex behaviour. *bioRxiv*, Published online July 4, 2023. https://doi.org/10.1101/2023.07.04.547681
50. Haxby, J.V. *et al.* (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430
51. Haxby, J.V. (2012) Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* 62, 852–855
52. Huth, A.G. *et al.* (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458
53. McCarthy, G. *et al.* (1997) Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* 9, 605–610
54. Epstein, R. and Kanwisher, N. (1998) A cortical representation of the local visual environment. *Nature* 392, 598–601
55. Liang, M. *et al.* (2019) Spatial patterns of brain activity preferentially reflecting transient pain and stimulus intensity. *Cereb. Cortex* 29, 2211–2227
56. Kohoutova, L. *et al.* (2022) Individual variability in brain representations of pain. *Nat. Neurosci.* 25, 749–759
57. Losin, E.A.R. *et al.* (2020) Neural and sociocultural mediators of ethnic differences in pain. *Nat. Hum. Behav.* 4, 517–530
58. Kahnt, T. (2018) A decade of decoding reward-related fMRI signals and where we go from here. *Neuroimage* 180, 324–333
59. Naselaris, T. *et al.* (2011) Encoding and decoding in fMRI. *Neuroimage* 56, 400–410
60. Friston, K.J. *et al.* (1995) Characterizing dynamic brain responses with fMRI: a multivariate approach. *Neuroimage* 2, 166–172
61. Nakai, T. and Nishimoto, S. (2022) Representations and decodability of diverse cognitive functions are preserved across the human cortex, cerebellum, and subcortex. *Commun. Biol.* 5, 1245
62. Misra, J. *et al.* (2021) Learning brain dynamics for decoding and predicting individual differences. *PLoS Comput. Biol.* 17, e1008943
63. Davis, T. *et al.* (2014) What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage* 97, 271–283
64. Rosenblatt, J.D. *et al.* (2021) Better-than-chance classification for signal detection. *Biostatistics* 22, 365–380
65. Allefeld, C. *et al.* (2016) Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *Neuroimage* 141, 378–392
66. Allefeld, C. and Haynes, J.D. (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *Neuroimage* 89, 345–357
67. Woo, C.W. *et al.* (2017) Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20, 365–377
68. Zunhammer, M. *et al.* (2016) Issues in pain prediction - more gain than pain. *Trends Neurosci.* 39, 639–640
69. Mouraux, A. and Iannetti, G.D. (2018) The search for pain biomarkers in the human brain. *Brain* 141, 3290–3307
70. van der Miesen, M.M. *et al.* (2019) Neuroimaging-based biomarkers for pain: state of the field and current directions. *Pain Rep.* 4, e751
71. Davis, K.D. *et al.* (2020) Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nat. Rev. Neurol.* 16, 381–400

72. Hashmi, J.A. *et al.* (2013) Shape shifting pain: chronification of back pain shifts brain representation from nociceptive to emotional circuits. *Brain* 136, 2751–2768
73. Shirvalkar, P. *et al.* (2023) First-in-human prediction of chronic pain state using intracranial neural biomarkers. *Nat. Neurosci.* 26, 1090–1099
74. Koban, L. *et al.* (2023) A neuromarker for drug and food craving distinguishes drug users from non-users. *Nat. Neurosci.* 26, 316–325
75. Altman, D.G. (1994) The scandal of poor medical research. *BMJ* 308, 283–284
76. Chalmers, I. and Glasziou, P. (2009) Avoidable waste in the production and reporting of research evidence. *Lancet* 374, 86–89