Multivariate pattern analysis and the search for neural representations

Bryce Gessell [1], Benjamin Geib [2,3], and Felipe De Brigard [2,3,4]

1. Department of Philosophy, Southern Virginia University

2. Department of Psychology and Neuroscience, Duke University

3. Center for Cognitive Neuroscience, Duke University

4. Department of Philosophy, Duke University

Acknowledgments:

Previous versions of this paper were presented at the 43rd Annual Meeting of the Society for Philosophy and Psychology at Johns Hopkins University, the workshop on *Locating Representations in the Brain* at Stanford University, and the meeting of the Deep South Philosophy and Neuroscience Workgroup at the Central APA. We thank the participants in these events for their feedback. We also thank the members of the Imagination and Modal Cognition lab at Duke University for many helpful discussions.

Corresponding author:

Felipe De Brigard

203A West Duke Building

Duke University

Durham, NC 27708

felipe.debrigard@duke.edu

Multivariate pattern analysis and the search for neural representations

Abstract

Multivariate pattern analysis, or MVPA, has become one of the most popular analytic methods in cognitive neuroscience. Since its inception, MVPA has been heralded as offering much more than regular univariate analyses, for—we are told—it not only can tell us which brain regions are engaged while processing particular stimuli, but also which patterns of neural activity *represent* the categories the stimuli are selected from. We disagree, and in the current paper we offer four conceptual challenges to the use of MVPA to make claims about neural representation. Our view is that the use of MVPA to make claims about neural representation is problematic.

Key words

MVPA; mental representation; neural representation; misrepresentation; neuroscience.

1. Introduction

Multivariate pattern analysis, or MVPA, has become one of the most popular analytic methods in cognitive neuroscience (Haxby et al., 2014; Weaverdyck et al., 2020). Since its inception, MVPA has been heralded as offering much more than regular univariate analyses, such as general linear modeling (GLM) approaches, which can merely associate a certain neuroimaging measure, say, increased BOLD signal in a particular brain region, with one experimental condition relative to another one. By contrast—the story goes—MVPA could tell us whether a particular pattern of brain activity carries information about a specific stimulus. Indeed, many researchers take MVPA to be able to provide evidence that a particular pattern of brain activity *represents* the categories the stimuli are selected from. This was the moral we were invited to draw from the first use of MVPA with fMRI signal, where it was employed to discriminate—according to the paper's title (Haxby et al, 2001)—"Distributed and overlapping *representations* of faces and objects in ventral temporal cortex" (emphasis added). Year after year we see papers with similar titles, all of which support their claims by employing MVPA approaches on neuroimaging data. For example, we have learned of patterns of brain activity that correspond to "categorical *representations* of objects" (Carlson et al., 2003; emphasis added), or "*representation* of behavioral choice from motion" (Serences & Boynton, 2007; emphasis added), or even "abstract and concrete concept *representations*" (Wang et al., 2013; emphasis added). MVPA has even been employed for "Decoding individual natural scene representations during perception and imagery" (Johnson & Johnson, 2014) because, we are told, MVPA can be used to "directly probe how information is *represented* in visually responsive brain areas" (Johnson & Johnson, 2014; emphasis added). In sum, MVPA does not seem to be simply a gradual increment over typical GLM approaches, for it has "revolutionized fMRI research" by allowing researchers to answer new questions that were hitherto unanswerable: “Instead of asking what a region’s function is, in terms of a single brain state associated with global activity, fMRI investigators can now ask what information is *represented* in a region, in terms of brain states associated with distinct patterns of brain activity, and how that information is encoded and organized” (Haxby et al., 2014: 436; emphasis added).

Roughly speaking, MVPA can be seen as a four-stage process (Norman et al., 2006). The first stage, *feature selection*, starts when the experimenters choose stimuli from two or more categories of interest, and proceed to present them to participants while collecting neuroimaging data—typically functional magnetic resonance imaging (fMRI), but also electroencephalography (EEG), magnetoencephalography (MEG), single unit recordings, and so on. To illustrate, suppose researchers are interested in brain activity, as measured by blood oxygen level dependent (BOLD) signal with fMRI, while participants are seeing stimuli that belong to one of two categories: LIVING or NON-LIVING. Feature selection occurs when researchers choose a set of stimulus-dependent data, from a set of voxels of interest—which could be as large as all voxels or some subset of them—in order to include them in a subsequent classification analysis. While there are a number of different feature selection methods, they typically involve an a priori selection of a region of interest (ROI), and, often, the use of some multivariate statistic in order to allow the inclusion of voxel-wise activity related to the stimuli of interest, that would not normally meet the stringent univariate threshold of statistical significance—although, to be fair, sometimes univariate approaches are used as well (Fan & Chou, 2016). In our example, the ROI could be, say, lateral temporal cortex, and feature selection occurs when the researcher chooses a set of voxel-wise activity from this ROI to include in the next stage of analysis.

This second stage, *pattern assembly*, consists in sorting the selected data into discrete vectors or “patterns” that correspond to individual stimuli from each category, based upon the voxels chosen in the previous *feature selection* step. Thus, in our example, fMRI data from the selected ROI will be sorted into patterns of brain activity associated with each individual stimulus from the LIVING category and patterns of brain activity associated with each individual stimulus of the NON-LIVING category. Each one of these distinct patterns is thus labeled as either "living" or "non-living," and then they are randomly divided into two sets: a training set and a testing set.

Now comes the third stage, *classifier training*, whereby the data in the training set is fed to a machine learning algorithm that learns a function to map the voxel-based vectors or patterns onto the experimental conditions. There are a number of machine learning algorithms in the offing (Pereira et al., 2009), but most typically they involve approaches similar to logistic regression analysis in which the predicted variable is not continuous but discrete. A classifier algorithm may learn a function to map patterns of brain activity (i.e., vectors of voxel-wise measures of BOLD signal from the selected ROI) from the training set—the predictor variables—onto a stimulus discriminant category—the predicted variable—which, in our example, is either LIVING or NON-LIVING.

The fourth and final stage, *classifier testing*, consists in feeding the untrained data—i.e., the dataset that was previously set apart as the testing set—to the classifier in order to determine whether or not the function correctly classifies these new patterns of brain activity as corresponding to relevant categories. Thus, in our example, the testing data will be the set of brain activity from the ROI that was not used to train the classifier, and the test is successful if, when fed to the algorithm, the classifier predicts whether a brain pattern corresponds to a stimulus with either a "living" or a "non-living" label beyond some determined decision boundary—which could be chance or certain percent accuracy (Pereira et al., 2009).

Although, as mentioned, MVPA has grown in popularity for the last two decades, many have pointed out a number of limitations. Some of these limitations fall on the practical or methodological side of things. For instance, temporal autocorrelation often yields collinearities in the preprocessing of fMRI data that could potentially render two brain patterns as being more similar to each other than they should be, particularly if they belong to adjacent trials or are part of the same run (see Davis & Poldrack, 2013, for discussion). Others have pointed at discrepancies that should not exist between category structures recovered from MVPA analysis conducted with fMRI versus single unit recordings of the same brain regions (Dubois et al., 2015). And others have indicated various difficulties in disambiguating geometrical and spatial ambiguities inherit in MVPA, with the further difficulty that many of the proposed solutions come with additional costs (Naselaris & Kay, 2015).

