

Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience

J. Brendan Ritchie, David Michael Kaplan, and Colin Klein

Abstract

Since its introduction, multivariate pattern analysis (MVPA), or “neural decoding”, has transformed the field of cognitive neuroscience. Underlying its influence is a crucial inference, which we call 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. Although the Dictum is a widely held and well-motivated principle in decoding research, it has received scant philosophical attention. We critically evaluate the Dictum, arguing that it is false: decodability is a poor guide for revealing the content of neural representations. However, we also suggest how the Dictum can be improved on, in order to better justify inferences about neural representation using MVPA.

1. *Introduction*
2. *A Brief Primer On Neural Decoding: Methods, Application, And Interpretation*
 - 2.1 *What is MVPA?*
 - 2.2 *The informational benefits of MVPA*
3. *Why The Decoder’s Dictum Is False*
 - 3.1 *We don’t know what information is decoded*
 - 3.2 *The theoretical basis for the dictum*
 - 3.3 *Undermining the theoretical basis*
4. *Objections And Replies*
 - 4.1 *Does anyone really believe the dictum?*
 - 4.2 *Good decoding is not enough*
 - 4.3 *Predicting behaviour is not enough*
5. *Moving Beyond The Dictum*
6. *Conclusion*

1. Introduction

Since its introduction, multivariate pattern analysis (MVPA)—or informally, neural ‘decoding’—has had a transformative influence on cognitive neuroscience. Methodologically, it is a veritable multi-tool that provides a unified approach for analyzing data from cellular recordings, fMRI, EEG, and MEG, which can also be paired with computational modeling and behavioral paradigms (Kriegeskorte *et al.* [2008]). Theoretically, it is often presented as a means for investigating the structure and content of the brain’s population code, thereby unifying

40 psychological and neuroscientific explanations while predicting behavioral performance (Haxby
41 *et al.* [2014]; Kriegeskorte and Kievet [2013]). More ambitiously still, decoding methods are
42 advertised as a means of ‘reading’ the brain and ‘listening’ in on the mind (Haynes and Rees
43 [2006]; Norman *et al.* [2006]).

44
45 Underlying these bold pronouncements is a crucial inference, which we call the Decoder's
46 Dictum:

47
48 If information can be decoded from patterns of neural activity, then this provides
49 strong evidence about what information those patterns represent.

50
51 The Decoder's Dictum should interest philosophers for two reasons. First, a central philosophical
52 issue with neuroimaging is its use in ‘reverse inferences’ about mental function (Poldrack
53 [2006]; Klein [2010]). The Decoder's Dictum is a similar but more nuanced form of inference, so
54 it deserves careful scrutiny. Second, decoding results are some of the most compelling in
55 cognitive neuroscience, and offer a wellspring of findings that philosophers may want to tap into
56 when defending theoretical claims about the architecture of the mind and brain.¹ It is therefore
57 worth clarifying what decoding can really show.

58
59 We argue that the Decoder's Dictum is false. The Dictum is underwritten by the idea that
60 uncovering information in neural activity patterns, using ‘biologically plausible’ MVPA methods
61 that are similar to the decoding procedures of the brain, is sufficient to show that this information
62 is neurally represented and functionally exploitable. However, as we are typically ignorant of the
63 precise information exploited by these methods, we cannot infer that the information decoded is
64 the same information the brain exploits. Thus decodability is not (by itself) a reliable guide to
65 neural representation. Our goal is not to reprimand neuroscientists for how they currently employ
66 and interpret MVPA. Rather, what follows will clarify the conditions under which decoding
67 could provide evidence about neural representation.

68

¹ A recent example: in arguing against the encapsulation of the visual system, Ogilvie and Carruthers ([2016]) rely almost exclusively on decoding results about early vision since they believe it provides more convincing evidence than behavioural research.

69 By analogy, consider research on brain-machine interface (BMI) systems, which use decoding to
70 generate control signals for computer cursors or prosthetic limbs (Hatsopolous and Donoghue
71 [2009]). Largely because of BMI's engineering and translational objectives, however, little
72 attention is paid to the biological plausibility of decoding methods. Consequently, BMI research
73 does not involve inferences about neural function based on decodability. We believe that,
74 epistemically, decoding in cognitive neuroscience is typically no better off than in BMI research,
75 and so forms a thin basis for drawing inferences about neural representation.

76
77 Our focus is on how MVPA is used to investigate neural representations. Since talk of
78 representation is itself philosophically contentious, we assume a relatively lightweight notion
79 that is consistent with usage in the relevant sectors of neuroscience: a representation is any
80 internal state of a complex system that serves as a vehicle for informational content and plays a
81 functional role within the system based on the information that it carries (Bechtel [1998]).² As
82 we shall see, some researchers talk of decoding mental representations. We assume they have in
83 mind at least the notion of (distributed) internal representation we have articulated, so our
84 arguments apply to their claims as well.

85
86 We focus on neural representations that take the form of population codes. A population code
87 represents information through distributed patterns of activity occurring across a number of
88 neurons. In typical population coding models, each individual neuron exhibits a distribution of
89 responses over some set of inputs, and for any given input, the joint or combined response across
90 the entire neural population encodes information about the input parameters (Pouget *et al.*
91 [2000]).

92

² One may reasonably wonder whether this characterization captures scientific usage. Although foundational concepts like 'representation' are rarely explicitly defined by neuroscientists, there are exceptions. For example, Marr ([1982], pp. 20-1) defines a representation as 'a formal system for making explicit certain entities or types of information', and Eliasmith and Anderson ([2003], p. 5) state that: '[r]epresentations, broadly speaking, serve to relate the internal state of the animal to its environment; they are often said to "stand-in for" some external state of affairs.' Along similar lines, deCharms and Zador ([2000], p. 614) define a representation as a 'message that uses [...] rules to carry information' and define content as the 'information that a representation carries'. Our discussion of the theoretical basis for the Dictum (section 3.2) also illustrates that something close to the above notion is widely assumed by researchers in the field.

93 Our critique of the Dictum will take some setup. In section 2, we provide a brief introduction to
94 decoding methods. In section 3, we argue that the Dictum is false: the presence of decodable
95 information in patterns of neural activity does not show that the brain represents that
96 information. Section 4 expands on this argument by considering possible objections. In section 5,
97 we suggest a way to move beyond the Dictum. Section 6 concludes the paper.

98

99 **2. A Brief Primer On Neural Decoding: Method, Application, And Interpretation**

100 We begin by providing a brief introduction to basic decoding methods and their interpretation.
101 We focus primarily on research that has used MVPA with fMRI to investigate the visual system.
102 There are three reasons for this narrow focus. First, decoding research on vision is largely
103 responsible for popularizing MVPA. Second, it has also driven many of the methodological
104 innovations in the field. Third, it is instructive because we have a detailed understanding of the
105 functional organization of many visual brain regions along with good psychophysics (Haxby
106 [2012]). Thus, if the Dictum is viable at all, it should apply to decoding research on the visual
107 system.

108

109

2.1 What is MVPA?

110 Multivariate pattern analysis (MVPA) is a set of general methods for revealing patterns in neural
111 data.³ It is useful to separate MVPA into three distinct stages (Mur *et al.* [2009]; Norman *et al.*
112 [2006]), which we will illustrate via a simple (hypothetical) fMRI experiment. In this
113 experiment, fMRI BOLD responses are measured while participants view two gratings of
114 different orientations over a number of trials (Figure 1A). The goal of the experiment is to test
115 whether the activity patterns elicited in response to the two stimulus conditions can be
116 differentiated.

117

118 The first step of analysis, pattern measurement, involves collecting neuroimaging data that
119 reflects condition-dependent patterns of activity. This step has a number of components,

³ Some terminological points. First, ‘MVPA’ originally meant ‘multi-*voxel* pattern analysis’, rather than ‘*multivariate* pattern analysis’. The latter is preferable because it highlights the fact that the methods are not specific to fMRI (Haxby [2012]). Second, ‘MVPA’ and ‘decoding’ are sometimes used interchangeably (as we do), but strictly speaking decoding methods are a subset of MVPA methods (Naselaris *et al.* [2011]). And third, ‘decoding’ is often used in two distinct senses: a machine learning sense, in which it is basically a synonym for ‘classify’; and a neural sense, referencing the encoding and decoding of signals by the brain. We make use of both senses here.

