frontiers in Neuroscience

PERSPECTIVE
published: 28 April 2016
doi: 10.3389/fnins.2016.00190

Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior

J. Brendan Ritchie[1,2]* *and Thomas A. Carlson*[3,4]

[1] *Laboratory of Biological Psychology, Brain and Cognition Unit, KU Leuven, Leuven, Belgium,* [2] *Department of Philosophy, University of Maryland, College Park, MD, USA,* [3] *Perception in Action Research Centre, Department of Cognitive Science, Macquarie University, Sydney, NSW, Australia,* [4] *ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, NSW, Australia*

A fundamental challenge for cognitive neuroscience is characterizing how the primitives of psychological theory are neurally implemented. Attempts to meet this challenge are a manifestation of what Fechner called "inner" psychophysics: the theory of the precise mapping between mental quantities and the brain. In his own time, inner psychophysics remained an unrealized ambition for Fechner. We suggest that, today, multivariate pattern analysis (MVPA), or neural "decoding," methods provide a promising starting point for developing an inner psychophysics. A cornerstone of these methods are simple linear classifiers applied to neural activity in high-dimensional activation spaces. We describe an approach to inner psychophysics based on the shared architecture of linear classifiers and observers under decision boundary models such as signal detection theory. Under this approach, distance from a decision boundary through activation space, as estimated by linear classifiers, can be used to predict reaction time in accordance with signal detection theory, and distance-to-bound models of reaction time. Our "neural distance-to-bound" approach is potentially quite general, and simple to implement. Furthermore, our recent work on visual object recognition suggests it is empirically viable. We believe the approach constitutes an important step along the path to an inner psychophysics that links mind, brain, and behavior.

Keywords: inner psychophysics, MVPA, signal detection theory, visual object categorization, reaction times

OPEN ACCESS

Edited by:
Jeanette Mumford,
University of Texas at Austin, USA

Reviewed by:
Jonas Kaplan,
University of Southern California, USA
Rogier Kievit,
University of Amsterdam, Netherlands

***Correspondence:**
J. Brendan Ritchie
j.brendan.w.ritchie@gmail.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *21 January 2016*
Accepted: *18 April 2016*
Published: *28 April 2016*

Citation:
Ritchie JB and Carlson TA (2016) Neural Decoding and "Inner" Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. Front. Neurosci. 10:190. doi: 10.3389/fnins.2016.00190

1. MAPPING A ROUTE FROM MIND TO BRAIN: THE DREAM OF AN INNER PSYCHOPHYSICS

A fundamental challenge for cognitive neuroscience is to explain how the primitives of psychological theory are neurally implemented (Davis and Poldrack, 2013). Theories and models aimed at meeting this challenge are the modern manifestation of what Fechner (1860/1966) called "inner" psychophysics: the theory of the precise mapping between mental quantities and the brain. In Fechner's own time, inner psychophysics remained a dream (Scheerer, 1992). Even today, concrete proposals remain elusive. Indeed, many have wondered whether cognitive neuroscience is even up to the challenge (Price and Friston, 2005; Coltheart, 2006; Feldman Barrett, 2009; Poldrack, 2010). We side with those who have argued, more optimistically, that the field requires a shift in thinking for progress to continue (de Wit et al., 2016).

The key to inner psychophysics, we believe, is using psychological models, applied to neural activity, to predict behavior (Werner and Mountcastle, 1963; Britten et al., 1992). Consider the traditional motivation for using behavioral measures in ("outer") psychophysics. Since the mind cannot be measured directly, Fechner and others reasoned that behavior can serve as a proxy to estimate stimulus-driven variation in mental quantities and processes. Similar reasoning supports behavioral measures as the key to developing "linking" hypotheses from psychological theory to the brain (Brindley, 1960; Teller, 1984). If a neural component implements a primitive identified by some theory or model, then we should be able to predict behavioral variation from its functional organization (Forstmann et al., 2011). We focus on *representational* linking hypotheses: how are the representations posited by psychological theory implemented by the brain in a manner that predicts behavior?