Yet, other limitations are more theoretical or conceptual in nature, and a number of them have been voiced by philosophers of neuroscience. For instance, Anderson & Oates (2010) argued that inferring represented information on the basis of information available to the classifier is, at best, problematic. In support, they ran a series of simulations and showed that nearly any voxel, regardless of activation level, could end up in the "most informative" set of voxels and, thus, carry a heavy weight in the generation of the pattern classifier. As such, they warn us that the inference from being the set of voxels most relevant for the classifying prediction, to being the set of brain regions carrying information about the predicted categories, is unwarranted. Similarly, Wright (2018) argues that results from MVPA can only be taken as giving us direct evidence about patterns in the data, not about patterns in the phenomena from which such data is generated—which, in this case, would be the alleged represented information. Inferring the latter from the former is, according to Wright, an unwarranted leap. Along similar lines, Ritchie, Kaplan and Klein (2019) have recently argued against what they call the “Decoder’s Dictum,” which says that decoding success provides strong evidence about the information that patterns represent. Their argument is that, even in successful decoding, a gap remains between the information used by the decoder and that represented in the brain. The existence of this gap threatens the reverse inference from decoder information to neural information.

The current paper seeks to add four additional philosophical challenges to the use of MVPA to make claims about neural representation. The nature of these challenges depends in part on what notion of representation MVPA is supposed to provide evidence for. Inspired by Sullivan (2010), we distinguish between a strong and a weak sense of representation in neuroscience, which closely follow her *substantive* versus *minimal* distinction. In the strong sense, a representation plays an explanatory role in neuroscience in virtue of its content, whereas in the weak sense it does so by merely signaling co-variation with its referent. Thus, the first two challenges, which we call “the problem of cross-cut categories” and “the problem of misrepresentation,” apply to inferences from MVPA to neural representations in a strong sense, and are discussed in sections 2 and 3, respectively. The other two challenges, which we call “the problem of orthogonalization” and “the problem of null results,” pertain to inferences from MVPA to neural representations in a weak sense, and are discussed in sections 4 and 5, respectively. Finally, in section 6, we conclude with a critical assessment of the promises of MVPA to deliver evidence about representations in the brain.

2. The problem of cross-cut categories

To introduce this first challenge, consider a thought experiment. Imagine there is a neuroscientist who is interested in understanding how the brain represents things that are living and things that are non-living. As such, she runs an fMRI study in which images of items that are living and items that are non-living are employed as stimuli, and then she asks participants to simply make a categorical discrimination judgment while lying in the scanner. A selected sub-set of the resultant data is chosen on the basis of some a priori ROI analysis, and trial-based vectors are then labeled as “living” or “non-living”. In turn, these data are split into training and testing sets, and the former are used to train a classifier to tell apart trials labeled as “living” versus “non-living”. Happily, the classifier succeeds, and the neuroscientist concludes that she has identified patterns of brain activity that represent the categories of LIVING and NON-LIVING.

But now suppose that this neuroscientist dies, unexpectedly, before she has the chance to report the results. And suppose further that the experimental protocol gets lost as well as the Python script she used to analyze the data. Nothing is left but the uncategorized images employed as stimuli and the collected fMRI data. Imagine, though, that another neuroscientist discovers these data and the stimulus set, and tries to make sense of the experiment. She looks at the uncategorized images and finds that they all could be neatly divided into SMALL and LARGE, so she takes it that the experiment must have employed a size discrimination task. Now, coincidentally, the set of small items in the stimuli includes only living things, while the set of large items includes only non-living things; that is, the SMALL and LARGE categories cross-cut the old LIVING and NON-LIVING categories. Unaware of this coincidence, the experimenter carries on by sorting trials into labels “small” and “large”, and proceeds to divide these data into a training set and a testing set. Once again, she feeds a classifier with the training set and then tests it on the testing set. Lo and behold, the new classifier performs just as well, and it is able to identify voxel patterns as belonging to the categories SMALL or LARGE at an above-chance rate. Critically, the same brain patterns of activity the new classifier identifies as either “small” or “large” correspond to those the previous classifier identified as either “living” or “non-living”.

Let us ask now: what does the pattern of brain activity initially assigned to “living”, and now assigned to “small”, represent? Does the pattern carry information about items that fall under the category LIVING and, thus, represent living things, or does it carry information about items that fall under the category SMALL and, thus, represent small things? Does it represent both? Or does it represent neither? Philosophers—we hope—may find a family resemblance between this thought experiment and Quine’s famous doctrine of *indeterminacy of reference* (Quine, 1960)[1], which he introduces with the famous example of ‘Gavagai’. Quine invites us to imagine a linguist in a foreign land meeting a native who only speaks a completely unknown language. All of a sudden, “a rabbit scurries by, the native says ‘Gavagai’, and the linguists notes down the sentence ‘Rabbit’ (or ‘Lo, a rabbit’) as tentative translation” (Quine, 1960: 25). The linguist, however, is aware that the utterance is consistent with a number of other possible meanings; it may mean, for instance, ‘jumping’, ‘fast’, ‘food’ or ‘white’, let’s say. So she tests her initial hypothesis—the tentative translation—by uttering ‘Gavagai’ in the presence of jumps, fast things, foods and white objects, and records the native’s assent or dissent accordingly. This way, she manages to narrow down the possible meaning of ‘Gavagai’. The problem, however, is that there are a number of compatible translations that no amount of testing could falsify, such as ‘there is a rabbit’, or ‘Lo, undetached rabbit parts’, or even ‘rabbithood is manifested here now’. While all these co-instantiated entities constitute possible different referents of the term ‘Gavagai’, it is impossible, *from the behavioral/observational data alone*, to know for sure whether ‘Gavagai’ means ‘rabbit’

[1] Quine sometimes uses the expressions “ontological relativity”, “indeterminacy of translation” and “inscrutability of reference” to refer to the same underlying notion. Although some philosophers tried to pull apart these ideas as different doctrines, later in life Quine made clear that these terms expressed the same idea (Quine, 1986).

and not ‘undetached rabbit parts’. In other words, a precise translation of ‘gavagai’ is underdetermined by the observational data and, thus, its referent remains inscrutable.

We believe that a similar lesson applies to the problem of cross-cut categories in MVPA. Mutatis mutandis, a machine learning classifier is nothing but an automated linguist that is trying to assign proxy-functions from expressions in brain data to objects in the world. But as with the case of the native’s behavior, the observational data—i.e., the patterns of brain activity—can be mapped to at least two co-extensive sets of stimuli and, thus, can mean either ‘small’ or ‘living’, or ‘large’ or ‘non-living’. It would be premature for the second neuroscientist to claim that she has found a pattern of brain activity that means SMALL, just as it would have been premature for the first neuroscientist to claim that she has found a pattern of brain activity that means LIVING. These are at best—to use Quine’s trope—tentative translations.

‘What’s the big deal?’, one may rebuff at this point, ‘Couldn’t a third, clever neuroscientist notice the co-extensionality of the stimulus set and run another experiment to tell them apart—this time involving living things that aren’t small, and non-living things that aren’t large?’ Sure! In practice, this is likely the best methodological response. In fact, it is precisely what the linguist in Quine’s fable does: he collects further observations to confirm or disconfirm semantic hypotheses. But the point in principle persists, for there is always an alternative possible category the same brain data could predict such that it would make the alleged neural representation to be about something else altogether. Then, just as it would have been wrong for the first neuroscientist to say that a certain pattern of brain activity represents LIVING, since—as shown by the second neuroscientist—it could as well represent SMALL, this third, clever neuroscientist could also be unaware of another co-extensive category her data set could be equally predictive of.