120 including performing the actual recordings and preprocessing of the activity-dependent signal.
121 Our example uses fMRI, but other techniques (for example, EEG, MEG, or cellular recordings)
122 could also be employed. As in all fMRI experiments, we must make certain assumptions about
123 the connection between the recorded signals and underlying neural activity.⁴ Nevertheless, the
124 end result is the same: a set of data consisting of multiple distinct measurements of activity
125 occurring during each experimental condition.

126
127 The second step, pattern selection, involves focusing in on a subset of the measured signals for
128 further analysis. With fMRI, this involves a subset of all voxels or a ‘region of interest’ (ROI).
129 ROIs can be defined anatomically (using connectivity patterns or architectonic criteria) and/or
130 defined functionally (using neural response profiles or more traditional univariate fMRI
131 analyses). Pattern selection also depends on experimenter goals and recording technique. In our
132 experiment (Figure 1B) the ROI is parafoveal primary visual cortex (V1), defined anatomically
133 (Benson *et al.* [2012]).

134
135 The third, and crucial, step is pattern classification. Pattern classification allows one to measure
136 the discriminability of different patterns in multivariate data. For example, in our experiment we
137 want to see if the patterns of BOLD activity in parafoveal V1 for our two stimulus conditions can
138 be distinguished (Figure 1C). A number of classification methods are available. The simplest is
139 to divide the data in half for each stimulus condition and compute the within- and between-class
140 correlations of the patterns (Haxby *et al.* [2001]). If the patterns are discriminable, the within-
141 class correlation should be higher.

142
143 A more powerful (and widely used) technique employs machine learning classifiers, which treat
144 each element of the patterns of interest (e.g., each voxel) as a separate dimension, or ‘feature’, in

⁴ It is well-known that the signals measured with neuroimaging techniques such as fMRI and MEG/EEG depend on neural activity, but often in complicated and indirect ways (e.g., Logothetis [2008]; Nir *et al.* [2008]; Singh [2012]). For example, fMRI measures blood oxygenation level-dependent (BOLD) signals reflecting changes in cerebral blood flow (CBF), cerebral blood volume (CBV), and cerebral metabolic rate of oxygen consumption (CMRO₂) following neural activity. Although it remains controversial precisely which types of neural responses induce these haemodynamic changes (e.g., Logothetis *et al.* [2001]; Sirotin and Das [2009]; Lee *et al.* [2010]), applications of MVPA typically assume that neuroimaging techniques coarsely measure the spatial structure and temporal dynamics of local neuronal populations. It is therefore common to use the term ‘activity patterns’ to describe the multivariate data collected with these techniques, even though, strictly speaking, MVPA is not being used to analyse neural activity patterns directly. We also adopt this convention.

145 a high-dimensional space. Assuming our ROI includes N voxels, then each trial-wise stimulus
146 presentation elicits a pattern that occupies a point in an N -dimensional neural activation space.
147 The goal of the classifiers is to find a way to transform this high-dimensional space into one
148 where the voxel patterns associated with each condition are separable by a decision boundary
149 (Figure 1D).

150
151 Although a rich variety of classifiers are available, usually simple linear classifiers are used for
152 MVPA because they provide a principled means of estimating a linear boundary between classes
153 in activation space. To avoid overfitting, the decision boundary is estimated for a subset of the
154 data designated as ‘training’ data, and the classifier is subsequently ‘tested’ on the remaining
155 data (Figure 1D). The classifier assigns condition labels for the training data based on the
156 position of the activity patterns relative to the decision boundary. The performance of the
157 classifier is then a function of the accuracy of its label assignments (for example, % correct;
158 Figure 1D). Training and testing is done multiple times, with each data partition taking its turn as
159 the testing data, and the performance of the classifier is then averaged across iterations. If the
160 mean classifier performance is statistically better than chance, the patterns for the different
161 conditions are considered to be discriminable. Although applications are typically far more
162 complex than what we have presented here, at root all decoding analyses make use of either
163 correlations or machine learning classifiers.

164

165 **2.2 The informational benefits of MVPA**

166 Before we turn to the Dictum, it is worth considering the advantages of MVPA over more
167 traditional univariate analysis methods. To do this we adapt a distinction from Kriegeskorte and
168 Bandettini ([2007]) between activation-based and information-based analyses of neuroimaging
169 data. Activation-based analysis involves spatially averaging activity across all voxels within a
170 given ROI, yielding a single measure of overall regional activation to correlate with the tested
171 conditions. By contrast, information-based analysis looks for a statistical dependency between
172 experimental conditions and the detailed local spatiotemporal activity patterns distributed across
173 the set of individual voxels comprising the ROI (see, for example, Haxby *et al.* [2014]; Tong and
174 Pratte [2012]). Hence, what distinguishes the two approaches is whether or not they are sensitive
175 to spatial patterns in fMRI data. Information-based approaches are so-called because they are

176 sensitive to information contained in these spatial patterns. In contrast, the spatial averaging at
177 the heart of activation-based analyses obscures this information.

178
179 All MVPA methods are information-based. Consequently, whatever the status of the Dictum,
180 MVPA decoding holds an advantage over most univariate methods because it offers more
181 spatially sensitive dependent measures. Demonstrating that information is present in activity
182 patterns is also likely to have greater functional significance given the widely held assumption
183 that the brain is an information-processing system that uses population coding to implement its
184 internal representations (Pouget *et al.* [2000]; Panzeri *et al.* [2015]). For example, in fMRI
185 research, activation-based methods are often used to infer that a brain region is involved in some
186 mental process given its engagement during an experimental condition. But as a dependent
187 measure, mean BOLD activity itself likely has no obvious functional significance. Similarly, the
188 evoked responses that are the focus of traditional EEG and MEG analysis are not signals that the
189 brain itself processes. In contrast, if the brain uses population codes, searching for information in
190 patterns of activation means looking for the currency in which the brain makes its transactions.

191
192 As an illustration of the informational benefits of MVPA over univariate methods, consider the
193 early findings of Haxby *et al.* ([2001]). Traditional univariate methods had previously been used
194 to isolate the ‘fusiform face area’ (FFA) within the temporal cortex, which had been interpreted
195 as a highly specialized face-processing ‘module’ in the ventral visual stream (Kanwisher *et al.*
196 [1997]). Haxby *et al.* used MVPA to show that face information was discriminable in the ventral
197 stream even when FFA was removed from the analysed ROI. Hence, their results demonstrated
198 that decoding methods could reveal information present in brain activity that was otherwise
199 undetectable by traditional methods. The results of Haxby *et al.* not only illustrated the greater
200 sensitivity of decoding methods, but also made explicit the idea that decoding was potentially
201 useful for revealing distributed representations in the brain.

202
203 In summary, univariate ‘activation-based’ analyses often obscure the information latent in spatial
204 patterns of neural activity, while decoding affords a powerful tool for revealing this information.
205 If the brain uses population codes, then spatial patterns in neural data that differentiate between
206 conditions should be recoverable using information-based MVPA methods.

207

208

3. Why The Decoder's Dictum Is False

209 Significant decoding indicates that information is latent in patterns of neural activity. However,
210 researchers often draw a further inference: if there is decodable information, then there is strong
211 evidence that the information is represented by the patterns of activity used as the basis for the
212 decoding.

213

214 For example, Kriegeskorte and Bandettini ([2007], p. 658) claim that information-based analyses
215 including MVPA 'can help us look into [brain] regions and illuminate their representational
216 content'. and go so far as to define decoding as 'the reading out of representational content from
217 measured activity' (p. 659). Similarly, in comparing and contrasting different fMRI analysis
218 techniques, Davis and Poldrack ([2013], p. 120) state that '[w]hereas univariate analysis focuses
219 on differences in mean signal across regions of cortex, MVPA focuses on the informational
220 content of activation patterns coded in different regions'. We have dubbed this further inference
221 the Decoder's Dictum. Although the Dictum is commonplace, exceptions can be found where
222 decodability is observed but the interpretation of the results does not reflect this problematic
223 inference. Instead, decodability is taken as evidence of functionally specialized processing rather
224 than representational content (Davis and Poldrack [2013]).

