prioritizing existing, low-tech strategies for maintaining and improving neural health, rather than novel, high-tech strategies that are often very expensive and have uncertain or unproven benefits. A great deal of evidence exists that healthy diet, adequate exercise and sleep, and high-quality education are associated with improved and healthy cognitive function, and are safe and carry virtually no risk of harm. Public health measures such as lead-paint abatement and requirements for toxin-free workplaces support neural health. Continuing to advance science in these areas can help the public improve its understanding of optimal lifestyles and environmental conditions.

Second, we urged prioritization of treatment of neurological diseases and injuries, versus the development of new drugs and devices solely to make people smarter. The burden of neurological disorders is high and is projected to increase considerably in future years with an aging population. Neurological disorders are estimated to affect as many as a billion people globally, including millions of people in the USA alone [8]. One of the primary goals of neuroscience is to prevent and treat these disorders. Directing research funding towards treatment, rather than enhanced cognition, helps to improve the lives of millions of individuals, attends to justice, and honors the primary goal of scientific inquiry.

However, although we recognized the need to prioritize both low-tech strategies to improve neural function and new techniques to treat disease, we did not ignore newer, high-tech enhancement techniques. Our third recommendation was that research should be conducted on the prevalence, benefits, and risks of new neural modifiers to augment or enhance neural function. Very limited evidence of this type exists for the off-label use of stimulant drugs such as Adderall or the use of brain-stimulation techniques including transcranial direct stimulation to improve cognition in healthy people. Before society can make accurate ethical assessments of these novel enhancement techniques, we must understand them.

If research does demonstrate that particular novel neural enhancers are safe and beneficial, then stakeholders must seek justice in their distribution. In our fourth recommendation, we urged that policymakers ensure equitable access to beneficial neural enhancers. In our society, access to existing services and opportunities, such as education and nutrition, is not equal across individuals or groups. However, societal tolerance of inequity in access to other crucial goods does not make inequity right, nor should it hamper society's efforts to reduce or eliminate inequity where we can. If safe and effective novel forms of cognitive enhancement become available, they will present an opportunity to insist on a distribution that is fair and just. While not eliminating all other less tractable forms of injustice in the distribution of neural health and well-being, it is possible to ensure that any new forms of safe and beneficial neural modification do not worsen those injustices.

Concluding Remarks

By broadening the discussion of cognitive enhancement to include all forms of neural modification, the Bioethics Commission has expanded the scope of the current debate. Neural modification – to maintain or improve brain health within typical or statistically normal ranges, treat neurological disorders, and expand or augment neural function – raises a set of ethical considerations. Our recommendations are intended to serve as a resource for scientists, physicians, and policymakers. We hope they will spark a broader discussion of these issues and serve as an impetus for scientists to consider how the research they conduct today could transform society, for better or worse, in the years to come.

[1]Presidential Commission for the Study of Bioethical Issues, 1425 New York Ave NW, Washington, DC 20005, USA

[2]Office of the Provost, University of Pennsylvania, 120D College Hall, Philadelphia, PA 19104, USA

*Correspondence: nicolle.strand@bioethics.gov (N.K. Strand).

http://dx.doi.org/10.1016/j.tics.2015.08.001

References

1. Presidential Commission for the Study of Bioethical Issues (2015) *Gray Matters: Topics at the Intersection of Neuroscience, Ethics, and Society*, Presidential Commission for the Study of Bioethical Issues
2. Allhoff, F. *et al.* (2011) Ethics of human enhancement: an executive summary. *Sci. Eng. Ethics* 17, 201–212
3. Schwarz, A. (2015) Workers seeking productivity in a pill are abusing ADHD drugs. *New York Times* 18 April
4. Repantis, D. *et al.* (2010) Modafinil and methylphenidate for neuroenhancement in healthy individuals: a systematic review. *Pharmacol. Res.* 62, 187–206
5. Cakic, V. (2009) Smart drugs for cognitive enhancement: ethical and pragmatic considerations in the era of cosmetic neurology. *J. Med. Ethics* 35, 611–615
6. Maslen, H. *et al.* (2014) The regulation of cognitive enhancement devices: Extending the medical model. *J. Law Biosci.* 1, 68–93
7. Farah, M.J. *et al.* (2014) Cognitive enhancement. *WIREs Cogn. Sci.* 5, 95–103
8. World Health Organization (2006) *Neurological Disorders: Public Health Challenges*, World Health Organization

Forum

Resolving Ambiguities of MVPA Using Explicit Models of Representation

Thomas Naselaris[1,*] and Kendrick N. Kay[2,#]

We advocate a shift in emphasis within cognitive neuroscience from multivariate pattern analysis (MVPA) to the design and testing of explicit models of neural representation. With such models, it becomes possible to identify the specific representations encoded in patterns of brain activity and to map them across the brain.