Although our thought experiment appears far-fetched, there is some evidence that suggests that, actually, it may not be as unlikely as it seems. Consider the aforementioned study by Haxby and colleagues (2001), which showcased MVPA as a method to identify representations in the brain. Among other things, this paper claims to find evidence of distinct neural representations for faces, cats, and five categories of human-made objects: houses, chairs, scissors, shoes, and bottles. Their selected ROI is the ventral occipitotemporal cortex, which they had reason to believe was sensitive to these sorts of contents. Fast-forward a decade, when Konkle & Oliva (2012) scanned participants while displaying several pictures of the same retinal size depicting objects that, in reality, are either smaller than or larger than an average person. Their analysis showed, among other things, distinct patters of brain activity in the ventral surface of the occipitotemporal cortex as a function of whether the object depicted by the displayed picture was either smaller or larger than an average person. How are we to know that the occipitotemporal pattern of brain activity identified by Haxby and colleagues was representing, say, BOTTLE and not BOTTLE SMALLER THAN ME, as identified by Konkle and Oliva (2012)? Indeed, just a year later, Konkle and Caramazza (2013) identified yet another category—animacy—that could also cross-cut those of object type and relative size. Now the same batch of gray matter could represent BOTTLE, BOTTLE SMALLER THAN ME, or STATIC BOTTLE.

In fact, as shown recently by Goddard and colleagues (2018), similar concerns could even arise for studies that are "data-driven" or "hypothesis-free", as opposed to hypothesis-driven ones beginning from a set of pre-determined categories. Roughly, data-driven approaches typically employ dimensionality reduction strategies to generate a tractable feature-space—i.e., a multidimensional space in which coordinates correspond to features of the stimuli or experimental design—that's allegedly hypothesis-free or unbiased, and from which one can read-off its underlying representational-space, namely the structure of neural activity that carries the contents in the feature-space revealed by the dimensionality reduction analysis. Unfortunately, as Goddard and colleagues show, different dimensionality reduction strategies fail to converge on a single solution for a feature-space, even when the representational structure they are supposed to mirror is as simple as a two- or three-dimensional one. Part of the problem, as they elaborate, is that many dimensionality reduction strategies suffer from “rotation indeterminacy”: since one can arbitrarily rotate factor solutions, it is always possible to generate a different feature-space (see also Carlson et al, 2018). And some of those spaces, as they demonstrated, do not look at all like the alleged representational structure they were supposed to recover. Indeed, what typically occurs is that the researchers already have a representational structure in mind, and use their preferred feature-space ‘solution’ as evidence that it can be recovered in an unbiased manner, when in reality the read-off is hypothesis-driven too. This opens up the door for another gavagai-like situation, whereby two hypothetical researchers, employing two distinct dimensionality reduction analyses to the same data, extract different solutions that recover distinct representational structures. What the neural patterns represent is, once again, underdetermined by the data.

‘This is an old story!’, someone may gruffly interject at this point, ‘This is exactly what happened when we were told that the fusiform face area (FFA; Kanwisher et al., 1997) was not selective for faces, since it also responded to greebles (Gauthier et al., 1999) and to expertise in categorizing cars and birds (Gauthier et al., 2000). So what’s new?’. Excellent question! What is new is that neither Kanwisher and colleagues, nor Gauthier and collaborators, claimed that their findings gave them evidence to the effect that FFA *represented* faces, greebles or general expertise. They knew their lowly univariate methods could at best speak to stimulus-dependent processing differences, not that those brain regions represented faces, or greebles, or what-have-you. But this is not what MVPA is trying to sell us, for we have been told that this method can "revolutionize fMRI research" as it can move beyond stimulus-dependent processing differences to evidence about what the brain *represents*. Our point is simply that it can't—or at least not if we want to hold a view according to which neural representations are *about* something in particular, i.e., that they have determinate meanings. In sum, the problem of cross-cut categories simply seeks to suggest that MVPA falls short of delivering what it promises.

3. The problem of misrepresentation

Traditionally, philosophers who think of mental states as representational take such entities as exhibiting intentionality, i.e., mental representations are about or refer to things. That which a mental representation is about is often called its *intentional object*, the adjectival qualifier indicating that the object need not exist. One can think that Santa was stingy last Christmas even though (spoiler alert) Santa does not exist. Additionally, most representational views of mental states distinguish between the intentional object and the intentional *content* of the mental representation, on account that mental states are *opaque* or *intensional* (with an 's')[2]. Louis Lane likely can entertain lots of thoughts about Superman. She may think, for instance, that Superman is brave. She can also entertain thoughts about Clark Kent; she may think, for instance, that Clark

[2] The term intensional comes from mathematical logic, and refers to definitions that are not extensional. A set is extensionally defined when its elements are listed. Thus, the set of all even number is extensionally defined as the set containing 2, 4, 6, 8, 10 and so on. But one could also intensionally define the same set with the function $f(\text{n}) = \text{n} \,.\, 2$, where $\text{n} \in \mathbb{N}$. Chisholm (1957) famously argued that intentional statements are also intensional because (1) they don't admit existential generalization—viz., from "Ana believes Santa is stingy" it does not follow that there exists an x such that x is Santa and x is stingy—and (2) co-referential terms in intensional contexts cannot be substituted *salva veritate*—viz., I cannot substitute "Clark Kent" for "Superman" in "Louis Lane believes that Clark Kent is a coward" without changing the truth value of the whole statement. Critical for our discussion is the fact that two terms can be co-extensional and differ in their intensional definition. For instance, the set of all cordata (animals with hearts) is co-extensional with that of renata (animals with kidneys), but it is definitively different to entertain thoughts about cordata as opposed to renata (Quine, 1951). The elements in the set—the referent—do not determine how you think about them for, as Frege put it, "sense determines reference": two expressions with the same sense will have the same referent, but two expressions with the same referent need not have the same sense (Frege, 1892).

Kent is a coward, since he seems to run away every time there is trouble. But even though Superman and Clark Kent are the same individual—Kal El—Louis Lane is not contradicting herself when she entertains both thoughts. This is because when she thinks about Superman, she's thinking of him under a certain *mode of presentation* or in a particular *sense*, which is different from the sense or mode of presentation under which she thinks about Clark Kent. This is a difference in the *intentional content* of her mental states, as they both refer to the same intentional object.

The project of naturalizing intentionality, of understanding how the brain can instantiate the semantic properties we attribute to mental representations, requires not only that we can explain how brains like ours can have neural states that refer to things that may or may not exist—i.e., how they can relate to their intentional objects—but also how they can be the bearers of intentional contents capable of determining their referents under one or another mode of presentation. Indeed, one way of thinking about the problem of cross-cut categories mentioned above, is in terms of MVPA failing to tell us anything about the content of neural representations that happen to have co-extensional referents.

With this clarification in mind, let us look at the second challenge, which is very closely related to the first one.[3] According to the received view in the philosophy of mind, any attempt to naturalize intentionality needs not only to explain how neural states or processes can bear contents that represent their intentional objects, but also how sometimes they can *misrepresent* them

[3] So closely related they are, that a reviewer suggested to combine this challenge with the previous one. However, we decided to keep them separate. We reasoned that, for many philosophers of mind with a naturalistic bent, the issue of misrepresentation is critical (Neander, 1995), and it is likely that at least some of them may be unmoved by Quinean considerations about radical translation. In fact, a number of philosophers of mind and cognitive scientists have rejected Quine's concerns (e.g., Chomsky, 1968; Soames, 1999), and yet they remain open to the possibility of naturalizing intentionality and representation (e.g, Fodor, 1990). Thus, although the points that the problem of cross-cut categories and the problem of misrepresentation make are closely related, the motivation behind them differs. We thank a reviewer for inviting us to clarify this issue.