225

226 The many fMRI decoding studies looking at top-down effects of visual and cognitive processing
227 on primary visual cortex (V1) provide a good illustration. For example, Williams *et al.* ([2008])
228 presented simple object exemplars in the visual periphery, and found that object shape could be
229 decoded from foveal V1. Jehee *et al.* ([2011]) similarly found that if two orientation grating
230 stimuli were presented in the periphery, but only one was attended to, this resulted in greater
231 classification accuracy for the orientation of the attended stimulus. Both of these results were
232 interpreted as providing evidence of attention-driven feedback to primary visual cortex. In
233 another study, Kok *et al.* ([2012]) found that when the orientation of a grating corresponded with
234 an observer's expectations, this resulted in lower BOLD activity but higher classification
235 accuracy. Again, the focus was on showing that early visual processing can be modulated by
236 expectations. Finally, Harrison and Tong ([2009]) found that stimulus information in a working
237 memory task could be decoded from V1 over a prolonged period of time, suggesting a

238 recruitment of the region for preserving stimulus information for later recall. The common goal
239 of these studies is to reveal facts about functional processing or localization, not representational
240 content.

241
242 In what follows, we defend the strong claim that the Decoder's Dictum is false: successful
243 decoding of information does not provide reasonable grounds for the inference that patterns of
244 neural activity represent the conditions (or aspects of the conditions) about which they carry
245 information. For some philosophers, this might sound like a trivial point: of course we cannot
246 make inferences from information to representation, as there is more to representation than
247 merely carrying information. Fair enough. Yet the problem is not (just) that informational
248 content comes too cheaply in comparison to representational content (Fodor [1984]). For even if
249 we accept that neural representations have content that is partially, or wholly, determined by
250 information, there are several reasons for thinking that the Dictum fails to hold. In the rest of this
251 section, we argue that a fundamental methodological issue with MVPA—specifically, the
252 uncertainty regarding the information exploited by linear classifiers—shows why the Dictum is
253 false.

254

255 **3.1 We don't know what information is decoded**

256 The Dictum entails that if a classifier can discriminate between conditions, then it is picking up
257 on the same information encoded by underlying neural representations. The problem is that we
258 rarely know what information a classifier actually relies on. Indeed, this is most obvious in cases
259 where we know a good deal about what a brain region represents.

260

261 To illustrate, consider again V1, where we have a reasonably good understanding of how
262 orientation information is encoded (see, for example, Priebe and Ferster [2012]). Orientation-
263 related information is also highly decodable using fMRI and MVPA (Haynes and Rees [2005];
264 Kamitani and Tong [2005]). And yet, we do not know what information classifiers are extracting
265 from this region. Indeed, it is something of a mystery why fMRI decoding in the region even
266 works at all. A typical voxel during a functional scan has a much coarser spatial resolution ($> 2 \times$
267 2×2 mm) than the scale of the cortical columns that code for orientation in this region (~ 2 mm
268 in humans; ~ 1 mm in monkeys). This means that one plausible explanation about how decoding

269 works—that patterns of activity across orientation columns occur at a spatial scale roughly
270 commensurate with the resolution of fMRI—cannot be correct.

271
272 There are a number of competing hypotheses about how orientation decoding in V1 is possible.
273 Imperfect sampling of the underlying orientation columns might result in small biases at the
274 voxel level, which decoding exploits, resulting in ‘hyperacuity’ or sub-voxel resolution (Haynes
275 and Rees [2005]; Kamitani and Tong [2005]). Another possibility is that biases in the retinotopic
276 map in V1 (in particular, radial biases) enable successful orientation decoding (Mannion *et al.*
277 [2009]; Freeman *et al.* [2011]). Yet a third possibility is that activity patterns elicited by stimulus
278 edges, not sampling or retinotopic biases, provide a potential source of decodable information in
279 V1 (Carlson [2014]). Note here that the ‘biases’ appealed to in the explanations of orientation
280 decoding are (in some important sense) artifacts in the way the data presented to the classifier is
281 structured, rather than deep facts about the representational structure of the brain. So long as
282 there is any information that distinguishes the conditions at hand, a linear decoder stands a good
283 chance of finding it.

284
285 These issues are not restricted to decoding orientation in V1. For instance, it has been found that
286 motion information decoding is more robust in V1 than V5/MT+ (Kamitani and Tong [2006];
287 Seymour *et al.* [2009]). This result is surprising when one considers that the majority of MT+
288 cells encode motion direction, while < 50 % of V1 neurons exhibit motion sensitivity and the
289 region does not have cortical columns for motion direction as it does for orientation (Lu *et al.*
290 [2010]). Wang *et al.* ([2014]) observe a direction-selective response bias that appears to explain
291 this contrast between decoding and underlying functional organization—it is present in V1-V3
292 but not in MT+—suggesting that motion decoding in early visual cortex bears little relation to
293 the actual encoding structure of these regions.

294
295 Thus, the fact that decoding can pick up on information unused by the brain, even in regions
296 where there is a suitable representation that is used (for example, orientation representation in
297 V1), means that even when prior theory and decoding are in agreement, decoding results cannot
298 be reliably interpreted as picking up on the information that is neurally represented and used. All
299 the worse, then, when we do not have converging evidence and prior theory. This epistemic

300 uncertainty regarding the source of decodable information cuts to the core of the theoretical
301 rationale for the Dictum. It is for this reason it is false, as we will illustrate by reconstructing the
302 theoretical basis for the Dictum. Although appeals to the Dictum are commonplace in research
303 using MVPA (a point we will return to), the theoretical basis for the Dictum is often
304 underspecified. Here we reconstruct the rationale. Doing so demonstrates why epistemic
305 uncertainty regarding the source of decodable information is fatal for the Dictum.

306

307

3.2 The theoretical basis for the dictum

308 The Decoder's Dictum licenses inferences from decodability to facts about neural representation.
309 The principle is evidential: if we can decode, we have reasonably strong evidence about what is
310 represented in the measured patterns of neural activity. But why think the Dictum is true? Here
311 we reconstruct what we take to be the underlying theoretical basis for the Dictum.

312

313 The support for the Dictum starts with two seemingly uncontroversial claims. The first is that if
314 activity patterns occurring in different experimental conditions are discriminable, then
315 information about the conditions is latent in these patterns. The second is that if activity patterns
316 represent information about an experimental condition, then there must be some way to decode
317 that content from the neural patterns. In other words, if internal representations are implemented
318 in patterns of neural activity, and the brain is an encoder and decoder of its own neural signals,
319 then the information must be decodable—that is, after all, what makes it a code. While
320 substantive, these assumptions are not enough to get us to the Dictum. For all we have said,
321 representations present in the brain might not have the right relationship to information extracted
322 by MVPA when applied to the data recorded with standard neuroimaging techniques.

323

324 Two additional steps are required. The first secures the link between information and
325 representation. This requires something like an informational approach to internal
326 representations and their content. The presence of a statistical dependency or correlation is of
327 interest because it suggests a causal dependency between the patterns and the experimental
328 conditions (cf. Dretske [1983]). So charitably, the notion of information that researchers have in
329 mind is that of natural information, where an event carries natural information about events that
330 reliably cause it to occur (Scarantino and Piccinini [2010]). The view, which many in the field

331 endorse, is very similar to Dretske's ([1988]): a representation is a state that carries natural
332 information, appropriately formatted to function as a state carrying this information.

333

334 For example, Cox ([2014], p. 189) notes that decoding research on the visual system:

335

336 implicitly recognizes that the problem of vision is not one of information content, but of
337 format. We know that the activity of retinal ganglion cells contains all of the information that
338 the visual system can act upon, and that nonlinearity and noise in neuronal processing can
339 only decrease (and never increase) the absolute amount of information present. However, the
340 information present in the firing of retinal ganglion cells is not in a format that can be easily
341 read-out by a downstream neuron in order to guide action.