MVPA is a powerful analysis tool that is replacing activation (or subtraction-based) analysis as the go-to method for

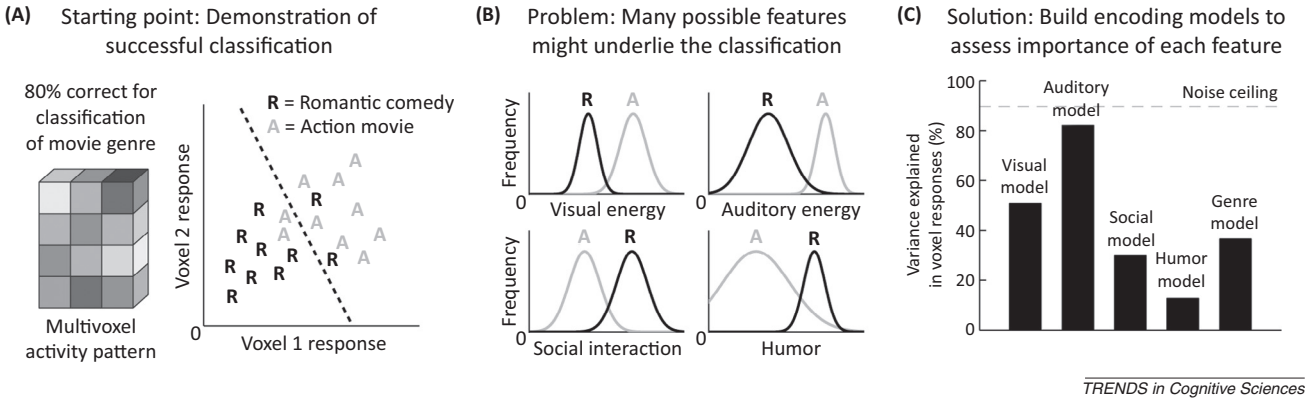

Figure 1. Resolving the Representational Ambiguity of Multivariate Pattern Analysis (MVPA). (A) An example MVPA experiment. Responses to several movie clips are measured. It is demonstrated that a linear classifier can predict the genre of a movie clip based on the multivoxel activity pattern elicited by that movie clip. (B) Representational ambiguity. Movie clips from different genres may differ with respect to one or more features. For example, clips from action movies (gray) may have larger amounts of visual energy than clips from romantic comedies (black). Thus, the voxels under consideration might represent a feature that is correlated with, but distinct from, the movie genre. (C) Building encoding models. To adjudicate between competing hypotheses about the features that are represented, each feature of interest is used to build an encoding model and the various models are fit to the data (Box 1, main text). By directly comparing the accuracy (variance explained) of different models, it is possible to determine the features encoded in the population activity.

interpreting functional magnetic resonance imaging (fMRI) data [1]. MVPA refers to the classification of patterns of brain activity into discrete experimental conditions (e.g., different stimuli, tasks, or cognitive states). The major appeal of MVPA is its sensitivity: it can identify populations of voxels that encode information about experimental conditions, even when the average amplitude of activity in the population does not vary across conditions.

Despite its appeal, MVPA has critical limitations as a tool for identifying the representations that are encoded in patterns of brain activity. There are three distinct kinds of ambiguity inherent to MVPA. The most benign kind is geometric ambiguity. This refers to the fact that activity patterns, interpreted as multivariate vectors, can be discriminated by MVPA on the basis of either their length (overall activation) or orientation. Although the pooling of length and orientation provides statistical sensitivity, these distinct features of the activity pattern cannot be disentangled when using MVPA alone. For example, the overall activity in a region may simply be larger in one condition compared with another, a fact that is missed in MVPA. Geometric ambiguity can be resolved by performing additional analyses (such as activation analysis).