(Dretske, 1986; Neander, 1995; Neander & Schulte, 2021). The idea comes from Grice's (1957) influential distinction between natural and non-natural meanings. The former refers to instances in which we say, of a meaning-bearer particular, X, that it means that *p*, only if *p* is the case. If we say "The spots on Tommy's face means he has measles, but he does not have measles", we are contradicting ourselves. The spots on Tommy's face mean that he has measles only if he has measles (Dretske, 1986). By contrast, non-natural meanings are such that X can mean that *p* even when *p* is not the case. If we say "The needle in the gas gauge means the tank isn't empty, but the tank is empty" we are not contradicting ourselves. Sediment in the tank may make the needle indicate that there is gas, for instance, even when the tank is empty. In other words: the needle in the gas gauge, unlike Tommy's face spots, can misrepresent. Given that many of the expressions of our thoughts convey non-natural meanings, then it follows that we are the sort of creature whose thoughts can misrepresent. Furthermore, if thinking consists in tokening a mental representation, then it follows that our mental representations can misrepresent.

What does it take for a mental representation to misrepresent? Consider the mental representation of DOG, the extension of which are dogs and only dogs. If one were to token the representation of DOG when pointing to a cat, say, then one would be misrepresenting the cat as a dog. To illustrate, suppose that you see, from a distance, what appears to you as a dog (but in fact is a cat) and then you confidently say to your companion 'Lo, a dog!'. A naturalistic analysis of what is going on would need to involve at least four components. First, there will be the tokening of a representation, whose vehicle would be some sort of neural state or process. Second, this representation would also have an intentional content which, in this case, would be DOG. Third, there would be an intentional object the representation is intended at—what Cummins (1996) calls the *target* of the representation. In this case, the target object would be a particular dog that would have fallen under the extension of DOG. Finally, the fourth component is the *causal* object of the tokening of the representation. In this case, the cause is a cat which does not fall under the extension of DOG. Thus, we have an instance of mismatch between the causal object and the target object and, thus, a case in which the referred object does not fall under the extension of its representational content. Explaining how this could be the case in purely descriptive terms is the challenge faced by those who want to offer naturalistic accounts of mental representation.

Advocates of MVPA argue that they can tell us how brains like ours can represent things like bottles, chairs and dogs. More precisely: some advocates of MVPA claim that they can give us evidence as to how the brain instantiates particular contents that represent intentional objects from categories such as BOTTLE, CHAIR, and DOG. And the relevant notion of representation here appears to be non-natural, as in the case of the aforementioned example of the mental representation of DOG. As a result, if MVPA is supposed to deliver evidence for representational contents in the non-natural sense, then it should be able to tell us when a neural representation gets its target object right and when it does not.

Unfortunately, we argue that if we take the success of a classifier at predicting a pattern of brain activity, X, given instances of category A as sufficient reason to claim that X represents A, then we may face a situation in which the classifier may classify an entirely different stimulus—say an instance of B—as being A without having a principled reason to assert that we have a case of misrepresentation. Let's work this out more carefully. Consider a variation on the thought experiment discussed above, when we presented the problem of cross-cut categories. Suppose, again, that Neuroscientist 1 collected fMRI data while participants were sorting stimuli as being either LIVING or NON-LIVING. She then trained a classifier over a subset of the data from a certain ROI, ran the classifier over the testing set and found that it identified that a particular brain pattern, X, predicted instances of testing stimuli labeled as “living” with extraordinarily high accuracy. Excitedly, Neuroscientist 1 claims that brain pattern X represents LIVING—i.e., that the content of this putative neural representation is such that it refers to living and only living things. Then, Neuroscientist 2 comes along, reclassifies the stimuli as either SMALL or LARGE, trains a classifier over a subset of the data from the same ROI, and finds that X predicts stimuli from the testing set labeled as “small” with incredibly high accuracy as well. Happily, Neuroscientist 2 announces that brain pattern X represents SMALL—i.e., that the content of the putative neural representation is such that its extension covers small and only small things.

But now suppose that Neuroscientist 1 takes a careful look at the testing data and discovers that a seemingly large looking animal—say, a cow—got classified as “non-living”. Being somewhat philosophically inclined, Neuroscientist 1 claims that this case of misclassification constitutes an instance of misrepresentation: the classifier misrepresented the cow as non-living while it should have been living. Neuroscientist 2 also decides to take a closer look at the testing data and notices that there is a particularly living-looking large item—say, a building’s façade[4]—that got classified as “small”. Neuroscientist 2, also in philosophical guise, claims that this case of misclassification constitutes an instance of misrepresentation: the classifier misrepresented a building as small when it should have been large. But here’s the problem: what reason do we have to accept the claim of Neuroscientist 1—namely, that the cow was misrepresented as NON-LIVING when it should have been LIVING—rather than the alternative interpretation, according to which it got misrepresented as LARGE when it should have been SMALL? Likewise, what reason do we have to accept the claim of Neuroscientist 2—namely, that the building was misrepresented as SMALL when it should have been LARGE—rather than the alternative

[4] Perhaps unsurprisingly, there are several websites dedicated to collecting pictures of buildings that look like living things (https://images.app.goo.gl/2NoL9eidmJtSPSQr7).

construal, whereby it got misrepresented as LIVING when it should have been NON-LIVING? Although both seem reasonable interpretations of misrepresentation, they cannot be true at the same time. Unfortunately, we don't seem to have a principled reason to prefer one rather than the other.

This concern harkens back to Fodor's (1987) observation that the causal dependency underlying causal theories of representation needs to be asymmetric: two items, A and B, may co-vary with tokens of mental representation X, but if we claim that X means A and not B, we need to be able to say *why it does not represent B even though it co-varies with A.* Alas, MVPA alone cannot reveal the asymmetry, because—as illustrated by our thought experiment—there is no principled reason to claim that the pattern of brain activity identified by the classifier represents LIVING rather than SMALL (or vice-versa). Moreover, to make things more interesting, it is even possible for a classifier to misclassify an item that the participant did not misrepresent, just as it is possible for a classifier to correctly identify an item a participant did misrepresent. Once again, the gap between the representation and the classification discussed by Klein et al (2019) and Wright (2018) reemerges in the context of misrepresentation. [5]

[5] A reviewer suggested an intriguing possibility to try to circumvent—or, at least, ameliorate—the problem of misrepresentation, namely the inclusion of error trials in the analysis. In fact, as pointed out by the reviewer, this is a strategy that has been employed, for instance, by Woolgar and colleagues (2019) in a challenging stimulus-response task with a stable error rate. The reviewer's suggestion is that these sorts of analyses may be able to tell us when a misrepresentation may occur by, say, "showing that patterns on trials where X is an error are similar to patterns where X is not". Although intriguing, we remain somewhat skeptical. For one, error trials are extremely difficult to analyze, as it is often not clear why a participant erred in each trial. Perhaps in one trial a participant was confused, in another one tired, or even simply pushed the wrong button by accident. Indeed, it is because of the difficulty of determining the reasons behind error trials that most researchers simply discard them. A second reason to be skeptical, is that the few instances (we are aware of) in which error trials have been included in MVPA analyses, is for extremely tight experimental paradigms where there are no more than a couple of options to hit or miss. It is not clear how these kinds of paradigms and analytic strategies can scale up to larger stimuli set. Finally, even if, for the sake of argument, we were to assume that we knew why a participant makes an error—say, a participant incorrectly identifies a dog as a cat—we may still face similar difficulties as those pointed out in the current paper. After all, such a result will only reverse the mapping for error trials (i.e., error trials will be thought to match the incorrect response and *not* the actual target), and thus problems like that of cross-cut categories and orthogonalization (discussed below) would still apply. Nevertheless, we acknowledge that our brief rebuttal is far from being a knock-down counterargument against this intriguing possibility. Further research is needed to evaluate the extent to which the inclusion of error trials can solve the problem of misrepresentation for MVPA.