We propose that multi-variate pattern analysis (MVPA), or neural "decoding," methods provide one starting point for the development of an inner psychophysics, and representational linking hypotheses. These methods have allowed researchers to investigate the information latent in neural activity patterns, and uncover the structure and content of the brain's population code (Kriegeskorte and Kievit, 2013; Haxby et al., 2014; Haynes, 2015). A cornerstone of these methods are linear classifiers applied to high-dimensional neural *activation spaces*. Here we present a simple approach for developing representational linking hypotheses based on the shared architecture of linear classifiers and observers under decision boundary models such as signal detection theory (Green and Swets, 1966). We also review our work on visual object categorization that lends empirical support to the approach (Carlson et al., 2014; Ritchie et al., 2015), and connect the approach to research on the neural loci of decision-making (Gold and Shadlen, 2007).

2. WHAT CAN DECODING CONTRIBUTE TO INNER PSYCHOPHYSICS? BIOLOGICAL VS. PSYCHOLOGICAL PLAUSIBILITY

The suitability of MVPA methods for investigating neural representation, and developing representational linking hypotheses, can be motivated in part by their biological and (potential) psychological plausibility.

A common assumption in cognitive neuroscience is that the brain utilizes "population codes": internal representations are implemented in distributed patterns of neural activity—incidentally, an idea somewhat anticipated by Fechner's (1882/1987) discussion of memory. If the brain uses population codes it may face a multivariate classification problem when differentiating these neural patterns. If this differentiation is achieved by a linear combinations of inputs, then we should be able to decode the contents of the encoding patterns of activity using classifiers that mirror the linear operations the brain employs. In decoding analyses, activation spaces are reconstructed from patterns of neural activity, and a linear classifier is trained to discriminate between the patterns for different experimental conditions. If the classifier performs significantly above chance, then minimally it can be inferred that information about the conditions is latent, and accessible, from the patterns of neural activity (Kriegeskorte and Bandettini, 2007). The biological plausibility of the linear classifiers also suggests that the information may be explicitly represented by the patterns.

While MVPA offers one starting point for developing an inner psychophysics, the biological plausibility of linear classifiers does not alone establish a connection between activation spaces and observer psychology. As de Wit et al. (2016) emphasize, that a classifier can learn to discriminate patterns of neural activity shows that information is latent, and perhaps represented, but not necessarily how it is being used, or is usable, by the observer (Cox and Savoy, 2003; Williams et al., 2007). In other words, the biological plausibility of linear classifiers is not enough to show that they are *psychologically* plausible, which also requires linking a psychological theory to an activation space. Fortunately, there is a long tradition in psychology of modeling the structure of psychological spaces to predict behavior (Attneave, 1950; Shepard, 1964; for a more recent perspective, see: Gärdenfors, 2000). All quantitative models of categorization within this tradition hold that tokens of a representation occupy different points in a space, and how these points are positioned in the space, based on some similarity or distance function, drives mental processes and behavior (Ashby and Maddox, 1993). For example, in prototype models (e.g., Posner and Keele, 1968) discriminability of a stimulus is determined by the distance of a representation to the central tendency of a category distribution in the space, and in exemplar models (e.g., Nosofsky, 1986) discriminability is determined by the relative similarity of the representation to all other exemplar representations in the space. The high-dimensional activation spaces reconstructed using MVPA may conform to similar principles of organization identified in these quantitative models of psychological space (Op de Beeck et al., 2008; Davis and Poldrack, 2013; Kriegeskorte and Kievit, 2013; Haxby et al., 2014). In which case, representational linking hypotheses can be developed by applying principles from models of psychological space to activation spaces.

One straightforward approach is to directly compare the structure of a psychological space to an activation space. Several studies using fMRI (Edelman et al., 1998; Mur et al., 2013; Charest et al., 2014; Sha et al., 2015; Bracci and Op de Beeck, 2016), cellular recordings (Op de Beeck et al., 2001) and MEG (Wardle et al., 2016), have constructed psychological spaces for stimuli from judgments of visual similarity, and compared them to activation spaces constructed using methods such as representational similarity analysis (RSA), which estimates the pair-wise (dis)similarity between patterns of neural activity for different conditions (Kriegeskorte et al., 2008a). A robust correlation between the two spaces suggests the activation spaces might implement the representations that are driving the similarity judgments. This similarity-based approach reflects the psychological plausibility of methods like RSA. Although seldom noticed in cognitive neuroscience,

linear classifiers are also psychologically plausible, as we will illustrate.