More problematic is spatial ambiguity. MVPA provides little information about how representations are organized across the cortical surface (e.g., the retinotopic organization of visual cortex [2]). This ambiguity results from the fact that different cortical organizations can give rise to identical classification performance [3]. Techniques for resolving spatial ambiguity in MVPA, such as the use of searchlights or examination of classifier weights, can be misleading [4–6]. For example, significant nonzero classifier weights can be obtained for voxels whose average response is the same across the experimental conditions of interest [5].

The most serious limitation of MVPA is representational ambiguity. Even moderately complex stimuli or task paradigms contain many distinct sources of variation. Each source corresponds to different stimulus features or cognitive states that might be encoded in brain activity. MVPA does not provide a framework for testing and distinguishing between different sources of variation [3].

Here is an illustration of how representational ambiguity can arise. Suppose we hypothesize that a brain region of interest (ROI) specializes in representing the genre of movie that one is watching (Figure 1). To test the hypothesis, we conduct an experiment in which subjects are scanned while viewing segments of movies of two different genres, say action movies and romantic comedies. If it turns out that a classifier is able to accurately discriminate the movie segments of each genre on the basis of measured brain activity, we will have established that something about the movie segments is indeed encoded in the activity patterns of our ROI. However, we will not have determined what that something is. A variety of alternative features correlated with movie genre might be encoded in the activity patterns, including visual and auditory energy (e.g., action movies contain more energy than romantic comedies do), amount of social interaction, amount of humor, amount of spoken language, and so on.

In some cases, it may be possible to discriminate features by using an experimental design that varies one and only one feature at a time. Although careful experimental design will always have an important role in studying brain representations, in the case of MVPA studies, experimental design can be surprisingly difficult. For example, consider a highly controlled experiment in which sinusoidal gratings

of different orientations are presented: successful orientation decoding may not necessarily derive from an encoding of the orientation of the stimulus, but may instead derive from an encoding of edge artifacts in the stimulus [3].

One might attempt to use MVPA to compare the movie-genre hypothesis against alternative hypotheses by comparing classification performance obtained for different features. For example, we might divide movie segments into low/high stimulus energy, low/high social interaction, low/high humor content, and so on, and then train a separate classifier to discriminate each of these. According to this logic, if movie genre is discriminated with higher performance than other features, then the movie-genre hypothesis is affirmed. However, this approach is problematic because decoding performance does not directly indicate the amount of variance in activity that is attributable to a given feature. A feature may be perfectly decoded from population activity even though it is responsible for little variance in activity. For example, a purely visual representation might support highly accurate decoding of genre, even though genre *per se* explains little variance in the brain responses. Therefore, comparing classification performance across different kinds of feature is an 'apples-to-oranges' comparison that is likely to mislead.

Many researchers have adopted an alternative approach for identifying representations encoded in brain activity [2,3,7–11]. We refer to this approach as voxelwise modeling (VM). The hallmark of VM is an explicit model of representation, known as an encoding model. Formally, an encoding model proposes a set of sensory or cognitive features and specifies how these features are transformed into a prediction of brain activity for the experiment under consideration. A given set of features represents an explicit hypothesis about the representation encoded in the brain. This hypothesis is tested by evaluating how much variance in measured activity the encoding model explains (Box 1). Competing hypotheses can be adjudicated by comparing the amount of variance explained by different encoding models. Alternatively, hypotheses can be assessed by comparing how well a representational similarity matrix (e.g., a matrix with correlations between pairs of experimental conditions) constructed from a set of features matches the representational similarity matrix constructed from the measured activity. This approach, called 'representational similarity analysis', imposes fewer constraints on the mapping between features and brain activity [12]. In both cases, hypotheses are tested by evaluating explicit models of representation.

VM offers important advantages over MVPA. There is no notion of geometric or spatial ambiguity. Analyses are performed on individual voxels, so the length and orientation of an activity pattern are naturally separated and individual parameters can be mapped to the cortical surface at the native resolution of the data.

Importantly, VM provides the means to resolve representational ambiguity. Encoding models predict activity based on explicitly defined representations. This makes it possible to enumerate different potential sources of variation, test the explanatory power of each source of variation, and identify specific data points that are well or poorly predicted by a given model [7]. Finally, VM provides a quantitative benchmark of our understanding of neural representation. In an experiment where responses are measured to a range of stimulus or task conditions, a model that perfectly explains the observed variance in voxel activity in an ROI (or, alternatively, a similarity matrix constructed from the observed activity patterns [12]) could be offered as a complete theory of the ROI.