Notice, once again, that the problem of misrepresentation does not necessarily go away with further experimentation. Sure, Neuroscientist 3 may suggest running an experiment with a 2 x 2 design in which LIVING and NON-LIVING are pitted against SMALL and LARGE. But, again, this not only amounts to caving in and accepting that neither Neuroscientist 1 nor Neuroscientist 2 should have made claims about representation, but it also opens the possibility that there is another co-varying feature, unknown to Neuroscientist 3, for which there will still be cases of misclassification that do not reveal the asymmetry required for them to be accounted for as cases of misrepresentation. Finally, it is worth remarking that concerns about co-extensional stimuli are not new to users of MVPA, and worries about so-called "representational ambiguities" are discussed regularly in the literature (e.g., Carlson & Wardle, 2015). In fact, for the past decade or so, many neuroimagers have advocated for the use of "encoding models"—e.g. voxelwise—which basically impose top-down constrains on the way the acquired data should be structured (Naselaris & Kay, 2015). The problem is that such constrains are always based on the researcher's a priori hypotheses as to how one *should* expect the data to behave, so any read-off is always going to be limited by the experimenters' choices, leaving open the possibility that the labels used in the model are not the correct "labels" from the point of view of the brain. Moreover, the use of explicit encoding models introduces a normative element to the interpretation of the data that should give the naturalist a pause—but then again, perhaps Loewer (2017) is right: the more naturalistic the approach to mental representation is, the less it accounts for content, but the more it accounts for content, the less naturalistic the view is.

4. The problem of orthogonalization

At this point, advocates of MVPA may bite the bullet and accept that the notion of representation they have in mind does *not* conform to the strict notion philosophers have in mind when arguing for naturalistic accounts of mental representation. Instead, they may argue that their notion of representation is weaker, that it admits contents to be underdetermined by reference, and that it need not demand an account of misrepresentation. This weaker notion of representation is close to the sense of *minimal* representation Sullivan (2010) contrasts against a more *substantive* sense of representation in neuroscience, whereby the latter plays explanatory roles in terms of the representation's content, while the former does so only in terms of co-variation. In other words, advocates of MVPA may reason that their use of representation implies nothing more than a reliable co-variation between the neural activity that allegedly carries a particular representational content and a particular stimulus, or class of stimuli, that correspond to the intentional object of said representation. Thus, despite misleading headlines to the effect that MVPA can unveil neural representations of emblematic instances of non-natural meanings, such as complex concepts and semantic categories, in reality we should understand MVPA as providing evidence for a weaker sense of representation, one that reveals mere reliable co-variations between neural activity and stimuli, not unlike typical cases of natural meanings such as the rings of a tree representing its age and, yes, the spots in Tommy's face indicating that he has measles (Dretske, 1986).

Unfortunately, we believe that even if we relax the notion of representation to its minimal, co-variational use, there are still situations in which MVPA does not give us reliable information about the sort of object that is allegedly represented in the pattern of brain activity. To see why, consider a third challenge, which we call "the problem of orthogonalization". The vast majority of machine learning algorithms in cognitive neuroscience classify between two classes. Most algorithms, including the popular support vector machine (SVM), project acquired data into a multi-dimensional space and draw a hyperplane to maximally divide the classes. Moreover, the majority of multiclass classifiers typically utilize something known as a softmax loss function, meaning that the summed probabilities of the output classes equals one. The strategy of dividing between the two classes is done in order to orthogonalize the data so as to maximally classify them. In general, this approach makes good sense, as in almost all cases the neuroscientist's primary goal when using MVPA is to clearly distinguish between categories of interest, and applying a softmax function to a classifier such as SVM tends to contribute to their high accuracy. But it also comes with some drawbacks, especially when the same classifier is applied to a new set of data.

First of all, most MVPA algorithms are likely to ignore features that are identical between classes. For example, if we wanted to classify items as belonging to CAT versus DOG, features such as 'living', 'household pet' and 'furry' are—from the point of view of the classification algorithm—irrelevant, whereas features such as 'aloof' and 'friendly' are critical. In nearly all real-world cases the researcher is blind to these features. Recall the problem of cross-cut categories above where small / large can be conflated with living / non-living dimensions. A key function of the classifier is to extract which features maximally separate categories. (Notice that, in order to better articulate the orthogonalization problem, the features are transparently laid out here; we get back to this point soon). Accordingly, a classifier trained on this information will be sensitive to these critical, unique features. As a consequence, when/if the classifier is applied to a new item, it is going to be forced to classify it as either CAT or DOG. For instance, if the classifier is given a new item—say, a friendly looking teddy bear—it is very likely that it will classify it as DOG with high accuracy, since teddy bears are much more friendly than they are aloof—even though a teddy bear is neither a living being nor a household pet. While arguably a teddy bear is no more a dog than it is a cat, the classifier still classifies it as a dog. But now consider a different new item: a rabbit. As dogs and cats, a rabbit also has the common features ‘living’, ‘household pet’, and ‘furry’ but, critically, rabbits are equally friendly and aloof (or let’s suppose, for the sake of the argument). If this is the case, then the classifier would be greatly challenged, and its output will be no more effective than that of a simple guess.

The upshot of this basic example is that we end up with high classification accuracy for the teddy bear but just chance-level accuracy for rabbit, despite rabbits being much more similar to either cats or dogs than teddy bears are. This highlights a critical point with respect to classifiers: their function is to achieve maximal accuracy, and the best way to do this is to focus on features that maximally separate the categories. As a consequence, when new items are highly similar to the categories the classifier was trained on, the classifier struggles. Furthermore, as the features that the classifier extracts as meaningful are often hidden from us (see footnote 4), the classifier has the capacity to make high-confidence judgements in cases where classification should not have been clear-cut.