3. A STEP IN THE RIGHT DIRECTION: A PSYCHOLOGICALLY PLAUSIBLE NEURAL DISTANCE-TO-BOUND APPROACH

As used in MVPA, linear classifiers specify a decision boundary through an activation space in order to discriminate between neural patterns produced by different experimental conditions. In general form this decision process is equivalent to that of the human observer as posited by (linear) decision boundary models of categorization for multivariate stimuli (Ashby and Gott, 1988; Ashby and Maddox, 1990). To illustrate the close correspondence of decoding methods with these models, consider that Naïve Bayes classifiers, applied after linear discriminate analysis (LDA), and observers under signal detection theory (SDT) share a common organization. More specifically, they make the same assumptions concerning: (i) distributions of evidence/data, and (ii) the evaluation of evidence/data by the classifier/observer.

At an abstract level, SDT specifies a number of primitives that mediate the relationship between stimulus and behavior (**Figure 1A**). Consider a simple task in which an observer must discriminate and map two stimuli (*stim1, stim2*) to two responses (*resp1, resp2*). The input produced by a stimulus is characterized as a sample from one of two distributions (f_{stim1} , f_{stim2}) along an evidence dimension, e.g., brightness, with response choice resulting solely from a rule applied to a decision variable (assuming equal stimulus probabilities and outcome utilities). Traditionally, the decision variable was the value of the log-likelihood ratio of the evidence, given the available hypotheses (i.e., the logarithm of the ratio of the height of f_{stim1} and f_{stim2} at a point on the evidence dimension). Assuming no response bias, the decision rule states that the observer selects the response with greater value in the ratio, resulting in a decision boundary along the evidence dimension. Under the usual distributional assumptions of normalcy and equal variance, the measure of observer sensitivity generated by the model, *d'*, is the difference (or distance) between the means of f_{stim1} and f_{stim2} (it is also closely related to Fechner's own measure of sensitivity; Link, 1994). Architecturally, this model requires an internal stimulus representation along with a decision process that determines choice behavior given the information made explicit by the representation.

LDA is a technique for transforming a space to maximize between class variance (Duda et al., 2001). In the simplest case, a 2D-space is replaced with a single discriminant axis onto which each data point is projected. Importantly, it is assumed that the distributions for the classes along each dimension (or "feature") are normal and of equal variance. If we further assume the dimensions are independent, then we have a Naïve Bayes classifier, which guesses based on the summed ratios of the posterior probabilities for each class along each dimension. If we take the logarithm of the ratio, and assume equal prior probabilities, then the classifier uses a decision rule applied to the log-likelihood ratio (Pereira et al., 2009). When a 2D space is projected to a single discriminant axis, the architectures of