342

343 In other words, vision repackages the information latent in the retinal input to make it
344 functionally available for downstream perceptual and cognitive processing. A simple
345 informational theory of representational content has as a corollary the idea that we can
346 distinguish between implicit and explicit information (Kirsh [1990]), where being 'implicit' or
347 'explicit' is understood as being relative to some procedure for reading-out the information based
348 on how a code is structured. Why should we think that successful decoding allows us to make an
349 inference about what information is explicitly represented by a population code? This question
350 brings us to the second additional assumption: the biological plausibility of MVPA methods in
351 general, and linear classifiers in particular.

352

353 Many views of population coding assume that information can be read out by some sort of linear
354 combination of components to the code. If so, then properties of the code can be made salient in
355 the appropriate activation space. As Kriegeskorte and Kievet ([2013], p. 401) put it:

356

357 We interpret neuronal activity as serving the function of representing content, and of
358 transforming representations of content, with the ultimate objective to produce successful
359 behaviors [...] The population of neurons within an area is thought to jointly represent the
360 content in what is called a neuronal population code. It is the pattern of activity across
361 neurons that represents the content [...] We can think of a brain region's representation as a

362 multidimensional space [...] It is the geometry of these points that defines the nature of the
363 representation.

364
365 Now comes the crucial step. If population coding does indeed involve linear combination of
366 elements, then MVPA is a plausible way to extract that information. For ultimately, a linear
367 classifier is a biologically plausible yet abstract approximation of what the brain itself does when
368 decoding its own signals (DiCarlo and Cox [2007]; King and Dehaene [2014]). In other words,
369 because of the biological plausibility of linear classifiers, significant decodability is taken as
370 evidence that the latent information in the data is also explicitly represented in the brain.

371
372 It is explicitly assumed in the field that linear decodability suffices to reveal an explicit
373 representation. In fact, Kriegeskorte and Kievet ([2013], p. 402) go so far as to define explicit
374 representation in such terms, claiming that ‘if the property can be read out by means of a linear
375 combination of the activities of the neurons [...] the property is explicitly represented.’

376
377 Misaki *et al.* ([2010], p. 116) offer a similar characterization of when information is explicit:

378
379 Linear decodable information can be thought of as “explicit” in the sense of being amenable
380 to biologically plausible readout in a single step (i.e. by a single unit receiving the pattern as
381 input) [...] Linearly decodable information is *directly* available information [...]

382
383 So the decoding of a linear classifier serves as a surrogate for the decoding of the brain. If the
384 linear classifier can use information latent in neural activity, then this information must be used
385 (or usable) by the brain: decoding provides evidence of an encoding.

386
387 In summary, one gets to the Decoder’s Dictum by endorsing several claims: (1) that MVPA
388 reveals information latent in neural activity; (2) that an underlying neural population code
389 implies decodability; (3) an informational view of neural representations and their contents; and
390 (4) the hypothesis that biologically plausible linear classifiers are sufficiently similar in
391 architecture to the decoding procedures employed by the brain. The latter is what lets us infer
392 that decodable information is appropriately formatted for use by the brain, even when we do not
393 necessarily know what that format is. So (5): if we can decode information from patterns of

394 activity using MVPA, this provides good evidence in favor of the hypothesis that the patterns
395 represent the information. Which is just a restatement of the Dictum.

396

397

3.3 Undermining the theoretical basis

398 We are now in a position to see precisely why the Dictum is false. For starters, note that a
399 version of the Dictum appealing to nonlinear classifiers would be summarily rejected by
400 researchers, as one cannot make an inference about what information is represented by patterns
401 of neural activity using overpowered, biologically implausible nonlinear methods. For example,
402 Kamitani and Tong ([2005], p. 684) were the first to caution against the use of nonlinear
403 classifiers:

404

405 [...] nonlinear methods may spuriously reflect the feature-tuning properties of the pattern
406 analysis algorithm rather than the tuning properties of individual units within the brain. For
407 these reasons, it is important to restrict the flexibility of pattern analysis methods when
408 measuring ensemble feature selectivity.

409

410 Along the same lines, Naselaris *et al.* ([2011]) point out that nonlinearity should be avoided
411 precisely because it is too powerful: it allows us to pull out information that is present in the
412 brain, but that could not be exploited by the brain itself. Hence even though:

413

414 [i]n theory a sufficiently powerful nonlinear classifier could decode almost any arbitrary
415 feature from the information contained implicitly within an ROI...a nonlinear classifier can
416 produce significant classification even if the decoded features are not explicitly represented
417 within the ROI. (Naselaris *et al.* [2011], p. 404).

418

419 The concern is that information relied on by nonlinear classifiers might bear little relationship to
420 what is actually represented by the brain. In other words, nonlinear classifiers are too
421 informationally greedy, and so cannot serve as surrogates for the decoding procedures of the
422 brain. Hence, a version of the Dictum appealing to nonlinear classifiers would clearly be false:
423 nonlinear decoding does not provide evidence for what neural activity patterns represent. In
424 contrast, the standard version of the Dictum seems to assume that linear classifiers are relatively
425 conservative in terms of the information they can exploit (that is, they are biologically plausible),

426 and so provide a safe (if defeasible) basis for making claims about representational content. The
427 fact that a linear classifier can discriminate between activity patterns from different conditions is
428 taken to provide good evidence that information about the conditions is both latent in the brain
429 and functionally available.

430
431 Critically, our earlier discussion of the uncertainty surrounding the source of (linearly) decodable
432 information shows the flaw in this reasoning. The fact that linear classifiers are biologically
433 plausible does not preclude them from also being informationally greedy. Linear classifiers are
434 surprisingly good at finding some linear combination of input features which discriminates
435 between conditions in a multivariate data set. As we saw in our discussion of orientation
436 decoding in V1, even when we do know the underlying functional architecture, how a classifier
437 exploits information in neural data is deeply opaque. To further illustrate the greed of linear
438 classifiers, consider that in psychology some have noted that linear decision-making models can
439 be surprisingly good even when feature weightings are assigned more or less arbitrarily (Dawes
440 [1979]). To emphasise a similar point, when using MVPA there is not even a guarantee that
441 classifiers are detecting multivariate signals. In a simulation study, Davis *et al.* ([2014])
442 produced a univariate fMRI signal that could not be detected by activation-based analyses, but
443 could nonetheless be decoded reliably.

444
445 Although a classifier (linear or nonlinear) may, through training, come to discriminate
446 successfully between activity patterns associated with different experimental conditions, the
447 information the classifier uses as the basis for this discrimination is not constrained to be the
448 information the brain actually exploits to make the distinction (that is, they are informationally
449 greedy). Importantly, it is evidence about the latter and not the former that is critical for zeroing
450 in on the contents of neural representations. Hence, decodability does not entail that the features
451 being combined, or their method of combination, bears any connection to how the brain is
452 decoding its own signals. At best, MVPA-based decoding shows that information about
453 experimental conditions is latent in neural patterns, but it cannot show that it is used, or even
454 usable, by the brain. This is the deep reason why the Dictum is false.

455

456

4. Objections And Replies

457 We have argued that the Decoder's Dictum is false. In this section we consider and respond to
458 some objections to our criticism.

459

460 **4.1 Does anyone really believe the Dictum?**

461 When criticizing inferences in cognitive neuroscience, it is common for the philosopher to be
462 informed that no working scientist really makes the sort of inference. Such an assertion is often
463 meant to be a normative claim as much as a descriptive one ('no good scientist argues thus'). Yet
464 it is the descriptive claim which really matters—for philosophical critique matters only insofar as
465 it identifies areas of actual methodological friction.