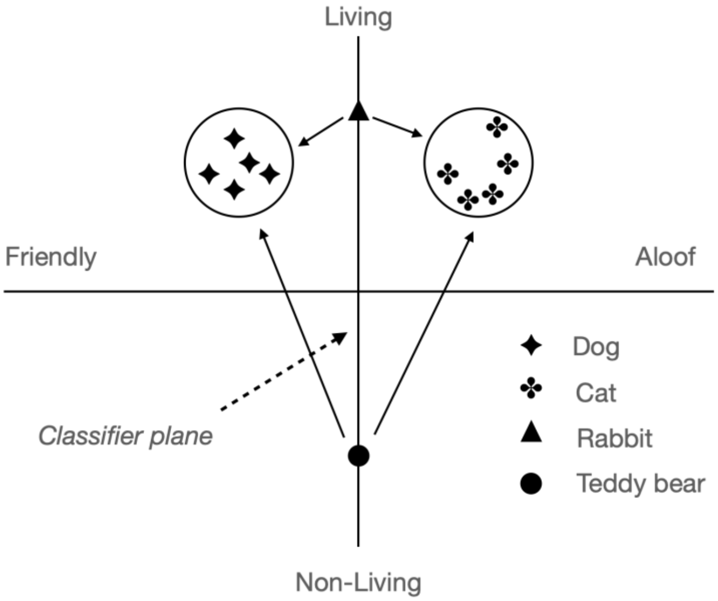

Figure 1: Graphical depiction of a binary classifier on a low dimensional space. The property ‘Friendliness’ varies along the x-axis. X = 0 at the intersection of the classifier plane and the x-axis.

It may be helpful to think of this issue graphically. Assume a low dimensional space, such as that depicted in Figure 1, in which the ‘Friendliness’ dimension ranges over the x axis. Classes are divided where x = 0. Thus, in our case, if x < 0, then the item is ‘Friendly’ and gets classified as DOG. By contrast, if x > 0, then the item is ‘Aloof’ and gets classified as CAT. Moreover, the bigger the magnitude of x and the farther the item is from 0, the more confident the classifier is that the item is either a dog or a cat. But now imagine an item that falls exactly at mid-point, where x = 0—let’s say, a rabbit that is equally friendly and aloof. Here, classification accuracy would be near-chance, as the classifier cannot decide if the rabbit’s x-coordinate is closer to the dog side (more friendly value along the x axis) or closer to the cat side (more aloof value along the x axis). The problem is that the same is true of another item, a teddy bear, who also happens to be equally friendly and aloof and, thus, falls squarely in the mid-point where x = 0. Of course, from the point of view of the y-axis, the point representing the teddy bear is clearly among the non-living items, whereas the point that represents the rabbit is clearly among the living items. Unfortunately, from the classifier’s perspective, both points (i.e., rabbit, teddy bear) are equally distant from the categories of interest (i.e., cat, dog) and, as such, the classifier outputs the same classification uncertainty for both items, despite one of them—the rabbit—being clearly much more related to the categories CAT and DOG than the teddy bear.

This common orthogonalization approach to the classification of new items presents, therefore, a problem of interpretation: low classification accuracy can occur because the new item looks nothing like the old classes *or* because it looks very much like both classes. Unfortunately—and this is critical—we simply cannot tell which from the classifier output alone. Often times, these classification errors are innocuous, but in certain circumstances they could have serious consequences. Consider fields such as associative or reward learning, where MVPA is commonly used (e.g., Visser et al., 2011). Suppose, for instance, that MVPA is used in a learning paradigm, during memory retrieval, to classify a pair-associate. Now imagine a particular instance in which the classifier accuracy is low. What should we conclude from that finding? Should we conclude that the participant was confused and thus retrieved both associated items—and, thus, the relevant brain pattern represents both items—or should we rather say that she was forgetful and retrieved neither—in which case the relevant brain pattern does not represent either associate? Once again, from the classifier output alone we cannot tell what the brain pattern represents.

To be sure, the severity of the problem of orthogonalization has been recognized by several researchers in the past few years, and many have suggested that a way to ameliorate this concern is to add a third *neutral* class (e.g., Rose et al., 2016).[6] In essence, the inclusion of a third class provides researchers with a baseline against which to compare their classes of interest. Here is how it works in practice. Suppose we add a neutral class, roughly characterized as 'neither-cat-nor-dog', and test the algorithm from the previous example against our two new items: a rabbit and a teddy bear. With the neutral class to be tested against, the classifier may output that 'rabbit' has higher evidence for DOG versus NEUTRAL and CAT versus NEUTRAL, whereas 'teddy bear' has low evidence for DOG vs. NEUTRAL and CAT vs. NEUTRAL. We, as researchers, could then conclude that this evidence suggests that rabbit is more like a dog and a cat than a teddy bear is. However, this result comes at the cost of not really knowing on what grounds the classifier is cataloguing an item as more or less similar to the target category, since we, by design, are blind to the features by means of which items are classified as likely belonging to the neutral class. It is

[6] Indeed, today it may be really hard to get a paper accepted using a SVM algorithm with a softmax loss function without the use of a neutral class, even though this was a common strategy just a few years ago.

natural to assume that the neutral class may classify in terms of LIVING or NON-LIVING, but then again, that is the researchers' read-off of the data and, as before, we could find ourselves in the territory of cross-cut categories, for there may always be another, consistent way of interpreting what the neutral class ranges over that would still fit the classifier's output.

Finally, another solution is to get rid of binary classifiers, and accept that they are *not* a requirement for MVPA. This strategy invites us to reject the assumption that items need to be classified in terms of distance to a single category, A or B, and, instead, that they should be classified as either closer to two (or more) categories, A and B, than to an alternative category, C. Unfortunately, for the purposes of giving us information about the representational content of the relevant brain pattern, this strategy massively backfires, as we now can read the evidence as showing that a particular brain pattern represents A or B, but we cannot tell, from the read-off alone, whether it is A and not B, or B and not A. This not only brings back the problem of disjunction discussed by Fodor (i.e., a state represents that *p* only if *p*, not if *p* or *q* (Fodor, 1987)), for which the asymmetric dependency view was proposed as a solution, but also goes against the only requirement a weak construal of representation demands: that a brain state represents that *p* only if *p* is the case.

5. The problem of null results

We finished section 3 conceding that MVPA users who claim that their method gives them evidence for neural representations may not have a strong but rather a weak notion of representation in mind. According to this weak sense of representation, a brain pattern represents that *p* only if *p* is the case—or, to put it differently, that the brain pattern co-varies with its referents.

However, as we discussed in the previous section, the problem of orthogonalization makes it possible for a brain pattern to be identified as representing that *p* when *p* is not the case.

The final problem we discuss also challenges a weak sense of representation. This fourth challenge, which we call “the problem of null results”, is not new to fMRI research (e.g., Cremers, Wager, and Yarkoni, 2017) and was noted briefly by Ritchie, Kaplan and Klein (2019) with respect to MVAP. As they mention, many users of MVPA take it that “poor decodability (or even failure to decode) provides evidence that the information is not represented in that region” (Ritchie et al. 2019, 597). As Ritchie, Kaplan and Klein, we also believe that this claim is false, but we think that it is such a serious concern for MVPA that it deserves further development. Consider, once again, a toy example. Suppose there is a neuroscientist who ran an fMRI experiment in order to determine if a pattern of brain activity in region X represents LIVING or NON-LIVING. She collects a corpus of pictures of living and non-living stimuli and presents them to participants while they are in the scanner. Then the neuroscientist conducts a MVPA analysis on the data, with a focus on the ROI of interest, X, and finds out that the classifier is *unable* to accurately sort the testing data according to the labels ‘living’ or ‘non-living’. As a result, the neuroscientist concludes that brain area X does *not* represent either LIVING or NON-LIVING. Despite the fact that this practice is common among practitioners of MVPA, she would be wrong to do so, as there are a number of reasons why a classifier may fail to associate a pattern of brain activity with certain stimuli, and these reasons have nothing to do with whether or not such a brain structure carries representational information about it.[7]

[7] For example, the theory of *activity-silent* working memory has begun to replace the earlier established hypothesis that working memory is dependent upon sustained neural activity. The *activity-silent* working memory hypothesis arose from theoretical work at the cellular level (Mongillo et al., 2008), but it was supposed to be later confirmed with EEG (Wolff et al., 2015) and fMRI (Rose et al., 2016) MVPA-style analyses via *null results* (i.e., the inability to decode the contents of working memory). Our argument in this section does not imply that something like the activity-silent working memory hypothesis is wrong; we simply want to suggest that the structure of the inference from null results alone to corroborate such a view is, if not inadequate, at best insufficient.