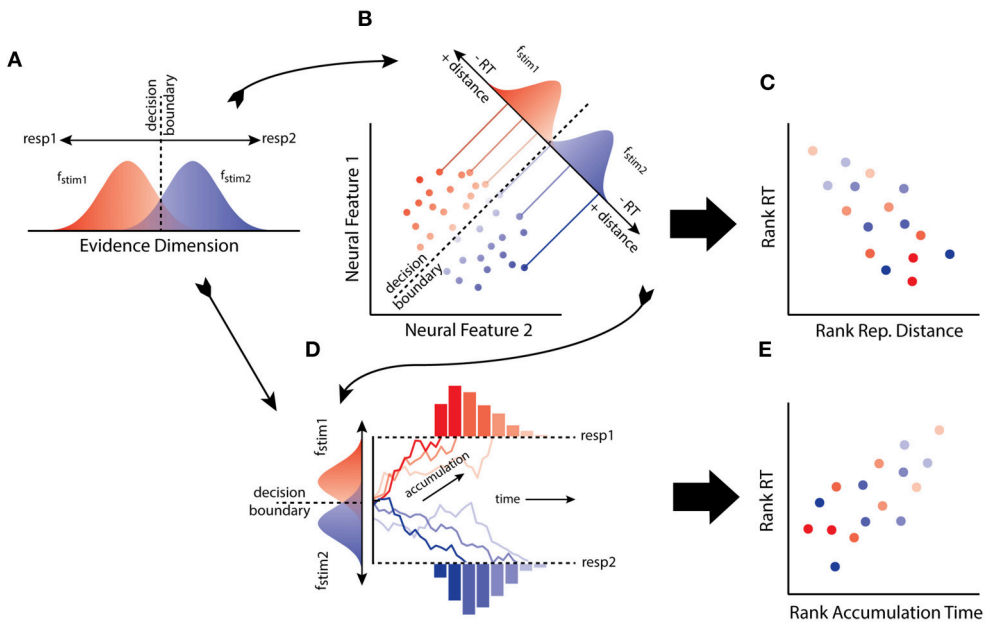

FIGURE 1 | (A) The model of the observer from signal detection theory (SDT) maps to **(B)** the distributions of patterns of neural activity as reconstructed using neural decoding. **(C)** From this connection it is predicted that RT will negatively correlate with distance from a decision boundary through activation space. **(D)** The model of the observer from SDT, and the distributed patterns of neural activity, can also be linked to the evidence accumulation process. **(E)** Thus another prediction from the hypothesis is that distances from a decision boundary in activation space reflects the mean accumulation rate, and hence RT should correlate with accumulation rates determined by representational distance.

the classifier and the SDT observer are identical. In the multi-dimensional case, the classifier is akin to the decision boundary observer under the multi-dimensional generalization of SDT, when dimensions are independent (Ashby and Townsend, 1986).

An initial implication of this equivalence is that one may be able, in principle, to achieve close correspondence between classifier and observer performance. For example, consider the results of Philiastides and Sajda (2006) who observed similar psychometric and "neurometric" functions for human and classifier performance, when using an LDA classifier applied to EEG data. Not only do their results take on a new theoretical significance in light of the above equivalence, but methodologically their application of a sensitivity measure to classifier performance seems even more appropriate since the measure presupposes the very architecture that the classifier possesses (Tanner, 1956).

A further implication of the equivalence relates to reaction time (RT) and the speed of transforming representations of a stimulus into a decision. A simple feature of perceptual decision-making, as first characterized by SDT, is that the quality of evidence for an observer varies in its uncertainty (Tanner and Swets, 1954). In particular, evidence close to the observers decision boundary, or criterion, is more ambiguous, reflecting the greater likelihood of the evidence under the alternative hypotheses. In contrast, evidence far from the boundary is less ambiguous, reflecting greater likelihood of the evidence under one of the hypotheses about the source of the stimulus. Thus, relative to some decision boundary, evidence quality tends to vary with distance. RT also tends to vary with the quality of evidence: lower quality evidence results in longer decision times compared to high-quality evidence. A simple consequence of this familiar picture from decision boundary models (e.g., SDT), as developed with distance-to-bound models of choice and RT, is that distance from a decision boundary will negatively correlate with RT (Pike, 1973; Ashby and Maddox, 1994).

LDA classifiers learn to discriminate between activity patterns by positioning a decision boundary along a discriminant axis. If an activation space provides the evidence being utilized by an observer (**Figure 1B**), then one possibility is that distance from a classifier boundary will predict RT (**Figure 1C**). Such a result would suggest that an activation space implements an explicit representation of stimulus information that is used by the observer in a psychologically plausible manner. When decision boundary models like SDT were first developed, it was presumed that the evidence utilized by an observer was some unknown state of neural activity (Swets et al., 1961; Werner and Mountcastle, 1963). The neural distance-to-bound approach we have described provides a means of making good on this presumption. The approach is potentially quite general and is simple to implement as it relies on familiar MVPA and behavioral methods. We have also conducted experiments to test the approach.

3.1. Neural Distance-to-Bound Predicts Reaction Time for Object Categorization

Two of our recent experiments on visual object categorization provide tangible evidence in support of neural distance-to-bound as a viable approach to inner psychophysics (Carlson et al., 2014; Ritchie et al., 2015). In both experiments, subjects were tasked with judging as quickly and accurately as possible whether object exemplars were animate or inanimate (i.e., "capable of self-movement"). RT for the task was then related to activation spaces reconstructed using fMRI and MEG decoding. Our prediction was that RT would negatively correlate with representational distances from a decision boundary computed using LDA (**Figure 1C**).

Inferior temporal cortex (IT) has been strongly implicated in object categorization in humans and primates (Logothetis and Sheinberg, 1996), and information about object categories—in particular, animacy—is highly decodable from this region using fMRI (Kriegeskorte et al., 2008b; Connolly et al., 2012; Konkle and Caramazza, 2013). In our first experiment we asked whether the animacy information latent in activity patterns in this region might be utilized when subjects performed an object categorization task (Carlson et al., 2014). Using the fMRI data of Kriegeskorte et al. (2008b), we computed the representational distances of activity patterns for 92 object exemplars (faces and bodies of humans and animals, as well as natural objects and human artifacts) from a decision boundary for animacy in IT. We then correlated these distances with the mean RT of separate subjects performing the animacy task. Despite using neural and behavioral data from completely different subjects, we observed a significant negative correlation between RT and the representational distances (**Figure 2A**). This result suggests that animacy information in activity patterns in the region may also be used by the observers performing the animacy task.