466

467 Do scientists really believe something like the Dictum? Our reconstruction of the theoretical
468 basis of the Dictum already suggests that they do. At the same time, enumeration is also
469 illuminating. Here are just a few (of many possible) illustrative examples where the Dictum is
470 either overtly referenced or strongly implied:

471

472 (1) Kamitani and Tong ([2005]) was one of the first studies showing that orientation
473 information is decodable from voxels in early visual cortex, including V1. They
474 state that their MVPA approach 'may be extended to studying the neural basis of
475 many types of mental content' (p. 684).

476 (2) Hung *et al.* ([2005]) was one of the first studies to pair MVPA with cellular
477 recordings. They showed that object identity and category could be decoded from
478 monkey IT as soon as ~125 ms post-stimulus onset. They state that their approach
479 'can be used to characterize the information represented in a cortical area [...]' (p.
480 865).

481 (3) In an early review of studies like Kamitani and Tong ([2005]) and Hung *et al.*
482 ([2005]), Haynes and Rees ([2006], p. 524) conclude that 'individual introspective
483 mental events can be tracked from brain activity at individual locations when the
484 underlying neural representations are well separated', where separation is
485 established by decodability with linear classifiers.

486 (4) Woolgar *et al.* ([2011]) used decoding to investigate the multiple-demand or 'MD'
487 regions of the brain, a frontoparietal network of regions that seem to be recruited

488 across cognitive tasks. They used decoding to investigate these regions because ‘[i]n
489 conventional fMRI the representational content of MD regions has been more
490 difficult to determine, but the question can be examined through multi-voxel pattern
491 analysis (MVPA)’ (p. 744).

492 (5) An important technique with time-series decoding is that of discriminant cross-
493 training, or ‘temporal generalization’: a classifier is trained on data from one time-
494 bin, and tested on another. In a review of this method, King and Dehaene ([2014], p.
495 1) claim it ‘provides a novel way to understand how mental representations are
496 manipulated and transformed’.

497 (6) More complex MVPA methods, which characterize the structure of an activation
498 space, or its ‘representational geometry’, have been promoted as ‘a useful
499 intermediate level of description, capturing both the information represented in
500 neuronal population code and the format in which it is represented’ (Kriegeskorte
501 and Kievet [2013], p. 401).

502

503 Some brief observations are worth making about these examples. First, they include both
504 individual studies (1, 2, 4) and reviews (3, 5, 6), spanning most of the period that decoding
505 methods have been utilized in neuroimaging, and were written by key figures responsible for
506 developing these methods. Second, the examples span fMRI (1, 4), EEG and MEG (4), and
507 cellular recordings (2, 3). The Dictum thus appears to be a fundamental and widespread
508 assumption in cognitive neuroscience, which has arguably played a key role in popularizing
509 MVPA because of what it promises to deliver.⁵

510

511 **4.2 Good decoding is not enough**

512 Another tempting reply to our argument goes as follows. Classifier performance is graded, so it
513 makes sense to talk about different brain regions having more or less decodable information. For
514 example, although early visual cortex contains some information about object category,
515 decodability is typically much worse than it is in inferior temporal cortex (IT), a region heavily
516 implicated in the representation of object categories (Kiani *et al.* [2007]; Kriegeskorte *et al.*

⁵ Of course, not all researchers using MVPA subscribe to the Dictum. As we have acknowledged, some embrace decoding because of its benefits over more conventional analyses, without drawing unjustified inferences about representational content.

517 [2008]). So perhaps the Dictum is true if we restrict ourselves to the best or most decodable
518 regions.

519
520 The problem with this reply is that it faces the same objection elaborated in detail above. What
521 makes a given region the best or most decodable might have little or nothing to do with the
522 information that is available to and used by the brain. This is why decoding results can be (and
523 often are) at odds with the answers derived from other methods. As pointed out earlier, visual
524 motion is more decodable from V1 than V5/+MT using fMRI (Kamitani and Tong [2006];
525 Seymour *et al.* [2009]), even though it is well-established that V5/+MT is a functionally
526 specialized region for representing and processing motion information. Seymour *et al.* ([2009])
527 similarly report classification accuracy of 86 % in V1 and 65 % in V5/+MT, though they
528 themselves refrain from drawing any strong conclusions due to the ‘potential differences
529 underlying functional architecture in each region’ (Seymour *et al.* [2009], p. 178).

530
531 Their caution appears to embody the same concern that decoding results may reflect arbitrary
532 differences to which the classifier is sensitive, without guaranteeing that these results track real
533 differences in neural representation. Decoding—excellent or otherwise—is not a reliable guide to
534 representation.

535
536 Another problem with this suggestion is that it entails that poor decodability (or even failure to
537 decode) provides evidence that the information is not represented in a region. But this is false.
538 Non-significant decoding does not entail the absence of information. One might have simply
539 chosen the wrong classifier or stimuli, or the particular code used by the brain might not be read
540 out easily by a linear classifier. Dubois *et al.* ([2015]) provide a nice illustration of this issue.
541 They compared single-unit recordings with fMRI decoding in the face patch system of the
542 macaque brain—an area known to possess face-sensitive neurons. In agreement with the single-
543 unit data, face viewpoint was readily decodable from these regions. However, in the anterior face
544 patches, face identity could not be decoded, even though single unit data shows that it is strongly
545 represented in the region. These results indicate how poor decodability provides a thin basis
546 upon which to mount negative claims about what a given region does not represent.

547

548 In sum, one cannot appeal to any level of classifier performance—good or bad—to preserve the
549 Dictum.

550

551 **4.3 Predicting behaviour is not enough**

552 Though not always carried out, the ability to connect classifier performance to behaviour has
553 been highlighted as one of the strengths of decoding methods (Naselaris *et al.* [2011]). To be
554 sure, a deep problem with the Dictum is that decodability fails to show that information is
555 formatted in a way that is used, or usable, by the brain (Cox and Savoy [2003]), while connecting
556 decoding to behaviour helps make the case for functional utilization (Tong and Pratte [2012]). If
557 behavioural performance can be predicted from the structure present in brain activation patterns,
558 this would provide more compelling evidence that decodable information is used (or at the very
559 least usable) by the brain, and hence neurally represented.

560

561 The simplest way to connect decoding and behaviour is to show that classifier and human
562 performance are highly correlated. Minimally, if this obtains for some activation patterns more
563 than others, this provides some (relatively weak) evidence that the patterns which correlate with
564 behaviour represents information that is used in the guidance of behaviour.

565

566 Williams *et al.* ([2007]) provided one of the earliest indications that not all decodable
567 information is ‘read-out’ in behaviour. They analysed the spatial pattern of the fMRI response in
568 specific task-relevant brain regions while subjects performed a visual shape discrimination task.
569 They hypothesized that if decodable shape category information is behaviourally relevant, then
570 decodability should be higher on correct trials than on incorrect trials. Critically, they showed
571 that although both retinotopic cortex and lateral occipital cortex (LOC) in humans contains
572 decodable category information, only the LOC shows a difference in pattern strength for correct
573 as compared to incorrect trials. Specifically, category information was decodable on correct but
574 not incorrect trials in the LOC. This was not true for retinotopic cortex. This pattern of results
575 suggests that only the information in LOC might drive behaviour.

576

577 It is also possible to quantify the relationship between decodability and behaviour more
578 precisely. For example, in an early EEG decoding study, Philiastides and Sajda ([2006]) were

579 able to show there was no significant difference between human psychometric and classifier
580 ‘neurometric’ functions, suggesting that the classifier performance was highly predictive of
581 observer performance when trained on time-series data of certain latencies.

582

583 While connection to behaviour supplies valuable evidence, we still think that it is not enough to
584 warrant inferences to representational content. As we noted earlier, there are cases where
585 decodability appears to show something about functional processing rather than the content of
586 neural representations. Again, V1 provides a useful test case. Since we know that V1 primarily
587 encodes information about low-level visual features (such as luminance or orientation) and does
588 not encode higher-level visual features (such as shape or object category) any decoding of
589 higher-level visual features is unlikely to reflect genuine representational content. This is true
590 even if decoded information can be linked with behavioural performance. For example, Haynes
591 and Rees ([2005]) found that V1 activity was predictive of whether or not subjects were
592 perceiving a visual illusion, and Kok *et al.* ([2012]) found that top-down effects of expectation
593 on V1 were predictive of behavioural performance. In these cases, the connection is that early
594 processing modulates later processing that determines behaviour.