First, a classifier may fail to associate a pattern of brain activity with a certain category because of a bias in the selection of the stimuli. No matter how hard our hypothetical neuroscientist must have worked to come up with a representative sample of pictures of living and non-living entities, it is always possible that the exact same classifier could fare much better, and even yield a much more accurate classification, with an entirely different set of pictures. Alternatively, the reason why the classifier failed to reveal a significant pattern in region X when presented with pictures of living animals may have to do with the fact that the stimuli were static rather than dynamic. Perhaps if the neuroscientist had chosen short clips rather than stills of living and non-living things the classifier would have revealed that activity in brain region X was, in fact, reliably associated with LIVING rather than NON-LIVING. This concern is reminiscent of the problem of cross-cut categories discussed above, but not identical, for the point is that any finding from MVPA, whether positive or negative, is read-off in relation to how researchers choose and label their stimuli. Null claims can just as easily be indicative of a biased sample as they can be a lack of a 'representation'.

A second reason to be skeptical of the neuroscientist's conclusion from a null result pertains to the nature of the classifier itself. MVPA does not refer to a single analytical strategy, but, in practice, to *any* multivariate analysis the researcher wishes to conduct. For example, many researchers utilize *linear* classifiers (e.g., SVMs) that fit straight lines (in 2-D space) to divide stimulus sets (as in the example in Figure 1). However, there is also the option to use *non-linear* classifiers (e.g., artificial neural networks) that can fit much more complex shapes (e.g., curves, circles, and so forth in a 2-D space). In practice, the more complex an algorithm is (i.e. the more parameters it has), the more complex the space it can define. More parameters often require more data to fit, but the general principle holds. But this brings out our critical point: does a null result arise because there is no signal *or* because an inadequate classifier was used? For example, the failure of an SVM says little or nothing about how an artificial neural network might perform. As consequence, suggesting that a region, X, does not ‘represent’ living or non-living items based upon the output of a single classifier is premature.

Finally, a third, well-known reason as to why MVPA can yield a null result concerns the issue of granularity inherent to fMRI. There are roughly 16 billion neurons in the cerebral cortex, and around 60 billion glial cells (Lent et al., 2012). Despite the fact that our most advanced MRI scanners have remarkable resolution, a typical 3 mm isotropic voxel in a functional scan contains approximately 630,000 neurons. That is a lot of neurons. As a result, a classifier may fail to output a successful discrimination simply because the resolution is too coarse. There are several ways in which different neuronal populations within a single voxel could co-vary with different stimuli (Figure 2). Unfortunately, because MVPA with fMRI data still takes as its input the average activity within a single voxel, all these more within-voxel, fine-grained distinctions would be missed by the classifier. That is, they could all produce a null result where a significant result would have been found, had the instrument’s resolution been higher. To complicate things a bit more, there is even suggestive evidence to the effect that a single neuron may be able to carry information from two different stimuli—a process known as multiplexing. If so, even a classifier with a single-cell resolution averaging over the activity of a single neuron would still produce a null result when it probably shouldn’t.

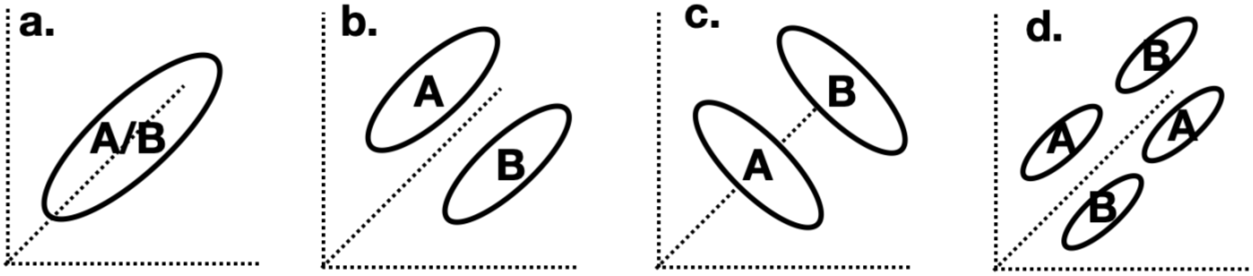

Figure 2: Examples of different ways in which the same average BOLD signal within a single voxel could result from different neuronal populations carrying information from distinct stimuli. a) Single neuronal population in which neurons multiplex for A and B. b) and c) Discrete neuronal populations, each one carrying information about different stimuli, A and B, but a much fine grained resolution than voxel-level. d) Four distinct neuronal populations, each one carrying information about different stimuli at an even lower level of granularity. The dotted lines represent the three dimensions of an isotropic functional voxel.

In sum, in section 4, we offered a possible situation in which the result of a multivariate pattern analysis invites the researcher to claim that a pattern of brain activity in a particular area X represents *p* when p is not the case, while in section 5 we discussed situations in which a researcher takes the null result from the MVPA in a particular area X as indicating that it does not represents that *p* when it is the case that it might. Together, both the problem of orthogonalization and the problem of null results give us reason to doubt that MVPA offers conclusive evidence for even a weak notion of neural representation.

6. Assessment and conclusion

Twenty years ago, MVPA was introduced as a revolutionary method that promised researchers to move beyond simple questions about brain areas preferentially associated with the processing of one or another kind of stimuli, to asking questions about "what information is *represented* in a region" (Haxby et al, 2014). If this is what the method promised, it is unsurprising to see its users making claims about their results in terms of providing evidence for neural representation. With the current paper, we sought to add to a chorus of skeptical voices by offering four challenges to the claim that MVPA does provide evidence for neural representation. The first two challenges—the problem of cross-cut categories and the problem of misrepresentation—target a strong or substantive notion of representation (Sullivan, 2010), while the other two—the problem of orthogonalization and the problem of null results—target a weaker or minimal notion of representation.[8] Our aim was simply to argue that MVPA alone does not provide conclusive evidence for neural representation in either of these two senses.

It is worth concluding with three questions that have been raised across different audiences when drafts of this manuscript have been presented. First, one may wonder whether our challenges do not also apply to pretty much any other neuroimaging technique, in which case one may ask why are we singling out MVPA. The short answer is: yes; it is very likely that all of our challenges—or, at least, an appropriately modified version of them—could apply to just about any neuroimaging technique in the offing. But there is a reason why we chose MVPA: because many of its advocates, as we mention throughout, have singled out MVPA as a method that has "revolutionized fMRI research" (Haxby et al, 2014), for it, unlike other techniques, does provide evidence for neural representations. Our point is simply that it does not: MVPA just does not deliver what it promises.