While typically utilized with fMRI, MVPA is also increasingly being employed with EEG/MEG (King and Dehaene, 2014). It has been shown that significant decoding for object categories, and in particular animacy, occurs as early as 60 ms post-stimulus onset, with peak classifier performance occurring at greater latencies for more abstract categories (Carlson et al., 2013; Cichy et al., 2014). In our second experiment we sought to determine when in time we might observe a negative correlation between RT and distance-to-bound (Ritchie et al., 2015). Peak decoding reflects the time at which information about stimulus categories is most discriminable in the brain, thus we predicted representational distance would negatively correlate with RTs during the period of peak decoding. We estimated the representational distances at each 20 ms time-point -100–600 ms post-stimulus onset for 24 object exemplars (same subordinate groupings as in our previous experiment). While in the MEG, subjects performed the animacy task, and their median RTs were correlated with the representational distances at each time point. This allowed us to see when in time there was a significant correlation between representational distance and RT. As predicted, we observed a significant correlation during the period of peak decoding, as well as at later time points (**Figure 2B**). More generally we found that the relationship between representational distance and RT followed the time-course of decoding.

Taken together, these two results provide compelling evidence in support of the viability of the neural distance-to-bound approach.

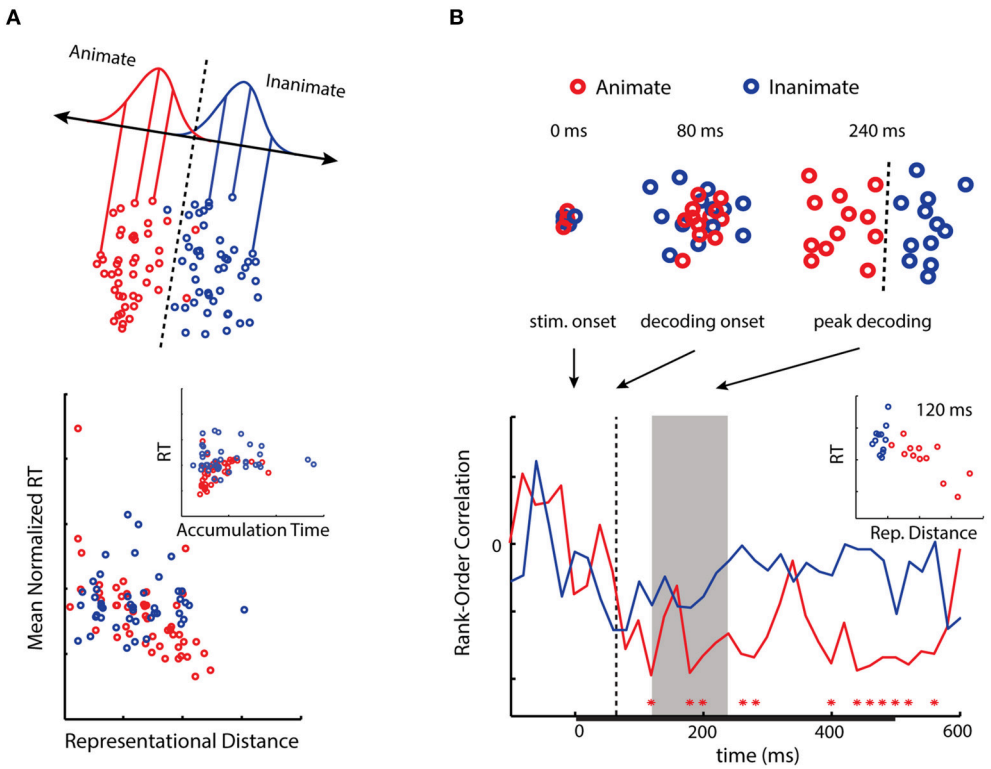

FIGURE 2 | (A) A 2D activation space for animacy in human IT. Distances from a decision boundary for animacy in human IT negatively correlate with RT as predicted from the neural distance-to-bound approach. When distance is transformed into a drift rate parameter, RT and evidence accumulation time positively correlate (modified from Carlson et al., 2014; © 2014 by the Massachusetts Institute of Technology). **(B)** Emergent activation spaces for animacy as reconstructed using MEG. Distances from decision boundaries at peak decoding (gray area) negatively correlate with RT (modified from Ritchie et al., 2015). Interestingly, for both results the correlations are primarily driven by the animate exemplar stimuli (for discussion, see Carlson et al., 2014).