595

596 Note that the problem is not one of spurious correlation. In an important sense, it is quite the
597 opposite problem. There is plenty of information, even in V1, which a clever decoding algorithm
598 can often pick up on. More generally, a brain region might carry information which is reliably
599 correlated with the information that is actually used, but which is not itself used in behaviour.
600 This is because the information in a region might need to be transformed into a more appropriate
601 format before it is read out. As DiCarlo and Cox ([2007], p. 335) put it, ‘[...] the problem is
602 typically not a lack of information or noisy information, but that the information is badly
603 formatted[...]’. But even ‘badly formatted’ information might correlate with behaviour. In
604 summary, merely predicting behaviour using decodable information is not enough to revive the
605 Dictum.

606

607

5. Moving Beyond The Dictum

608 We have argued that the Decoder’s Dictum is false. However, we are not pessimists about
609 decoding. Rather, we think the right conclusion to draw is that decoding must be augmented in

610 order to provide good evidence about neural representation. If linear classifiers are greedy, then
611 they cannot function as a surrogate for the sort of linear read-out carried out by the brain.
612 Instead, we need some additional assurance that a particular decoding result relies on information
613 stemming from neural representations. This need not be knock-down evidence, but decodability
614 alone is not enough to do the job (as the Dictum suggests).

615
616 In the previous section, we considered one form of augmentation—linking decoding results to
617 behavioural outcomes—and argued that it was insufficient. The problem was that linkages to
618 behaviour do not show that the information is actually formatted in a useable way. Framing it
619 this way, however, already suggests a solution. The Dictum relies on the idea that the biological
620 plausibility of linear classifiers allows them to function as a kind of surrogate—the classifier-as-
621 decoder takes the place of the brain-as-decoder in showing that information that is latent in
622 neural activity is used, or usable (cf. de Wit *et al.* [2016]). We have shown that it cannot play this
623 role. But if the information latent in patterns of neural activity can be used to predict observer
624 behaviour based on a psychological model, then we would have a more sound evidential basis
625 for drawing conclusions about neural representation. For unlike classifier performance, observer
626 behaviour is clearly dependent on how the brain decodes its own signals. In other words, this
627 approach depends on offering a psychologically plausible model of how observers (through
628 down-stream processing) exploit the information found in patterns of neural activity (cf. Ritchie
629 and Carlson [2016]). And as it happens, such an approach is already on offer.

630
631 There is a long tradition in psychology of modeling behavioural performance using
632 psychological spaces (Attneave [1950]; Shepard [1964]). Here by ‘psychological’ space we mean
633 a space in which dimensions reflect different features or combinations of features of stimuli, as
634 reconstructed from comparative similarity judgments of observers of stimuli/conditions. Models
635 within this tradition characterize representations for individual stimuli or experimental conditions
636 as points in a space, and observer behaviour (such as choice or reaction time) is modeled based
637 on the relationship between different representations in these spaces. Thus, familiar
638 categorization models from cognitive psychology such as prototype models, exemplar models,
639 and decision boundary models all predict observer behaviour based on different distance metrics
640 applied to a reconstructed psychological space (Ashby and Maddox [1993]). A virtue of some

641 MVPA methods like Representational Similarity Analysis (RSA) is that they help to focus
642 attention on structure in activation spaces (Haxby *et al.* [2014]; Kriegeskorte and Kievet [2013]).
643 In RSA the pair-wise (dis)similarity for patterns of activity for different conditions is computed,
644 which can be used to reconstruct an activation space from multivariate neural data. A hypothesis
645 that many have considered is that if an activation space implements a psychological space, then
646 one can apply psychological models or hypotheses to the activation space directly in order to
647 predict behaviour (Edelman *et al.* [1998]; de Beeck *et al.* [2001], [2008]; Davis and Poldrack
648 [2014]). Note that this approach is importantly different from the Dictum, as it does not rely on
649 using linear classifiers as a surrogate. Furthermore, the approach achieves both biological and
650 psychological plausibility through a linkage between the structure of the decoded activation
651 space and the structure of behaviour (Ritchie and Carlson [2016]). And since it makes use of
652 MVPA in conjunction with established techniques for modeling behaviour, it also takes
653 advantage of some of the strengths of MVPA we have already mentioned. Here we offer two
654 examples of research that adopt this sort of approach.

655
656 First, a popular and theoretically simple approach involves directly comparing the similarity
657 structure of activation spaces with psychological spaces reconstructed from subjects' similarity
658 judgments of stimuli (e.g. Mur *et al.* [2013]; Bracci and de Beeck [2016]; Wardle *et al.* [2016]).
659 One illustration of this approach is provided by the results of Sha *et al.* ([2015]), who collected
660 similarity ratings for a large number of exemplar images for several animate or inanimate object
661 categories. The similarity space constructed from these judgments was then directly related to the
662 similarity structure of activation spaces from throughout the brain measured using fMRI. They
663 found that activation spaces that correlated with the behavioural similarity space were best
664 accounted for by a single dimension, which seemed to reflect an animacy continuum rather than
665 a categorical difference between the neural patterns for animate and inanimate objects (Kiani *et*
666 *al.* [2007]; Kriegeskorte *et al.* [2008]).

667
668 Second, some work has focused on the psychological plausibility of activation spaces by using
669 them to predict the latency of behaviour. For example, in two studies using fMRI and MEG
670 decoding, Carlson and Ritchie (Carlson *et al.* [2014]; Ritchie, Tovar, and Carlson [2015])
671 showed that distance from a decision boundary for a classifier through activation space was

672 predictive of reaction time (RT). In their experiments they were explicitly motivated by the idea
673 that linear classifiers are structurally identical to the model of an observer under signal detection
674 theory (Green and Swets [1966]). A natural extension of signal detection theory is that distance
675 from an evidential boundary negatively correlates with RT (Ashby and Maddox [1994]). As
676 predicted, they found that RT negatively correlated with distance from the decision boundaries,
677 suggesting a level of psychological plausibility to even simple linear classifiers.

678
679 Crucially, in these sorts of studies it is implausible to suppose that the information is present but
680 not correctly formatted, because the decoded format of the information in activation space is
681 precisely what is being used to predict behaviour in a psychologically plausible manner. We do
682 not mean to suggest that the results we have summarized suffice for drawing conclusions about
683 neural representation, but we do believe that they help point the way forward.

684

685 **6. Conclusion**

686 The Decoder's Dictum is false. Significant decoding, even when supplemented by other results,
687 does not warrant an inference that the decoded information is represented. However, we do
688 believe that if behaviour can be connected to the structure of activation space in a
689 psychologically plausible manner, then this may warrant the sort of inference researchers have
690 had in mind. And we should stress that we do not think the above shows that decoding is
691 immaterial. Indeed, as we have suggested, MVPA is crucial for connecting activation spaces to
692 behaviour. Rather, as we have argued, appealing to the Dictum obscures not only the true import
693 of decoding as a tool in cognitive neuroscience, but also what sort of evidence is required for
694 making claims about neural representation.

695

696 **Acknowledgements**

697 Thanks to two anonymous reviewers for helpful comments on a previous draft, and to audiences
698 at Macquarie University and the 2014 Australasian Society for Cognitive Science for feedback
699 on earlier versions of this work. Funding for this research was provided by the Australian
700 Research Council (FT140100422 to Colin Klein).