Second, does that mean that people should stop using MVPA? Not at all. MVPA is a clever method to interrogate neuroimaging data. The fact that it allows us to use principles from machine learning to test classifiers with untrained data certainly helps to find signal in the rather noisy world of neuroimaging data—a signal that likely otherwise would not have been uncovered. But this development does not constitute the revolution some sell it to be. It is just another useful yet gradual advancement that provides users with another analytic tool, not unlike—say—partial least

[8] To be sure, the first two challenges likely apply to a weaker notion of representation as well. We are assuming here that partisans of a weaker notion of representation may feel unaffected by the first two challenges, but it is possible that don't. We thank Zina Ward for mentioning this point.

squares (PLS) analysis (Krishnan et al., 2011) or any other of the several multivariate techniques available to neuroimaging researchers.

Finally, some may wonder at this point whether any technique could ever deliver evidence for neural representation, even in a weaker sense. This question is complex, the answer to which requires much more space than what is available here. But here is our two cents. One the one hand, it is likely that robust evidence for neural representation is going to require more than just *neuroimaging* approaches, and that intervention methods would be required—be they optogenetic, intracranial stimulation, or even some yet-to-be-developed ones. This is partly because of the asymmetric dependency of mental/neural representations mentioned above: unless we can move from mere statistical co-variation to causation we likely won't be able to get rid of the challenges we offer here. On the other hand, the solution won't come from technological developments alone.. There is also a need for theory development. Consider the sacrosanct example of Hubel & Wiesel (1959), in which their results are interpreted as providing evidence that single neurons in the striate cortex code for orientation. These conclusions are open to question under a different theoretical framework, such as that of predictive coding, according to which the neuron does not represent orientation but the mismatch between the expected model and the incoming retinal signal (Rao and Ballard, 1999). Once again, we find ourselves in gavagai territory. Undoubtedly, the path of contemporary cognitive neuroscience toward neural representation faces formidable challenges, and it is time to recognize that many of them had been anticipated, years ago, by philosophers who understood how difficult it would be to ever naturalize intentionality.

References

Anderson, M. L., & Oates, T. (2010). A critique of multi-voxel pattern analysis. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32), 7.

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage*, *180*, 88-100.

Carlson, T. A., Schrater, P., & He, S. (2003). *Patterns of Activity in the Categorical Representations of Objects*. *15*(5), 21.

Carlson, T. A., & Wardle, S. G. (2015). Sensible decoding. *NeuroImage*, *110*, 217–218. https://doi.org/10.1016/j.neuroimage.2015.02.009

Chisholm, R.M. (1957). *Perceiving: A Philosophical Study*. Ithaca: Cornell University Press.

Chomsky, Noam (1968). *Language and Mind*. Cambridge University Press.

Cremers, H.R., Wager, T.D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *Plos ONE*, 12(11): e0184923.

Cummins, R. (1996) *Representations, Targets and Attitudes*, Cambridge, Mass: MIT Press.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: Practices and pitfalls: Representational analysis using fMRI. *Annals of the New York Academy of Sciences*, *1296*(1), 108–134. https://doi.org/10.1111/nyas.12156

Dretske, F. (1986), "Misrepresentation", in Radu Bogdan (ed) Belief: Form, Content and Function, New York: Oxford: 17–36.

Dubois, J., de Berker, A. O., & Tsao, D. Y. (2015). Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. *Journal of Neuroscience*, *35*(6), 2791–2802. https://doi.org/10.1523/JNEUROSCI.4037-14.2015

Fan, M., & Chou, C.-A. (2016). Exploring stability-based voxel selection methods in MVPA using cognitive neuroimaging data: A comprehensive study. *Brain Informatics*, *3*(3), 193–203. https://doi.org/10.1007/s40708-016-0048-0

Frege, F.L.G. (1892) 'Über Sinn und Bedeutung'. in Zeitschrift für Philosophie und philosophische Kritik, 100: 25–50

Fodor, J.A. 1987, Psychosemantics: The Problem of Meaning in the Philosophy of Mind, Cambridge, MA: MIT Press, Bradford Books.

Fodor, Jerry A. (1990). *A Theory of Content and Other Essays*. MIT Press.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 8.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–573. https://doi.org/10.1038/9224

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., & Carlson, T. A. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *NeuroImage, 180*, 41-67.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. https://doi.org/10.2307/2182440

Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haxby, James V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Johnson, M. R., & Johnson, M. K. (2014). Decoding individual natural scene representations during perception and imagery. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00059

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). *The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception*. *17*(11), 4302–4311.

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Konkle, Talia, & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124. https://doi.org/10.1016/j.neuron.2012.04.036

Lent, R., Azevedo, F. A. C., Andrade-Moraes, C. H., & Pinto, A. V. O. (2012). How many neurons do you have? Some dogmas of quantitative neuroscience under revision. *European Journal of Neuroscience*, *35*(1), 1–9. https://doi.org/10.1111/j.1460-9568.2011.07923.x

Loewer, B. (2017). A Guide to Naturalizing Semantics. In B. Hale, C. Wright, & A. Miller (Eds.), *A Companion to the Philosophy of Language* (pp. 174–196). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118972090.ch8

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Naselaris, T., & Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends in Cognitive Sciences*, *19*(10), 551–554. https://doi.org/10.1016/j.tics.2015.07.005

Neander, K. (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, *79*(2), 109–141. https://doi.org/10.1007/BF00989706

Neander, K. and Schulte, P. (2021). "Teleological Theories of Mental Content", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/spr2021/entries/content-teleological/>.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1, Supplement 1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Quine, W.V.O. (1951) Two Dogmas of Empiricism. *Philosophical Review*, 60: 20–43;

Quine, W.V.O. (1960). *Word and Object*, Cambridge, Mass.: M.I.T. Press, 1960.

Quine, W.V.O. (1986). "Reply to Paul A. Roth" in Hahn and Schilpp (eds.), T*he Philosophy of W. V. Quine*. Peru, IL: Open Court. pp 469-461.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science*, *70*(2), 581–607. https://doi.org/10.1093/bjps/axx023

Rao, R.P.N. & Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 2(1): 79-87.

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Serences, J. T., & Boynton, G. M. (2007). The Representation of Behavioral Choice for Motion in Human Visual Cortex. *Journal of Neuroscience*, *27*(47), 12893–12899. https://doi.org/10.1523/JNEUROSCI.4021-07.2007

Soames, Scott (1999). The indeterminacy of translation and the inscrutability of reference. *Canadian Journal of Philosophy* 29 (3):321-370.

Sullivan, J. A. (2010). A Role for Representation in Cognitive Neurobiology. *Philosophy of Science*, *77*(5), 875–887. https://doi.org/10.1086/656818

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative Learning Increases Trial-by-Trial Similarity of BOLD-MRI Patterns. *Journal of Neuroscience, 31*(33), 12021–12028. https://doi.org/10.1523/JNEUROSCI.2178-11.2011

Wang, J., Baucom, L. B., & Shinkareva, S. V. (2013). Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*, *34*(5), 1133–1147. https://doi.org/10.1002/hbm.21498

Weaverdyck, M. E., Lieberman, M. D., & Parkinson, C. (2020). Tools of the Trade Multivoxel pattern analysis in fMRI: A practical introduction for social and affective neuroscientists. *Social Cognitive and Affective Neuroscience*, *15*(4), 487–509. https://doi.org/10.1093/scan/nsaa057

Woolgar, A., Dermody, N., Afshar, S., Williams, M.A., & Rich, A.N. (2019). Meaningful patterns of information in the brain revealed through analysis of errors. bioRxiv 673681: doi: https://doi.org/10.1101/673681

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00123