3.2. Distance-to-Bound and the Neural Basis of Decision-Making

The neural distance-to-bound approach has two important implications for research on the neural basis of decision-making. First, it has been suggested, in part based on decoding methodology, that the line between representing and deciding in the brain is blurred (DiCarlo and Cox, 2007). Our approach provides theoretical and empirical support for this perspective. If an observer's decision boundary extends through an activation space, then at least in some circumstances stimulus representations and decision variables may be implemented in the same neural component (Carlson et al., 2014). This contrasts with perspectives according to which stimulus representations and decision variables are generally associated with distinct brain regions (Schall, 2001; Heekeren et al., 2004; Shadlen et al., 2007; Filimon et al., 2013).

Second, evidence accumulation models of choice and RT are a popular means of investigating the neural underpinnings of decision-making (Smith and Ratcliff, 2004; Gold and Shadlen, 2007; Shadlen and Kiani, 2013). While there are several versions of these models to choose from Ratcliff and Smith (2004) they all share certain features: evidence is sampled from a random variable; at each iteration of the model, the evidence is used to update a decision variable; and when the decision variable reaches a stopping value, or threshold, the observer makes a decision. Typically, choice and RT effects are modeled as resulting from differences in accumulation rate of the decision variable between experimental conditions.

Distance-to-bound and evidence accumulation models of RT have sometimes been contrasted with each other (Pike, 1973; Thomas, 2006). However, this opposition is not obligatory, since distance-to-bound can be related to accumulation rate (Ratcliff, 1985; Ashby, 2000). We assume the simplest RT-distance relationship possible: a monotonic decrease in RT correlating with a monotonic increase in distance. So neural distance-to-bound can also be thought of as an approach for characterizing the distance between the accumulate rate distributions (**Figures 1D–E**), for any evidence accumulation model that assumes a monotonic decrease in RTs as accumulation rate increases. Thus, neural distance-to-bound also provides a method for connecting neural decoding and neural evidence accumulation approaches.

To illustrate the connection, we used distances from the animacy boundary in human IT to simulate accumulation rates for the sequential probability ratio test (SPRT; Wald, 1945), which has been used to relate spike rates to evidence

accumulation (Gold and Shadlen, 2002). SPRT is a dynamic extension of classic SDT: the observer selects a response based on the log-likelihood ratio, but if the value of the ratio has not yet reached threshold, the observer receives another unit of evidence. The decision variable is the running tally of the ratio, which accumulates until the threshold is reached. In our study, we simulated SPRT using the representational distances in human IT for each individual exemplar, transforming the distances into accumulation rates (**Figure 2A**; Carlson et al., 2014).

4. THE PATH TO AN INNER PSYCHOPHYSICS: STILL A LONG WAY TO GO

We believe neural distance-to-bound has considerable potential as an (easy to apply) approach for developing representational linking hypotheses. Still, we stress that it is just one possible approach for furthering the study of inner psychophysics. In some domains of perception and cognition, it might not be viable at all. For example, it is unclear how well it will fair in a domain like neurolinguistics, where there is considerable difficulty in linking the primitives of linguistic theory to the brain (Poeppel, 2012). Furthermore, other approaches that may be superior at modeling RT, such as exemplar models (Nosofsky and Stanton, 2005), could provide a better connection between activation spaces and evidence accumulation.

More fundamentally, failure to observe a negative RT-distance correlation does not necessarily entail that the information in neural activity is unused. Instead, we might have the wrong model for *how* it is used. For instance, crucially our approach assumes linear separability, as using nonlinear classifiers for decoding is typically discouraged on the grounds that they are overpowered and lack biological plausibility (Kamitani and Tong, 2005; DiCarlo and Cox, 2007). However, many quantitative models of categorization do not share this assumption. For example, exemplar models have often been tested using stimulus sets that do not allow for linear separation in perceptual space (e.g., Nosofsky, 1986). Thus, the existence of psychologically plausible nonlinear categorization models may warrant revisiting the use of nonlinear classifiers in MVPA.