701

702 J. Brendan Ritchie

703 *Laboratory of Biological Psychology*
704 *KU Leuven*
705 *Tiensestraat 102 - Box 3714*
706 *3000 Leuven , Belgium*
707 *brendan.ritchie@kuleuven.be*

708
709 *David Michael Kaplan*
710 *Department of Cognitive Science*
711 *Perception in Action Research Centre*
712 *ARC Centre of Excellence in Cognition and its Disorders*
713 *Macquarie University NSW 2109*
714 *Sydney, Australia*
715 david.kaplan@mq.edu.au

716
717 *Colin Klein*
718 *Department of Philosophy*
719 *ARC Centre of Excellence in Cognition and its Disorders*
720 *Macquarie University NSW 2109*
721 *Sydney, Australia*
722 colin.klein@mq.edu.au

723
724 **References**

- 725
726 Ashby, F. G. and Maddox, W. T. [1994]: ‘A response time theory
727 of separability and integrality in speeded classification’, *Journal of Mathematical*
728 *Psychology*, **38**, pp. 423–66.
729 Attneave, F. [1950]: ‘Dimensions of similarity’, *The American Journal of Psychology*, **63**,
730 pp. 516–56.
731 Bechtel, W. [1998]: ‘Representations and cognitive explanations: Assessing the
732 dynamicist’s challenge in cognitive science’, *Cognitive Science*, **22**, pp. 295–318.
733 Benson, N. C., Butt, O. H., Datta, R., Radoeva, P. D., Brainard, D. H. and Aguirre, G. K.
734 [2012]: ‘The retinotopic organization of striate cortex is well predicted by surface
735 topology’, *Current Biology*, **22**, pp. 2081–5.
736 Bracci, S. and Beeck, H. O. [2016]: ‘Dissociations and associations between shape and
737 category representations in the two visual pathways’, *The Journal of Neuroscience*, **36**,

- 738 pp. 432–44.
- 739 Carlson, T. A., Ritchie J. B., Kriegeskorte, N., Durvasula, S. and Ma, J. [2014]: ‘Reaction
740 time for object categorization is predicted by representational distance’, *Journal of*
741 *Cognitive Neuroscience*, **26**, pp. 132–42.
- 742 Cox, D. D. and Savoy, R. L. [2003]: ‘Functional magnetic resonance imaging fMRI brain
743 reading: detecting and classifying distributed patterns of fMRI activity in human
744 visual cortex’, *Neuroimage*, **19**, pp. 261–70.
- 745 Cox, D. D. [2014]: ‘Do we understand high-level vision?’, *Current Opinion in Neurobiology*,
746 **25**, pp. 187–93.
- 747 Davis, T., LaRocque, K. F., Mumford J. A., Norman, K. A., Wagner, A. D. and Poldrack, R.
748 A. [2014]: ‘What do differences between multi-voxel and univariate analysis mean?
749 How subject-, voxel-, and trial-level variance impact fMRI analysis’, *NeuroImage*,
750 **97**, pp. 271–83.
- 751 Davis, T. and Poldrack, R. A. [2013]: ‘Measuring neural representations with fMRI: practices
752 and pitfalls’, *Annals of the New York Academy of Sciences*, **1296**, pp. 108–34.
- 753 ———. [2014]: ‘Quantifying the internal structure of categories using a neural typicality
754 Measure’, *Cerebral Cortex*, **24**, pp. 1720–37.
- 755 Dawes, R. M. [1979]: ‘The robust beauty of improper linear models in decision making’,
756 *American psychologist*, **34**, 571–82.
- 757 de Beeck, H. O., Wagemans, J. and Vogels, R. [2001]: ‘Inferotemporal neurons represent low-
758 dimensional configurations of parameterized shapes’, *Nature Neuroscience*, **4**, pp.
759 1244–52.
- 760 deCharms, R. C. and Zador, A. [2000]: ‘Neural Representation and the Cortical Code’,
761 *Annual Review of Neuroscience*, **23**, pp. 613–47.
- 762 De Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. [2016] ‘Is neuroimaging measuring
763 information in the brain?’, *Psychonomic Bulletin & Review*, pp. 1–14.
- 764 DiCarlo, J. J. and Cox, D. D. [2007]: ‘Untangling invariant object recognition’, *Trends in*
765 *Cognitive Sciences*, **11**, pp. 333–41.
- 766 Dretske, F. I. [1983]: ‘Precis of Knowledge and the Flow of Information’, *Behavioral and Brain*
767 *Sciences*, **6**, pp. 55–63.
- 768 ———. [1988]: *Explaining behavior: Reasons in a world of causes*, Cambridge: MIT Press.
- 769 Dubois, J., de Berker, A. O. and Tsao, D. Y. [2015]: ‘Single-unit recordings in the macaque
770 face patch system reveal limitations of fMRI MVPA’, *The Journal of Neuroscience*,
771 **35**, pp. 2791–802.
- 772 Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. [1998]: ‘Toward direct
773 visualization of the internal shape representation space by fMRI’, *Psychobiology* **26**,
774 pp. 309–21.
- 775 Eliasmith, C. and Anderson, C. H. [2004] *Neural engineering: Computation, representation,*
776 *and dynamics in neurobiological systems*, Cambridge: MIT Press.
- 777 Fodor, J. A. [1984]: ‘Semantics, Wisconsin style’, *Synthese*, **59**, pp. 231–50.
- 778 Freeman, J., Brouwer, G.J., Heeger, D. J. and Merriam, E.P. [2011]: ‘Orientation decoding
779 depends on maps, not columns’, *The Journal of Neuroscience*, **31**, pp. 4792–804.
- 780 Green, D. M., and Swets, J. A. [1966]: *Signal detection theory and psychophysics*. New York:
781 Wiley & Sons.
- 782 Harrison, S. A., and Tong, F. [2009]: ‘Decoding reveals the contents of visual working
783 memory in early visual areas’, *Nature*, **458**, pp. 632–5.

- 784 Hatsopoulos, N. G. and Donoghue, J. P. [2009]: ‘The science of neural interface systems’,
785 *Annual Review of Neuroscience*, **32**, pp. 249–66.
- 786 Haxby, J. V. [2012] ‘Multivariate pattern analysis of fMRI: the early beginnings’,
787 *Neuroimage*, **62**, pp. 852–5.
- 788 Haxby, J. V., Connolly A. C. and Guntupalli J. S. [2014]: ‘Decoding neural representational
789 spaces using multivariate pattern analysis’, *Annual Review of Neuroscience*, **37**, pp.
790 435–56.
- 791 Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten J. L. and Pietrini, P. [2001]:
792 ‘Distributed and overlapping representations of faces and objects in ventral temporal
793 cortex’, *Science*, **293**, pp. 2425–30.
- 794 Haynes, J-D. and Rees, G. [2005]: ‘Predicting the orientation of invisible stimuli from activity
795 in human primary visual cortex’, *Nature Neuroscience*, **8**, pp. 686–91.
- 796 ———. [2006]: ‘Decoding mental states from brain activity in humans’, *Nature Reviews*
797 *Neuroscience*, **7**, pp. 523–34.
- 798 Hung, C. P., Kreiman, G., Poggio, T. and DiCarlo, J. J. [2005]: ‘Fast readout of object identity
799 from macaque inferior temporal cortex’, *Science*, **310**, pp. 863–6.
- 800 Jehee, J. FM., Brady, D. K. and Tong, F. [2011]: ‘Attention improves encoding of task-
801 relevant features in the human visual cortex’, *The Journal of Neuroscience*, **31**, pp.
802 8210–9.
- 803 Kamitani, Y. and Tong, F. [2005]: ‘Decoding the visual and subjective contents of the human
804 Brain’, *Nature Neuroscience*, **8**, pp. 679–85.
- 805 ———. [2006]: ‘Decoding seen and attended motion directions from activity in the human
806 visual cortex’, *Current Biology*, **16**, pp. 1096–102.
- 807 Kanwisher, N., McDermott, J. and Chun, M. M. [1997]: ‘The fusiform face area: a module in
808 human extrastriate cortex specialized for face perception’, *The Journal of*
809 *Neuroscience*, **17**, pp. 4302–11.
- 810 Kiani, R., Esteky, H., Mirpour, K. and Tanaka, K. [2007]: ‘Object category structure in
811 response patterns of neuronal population in monkey inferior temporal cortex’, *Journal*
812 *of Neurophysiology*, **97**, pp. 4296–309.
- 813 King, J. R. and Dehaene, S. [2014]. ‘Characterizing the dynamics of mental representations:
814 the temporal generalization method’, *Trends in Cognitive Sciences*, **18**, pp. 203–10.
- 815 Kirsh, D. [1990] ‘When is Information Explicitly Represented?’, In Philip P. Hanson(ed.),
816 *Information, Language and Cognition*, Vancouver: University of British Columbia
817 Press.
- 818 Klein, C. [2010]: ‘Philosophical issues in neuroimaging’, *Philosophy Compass*, **5**, pp. 186–98.
- 819 Kok, P., Jehee, J. FM. and de Lange, F. P. [2012]: ‘Less is more: expectation sharpens
820 representations in the primary visual cortex’, *Neuron*, **75**, pp. 265–70.
- 821 Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K. and
822 Bandettini, P. A. [2008]: ‘Matching categorical object representations in inferior
823 temporal cortex of man and monkey’, *Neuron*, **60**, pp. 1126–41.
- 824 Kriegeskorte, N., Mur, M. and Bandettini, P. [2008]: ‘Representational similarity analysis—
825 connecting the branches of systems neuroscience’, *Frontiers in Systems Neuroscience*,
826 **2**, pp. 4.
- 827 Kriegeskorte, N. and Bandettini, P. [2007]: ‘Analyzing for information, not activation, to
828 exploit high-resolution fMRI’, *Neuroimage*, **38**, pp. 649–62.
- 829 Kriegeskorte, N. and Kievit, R. A. [2013]: ‘Representational geometry: integrating cognition,