The ultimate import of our approach, then, is that it suggests more sophisticated representational linking hypotheses are possible. Recognizing this possibility is an important step along the path to an inner psychophysics, and the realization of Fechner's dream.

AUTHOR CONTRIBUTIONS

JR and TC contributed equally to developing the ideas presented in the paper. JR wrote, and TC helped edit, the manuscript.

FUNDING

This research was supported by an Australian Research Council Future Fellowship [FT120100816] to TC. The funders had no role in the preparation or decision to publish the manuscript.

REFERENCES

Ashby, F. G. (2000). A stochastic version of general recognition theory. *J. Math. Psychol.* 44, 310–329. doi: 10.1006/jmps.1998.1249

Ashby, F. G., and Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 14, 33. doi: 10.1037/0278-7393.14.1.33

Ashby, F. G., and Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 598. doi: 10.1037/0096-1523.16.3.598

Ashby, F. G., and Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *J. Math. Psychol.* 37, 372–400. doi: 10.1006/jmps.1993.1023

Ashby, F. G., and Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *J. Math. Psychol.* 38, 423–466. doi: 10.1006/jmps.1994.1032

Ashby, F. G., and Townsend, J. T. (1986). Varieties of perceptual independence. *Psychol. Rev.* 93, 154. doi: 10.1037/0033-295X.93.2.154

Attneave, F. (1950). Dimensions of similarity. *Am. J. Psychol.* 63, 516–556. doi: 10.2307/1418869

Bracci, S., and Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* 36, 432–444. doi: 10.1523/JNEUROSCI.2314-15.2016

Brindley, G. S. (1960). *Physiology of the Retina and the Visual Pathway*. London: Edward Arnold.

Britten, K. H., Shadlen, M. N., Newsome, W. T., and Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* 12, 4745–4765.

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13, 1. doi: 10.1167/13.10.1

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., and Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *J. Cogn. Neurosci.* 26, 132–142. doi: 10.1162/jocn_a_00476

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., and Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14565–14570. doi: 10.1073/pnas.1402594111

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex* 42, 323–331. doi: 10.1016/S0010-9452(08)70358-7

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., et al. (2012). The representation of biological classes in the human brain. *J. Neurosci.* 32, 2608–2618. doi: 10.1523/JNEUROSCI.5547-11.2012

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

Davis, T., and Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Ann. N. Y. Acad. Sci.* 1296, 108–134. doi: 10.1111/nyas.12156

de Wit, L., Alexander, D., Vebjorn, E., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychon. Bull. Rev.* doi: 10.3758/s13423-016-1002-0. [Epub ahead of print].

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification, 2nd Edn*. New York, NY: Wiley & Sons.

Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology* 26, 309–321.

Fechner, G. T. (1966). *Elements of Psychophysics (Vol.I)*. New York, NY: Holt, Rinehart & Winston.

Fechner, G. T. (1987). Some thoughts on the psychophysical representation of memories (1882). *Psychol. Res.* 49, 209–212. doi: 10.1007/BF00309028

Feldman Barrett, L. (2009). The future of psychology: connecting mind to brain. *Perspect. Psychol. Sci.* 4, 326–339. doi: 10.1111/j.1745-6924.2009.01134.x

Filimon, F., Philiastides, M. G., Nelson, J. D., Kloosterman, N. A., and Heekeren, H. R. (2013). How embodied is perceptual decision making? evidence for separate processing of perceptual and motor decisions. *J. Neurosci.* 33, 2121–2136. doi: 10.1523/JNEUROSCI.2334-12.2013

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., and Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends Cogn. Sci.* 15, 272–279. doi: 10.1016/j.tics.2011.04.002

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gold, J. I., and Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36, 299–308. doi: 10.1016/S0896-6273(02)00971-6

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Green, D., and Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley & Sons.

Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Ann. Rev. Neurosci.* 37, 435–456. doi: 10.1146/annurev-neuro-062012-170325

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* 87, 257–270. doi: 10.1016/j.neuron.2015.05.025

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature* 431, 859–862. doi: 10.1038/nature02966

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi: 10.1038/nn1444

King, J., and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* 18, 203–210. doi: 10.1016/j.tics.2014.01.002

Konkle, T., and Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *J. Neurosci.* 33, 10235–10242. doi: 10.1523/JNEUROSCI.0983-13.2013

Kriegeskorte, N., and Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage* 38, 649–662. doi: 10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17, 401–412. doi: 10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis–connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Link, S. W. (1994). Rediscovering the past: gustav fechner and signal detection theory. *Psychol. Sci.* 5, 335–340. doi: 10.1111/j.1467-9280.1994.tb00282.x

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Ann. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., and Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Front. Psychol.* 4:128. doi: 10.3389/fpsyg.2013.00128

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* 115, 39. doi: 10.1037/0096-3445.115.1.39

Nosofsky, R. M., and Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: contrasting exemplar-retrieval, decision-boundary, and prototype models. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 608. doi: 10.1037/0096-1523.31.3.608

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252. doi: 10.1038/nn767

Op de Beeck, H. P., Wagemans, J., and Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: a modeling study. *Vision Res.* 48, 598–610. doi: 10.1016/j.visres.2007.11.019

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Philiastides, M. G., and Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cereb. Cortex* 16, 509–518. doi: 10.1093/cercor/bhi130

Pike, R. (1973). Response latency models for signal detection. *Psychol. Rev.* 80, 53–68. doi: 10.1037/h0033871

Poeppel, D. (2012). The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cogn. Neuropsychol.* 29, 34–55. doi: 10.1080/02643294.2012.710600

Poldrack, R. A. (2010). Mapping mental function to brain structure: how can cognitive neuroimaging succeed? *Perspect. Psychol. Sci.* 5, 753–761. doi: 10.1177/1745691610388777

Posner, M. I., and Keele, S. W. (1968). On the genesis of abstract ideas. *J. Exp. Psychol.* 77, 353–363. doi: 10.1037/h0025953

Price, C. J., and Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275. doi: 10.1080/02643290442000095

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychol. Rev.* 92, 212. doi: 10.1037/0033-295X.92.2.212

Ratcliff, R., and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* 111, 333. doi: 10.1037/0033-295X.111.2.333

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Comput. Biol.* 11:e1004316. doi: 10.1371/journal.pcbi.1004316

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nat. Rev. Neurosci.* 2, 33–42. doi: 10.1038/35049054

Scheerer, E. (1992). "Fechner's inner psychophysics: its historical fate and present status," in *Cognition, Information Processing and Psychophysics. Basic Issue*, eds H.-G. Geissler, S. W. Link, and J. T. Townsend (Hillsdale, NJ: Lawrence Erlbaum), 3–22.

Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., et al. (2015). The animacy continuum in the human ventral vision pathway. *J. Cogn. Neurosci.* 27, 665–678. doi: 10.1162/jocn_a_00733

Shadlen, M. N., and Kiani, R. (2013). Decision making as a window on cognition. *Neuron* 80, 791–806. doi: 10.1016/j.neuron.2013.10.047

Shadlen, M. N., Kiani, R., Hanks, T. D., and Churchland, A. K. (2007). "Neurobiology of decision making: an intentional framework," in *Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions*, eds C. Engel and W. Singer (Cambridge, MA: MIT Press), 71–102.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *J. Math. Psychol.* 1, 54–87. doi: 10.1016/0022-2496(64)90017-3

Smith, P. L., and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends Neurosci.* 27, 161–168. doi: 10.1016/j.tins.2004.01.006

Swets, J., Tanner, W. P. Jr., and Birdsall, T. G. (1961). Decision processes in perception. *Psychol. Rev.* 68, 301. doi: 10.1037/h0040547

Tanner, W. P. (1956). Theory of recognition. *J. Acoust. Soc. Am.* 28, 882–888. doi: 10.1121/1.1908504

Tanner, W. P. Jr., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychol. Rev.* 61, 401. doi: 10.1037/h0058700

Teller, D. Y. (1984). Linking propositions. *Vision Res.* 24, 1233–1246. doi: 10.1016/0042-6989(84)90178-0

Thomas, R. D. (2006). Processing time predictions of current models of perception in the classic additive factors paradigm. *J. Math. Psychol.* 50, 441–455. doi: 10.1016/j.jmp.2006.05.006

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* 16, 117–186. doi: 10.1214/aoms/1177731118

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., and Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with meg. *NeuroImage* 132, 59–70. doi: 10.1016/j.neuroimage.2016.02.019

Werner, G., and Mountcastle, V. B. (1963). The variability of central neural activity in a sensory system, and its implications for the central reflection of sensory events. *J. Neurophysiol.* 26, 958–977.

Williams, M. A., Dang, S., and Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nat. Neurosci.* 10, 685–686. doi: 10.1038/nn1900

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.