- 830 computation, and the brain’, *Trends in cognitive sciences*, **17**, pp. 401–412.
- 831 Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D-S., Fenno, L. E.,
832 Ramakrishnan, C. and Deisseroth, K. [2010]: ‘Global and local fMRI signals driven
833 by neurons defined optogenetically by type and wiring’, *Nature*, 465, pp. 788–92.
- 834 Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. and Oeltermann, A. [2001]:
835 ‘Neurophysiological investigation of the basis of the fMRI signal’, *Nature*, **412**, pp.
836 150–157.
- 837 Logothetis, N. K. [2008]: ‘What we can do and what we cannot do with fMRI’, *Nature*, **453**,
838 pp. 869–78.
- 839 Marr, D. [1982]: *Vision: A computational investigation into the human representation and*
840 *processing of visual information*, New York: WH Freeman.
- 841 Mannion, D. J., McDonald, J. S., and Clifford, C. W. G. [2009]: ‘Discrimination of the local
842 orientation structure of spiral Glass patterns early in human visual cortex’,
843 *Neuroimage*, **46**, pp. 511–5.
- 844 Miski, M., Kim, Y., Bandettini, P. A. and Kriegeskorte, N. [2010]: ‘Comparison of
845 multivariate classifiers and response normalizations for pattern-information fMRI’,
846 *Neuroimage*, **53**, pp. 103–18.
- 847 Mur, M., Bandettini, P. A. and Kriegeskorte, N. [2009] ‘Revealing representational content
848 with pattern-information fMRI—an introductory guide’, *Social, Cognitive, and*
849 *Affective Neuroscience*, pp. 1–9.
- 850 Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A. and Kriegeskorte, N. [2013]:
851 ‘Human object-similarity judgments reflect and transcend the primate-IT object
852 representation’, *Frontiers in Psychology*, **4**, pp. 128.
- 853 Naselaris, T., Kay, K. N., Nishimoto, S. and Gallant, J. L. [2011]: ‘Encoding and decoding in
854 fMRI’, *Neuroimage*, **56**, pp. 400–10.
- 855 Nir, Y., Dinstein, I., Malach, R. and Heeger, D. J. [2008]: ‘BOLD and spiking activity’,
856 *Nature Neuroscience*, **11**, pp. 523–4.
- 857 Norman, K. A., Polyn, S. M., Detre, G. J. and Haxby, J. V. [2006]: ‘Beyond mind-reading:
858 multi-voxel pattern analysis of fMRI data’, *Trends in cognitive sciences*, **10**, pp. 424–
859 430.
- 860 Ogilvie, R. and Carruthers, P. [2016]: ‘Opening Up Vision: The Case Against Encapsulation’,
861 *Review of Philosophy and Psychology*, pp. 1–22.
- 862 Panzeri, S., Macke, J. H., Gross, J. and Kayser, C. [2015]: ‘Neural population coding:
863 combining insights from microscopic and mass signals’, *Trends in Cognitive Sciences*,
864 **19**, pp. 162–72.
- 865 Philiastides, M. G. and Sajda, P. [2006]: ‘Temporal characterization of the neural correlates of
866 perceptual decision making in the human brain’, *Cerebral Cortex*, **16**, pp. 509–518.
- 867 Poldrack, R. A. [2006]. ‘Can cognitive processes be inferred from neuroimaging data?’,
868 *Trends in Cognitive Sciences*, **10**, pp. 59–63.
- 869 Pouget, A., Dayan, P. and Zemel, R. [2000]: ‘Information processing with population codes’,
870 *Nature Reviews Neuroscience*, **1**, pp. 125–132.
- 871 Priebe, N. J. and Ferster, D. [2012]: ‘Mechanisms of Neuronal Computation in Mammalian
872 Visual Cortex’, *Neuron*, **75**, pp. 194–208.
- 873 Ritchie, J. B., Tovar, D. A. and Carlson, T. A. [2015]: ‘Emerging Object Representations in
874 the Visual System Predict Reaction Times for Categorization’, *PLoS Computational*
875 *Biology* **11**(6), e1004316.

- 876 Ritchie, J. B. and Carlson, T. A. [2016]: ‘Neural decoding and “inner” psychophysics: A
877 distance-to-bound approach for linking mind, brain, and behavior’, *Frontiers in*
878 *Neuroscience*, **10** (190).
- 879 Scarantino, A. and Piccinini, G. [2010]: ‘Information Without Truth’, *Metaphilosophy*, **41**,
880 pp. 313–30.
- 881 Seymour, K., Clifford, C. W. G., Logothetis, N. K. and Bartels, A. [2009]: ‘The coding of
882 color, motion, and their conjunction in the human visual cortex’, *Current Biology*, **19**,
883 pp. 177–83.
- 884 Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O. and
885 Connolly, A. C. [2015]: ‘The animacy continuum in the human ventral vision
886 pathway’, *Journal of cognitive neuroscience*, **27**, pp. 665–78.
- 887 Shepard, R. N. [1964]: ‘Attention and the metric structure of the stimulus space’, *Journal of*
888 *mathematical psychology*, **1**, pp. 54–87.
- 889 Singh, K. D. [2012]: ‘Which neural activity do you mean? fMRI, MEG, oscillations and
890 neurotransmitters’, *Neuroimage*, **62**, pp. 1121–30.
- 891 Sirotin, Y. B. and Das, A. [2009]: ‘Anticipatory haemodynamic signals in sensory cortex not
892 predicted by local neuronal activity’, *Nature*, **457**, pp. 475–79.
- 893 Tong, F. and Pratte, M. S. [2012]: ‘Decoding patterns of human brain activity’, *Annual Review*
894 *of Psychology*, **63**, pp. 483–509.
- 895 Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S-M. and Carlson, T. A.
896 [2016]: ‘Perceptual similarity of visual patterns predicts dynamic neural activation
897 patterns measured with MEG’, *NeuroImage*, **132**, pp. 59-70.
- 898 Williams, M. A., Baker, C. I., de Beek, H. P. O, Shim, W. M., Dang, S., Triantafyllou, C. and
899 Kanwisher, N. [2008]: ‘Feedback of visual object information to foveal retinotopic
900 cortex’, *Nature Neuroscience*, **11**, pp. 1439–45.
- 901 Williams, M. A., Dang, S. and Kanwisher, N. G. [2007]: ‘Only some spatial patterns of fMRI
902 response are read out in task performance’, *Nature Neuroscience*, **10**, pp. 685–6.
- 903 Woolgar, A., Thompson, R., Bor, D. and Duncan, J. [2011]: ‘Multi-voxel coding of stimuli,
904 rules, and responses in human frontoparietal cortex’, *Neuroimage*, **56**, pp. 744–752.
905
- 906
907
908