In conclusion, by improving detection sensitivity, MVPA is a powerful tool that has served the fMRI community well. In situations where prediction of stimulus or task states is of primary importance, MVPA will continue to have a useful role. However, MVPA provides fundamentally ambiguous results regarding the nature of brain representations. As research in cognitive neuroscience moves forward, we suggest that MVPA should be replaced by explicit models of representation.

Box 1. Steps in Building Encoding Models

- Design the experiment: typically, a large number of conditions are used to sample a variety of features, postponing commitment to the specific features that may be relevant to a given brain area.
- Collect the data: physiological responses are measured using multiple repetitions of each condition so that response variability (i.e., noise level) can be quantified.
- Select a model: the features hypothesized to be encoded in a given brain area are formally specified.
- Fit the model: free parameters of the model (e.g., weights in a linear model) are adjusted to best fit the data. This can entail ordinary least-squares estimation or regularized estimation procedures, such as ridge regression or the lasso.
- Summarize model parameters: parameters are summarized and compared across brain areas using simple metrics (e.g., mean or median) or more sophisticated methods (e.g., principal components analysis or model-based decoding). Reliability of parameter estimates is also assessed (e.g., by bootstrapping trials or subjects).
- Quantify model accuracy: to control for overfitting, model accuracy is assessed by cross-validating on new data (e.g., new trials, experimental conditions, or subjects). Accuracy is quantified as percent variance explained.
- Consider alternative models: the modeling procedure is repeated to determine whether the data might be better explained by a simpler or completely different model.

Acknowledgments

This work was supported by the McDonnell Center for Systems Neuroscience and Arts & Sciences at Washington University (K.N.K.) and by grant NEI R01 EY023384 (T.N.).

[1]Medical University of South Carolina, Charleston, SC, USA
[2]Washington University in St Louis, St Louis, MO, USA

*Correspondence: tnaselar@musc.edu (T. Naselaris).

http://dx.doi.org/10.1016/j.tics.2015.07.005

References

1. Haxby, J.V. *et al.* (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37, 435–456
2. Dumoulin, S.O. and Wandell, B. (2008) Population receptive field estimates in human visual cortex. *Neuroimage* 39, 647–660
3. Carlson, T.A. and Wardle, S.G. (2015) Sensible decoding. *Neuroimage* 110, 217–218
4. Etzel, J.A. *et al.* (2013) Searchlight analysis: promise, pitfalls, and potential. *Neuroimage* 78, 261–269
5. Haufe, S. *et al.* (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110
6. Davis, T. *et al.* (2014) What do differences between multivoxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage* 97, 271–283
7. Kay, K.N. *et al.* (2013) A two-stage cascade model of BOLD responses in human visual cortex. *PLoS Comput. Biol.* 9, e1003079
8. Naselaris, T. *et al.* (2015) A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage* 105, 215–228
9. Mitchell, T.M. *et al.* (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320, 1191–1195
10. Santoro, R. *et al.* (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* 10, e1003412
11. Brouwer, G.J. and Heeger, D.J. (2009) Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* 29, 13992–14003
12. Kriegeskorte, N. and Kievit, R.A. (2013) Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17, 401–412

Spotlight

Color Preferences Differ with Variations in Color Perception

Karen B. Schloss[1,*]

A recent study demonstrates that color preferences of red–green dichromats differ systematically from color preferences of typical trichromatic observers. These differences can be partially explained by variations in cone-opponent mechanisms of dichromatic and trichromatic observers, but they may also be explained from an ecological perspective.

It is well established that, on average, people with typical color vision prefer blues most and yellow–greens least, with moderate preference for hues in between [1–5]. However, if that were the complete story, then everyone would want to wear blue clothes, drive blue cars, and live in blue houses containing blue artifacts. Clearly that is not the case; people surround themselves with a wide variety of colors. What explains this disconnect between average color preferences and the ways people choose to color their world? The answer has at least two critical factors: individual differences and contextual effects. In this article, I highlight new discoveries on individual differences. (See [6] for a discussion of contextual effects for different kinds of objects.)

In their recent study, Álvaro, Moreira, Lillo, and Franklin [7] were the first to report color preferences of individuals with atypical color vision. They compared color preferences of typical males (trichromats) with those of two types of red–green dichromatic males: protanopes (missing long-wavelength sensitive photoreceptors; L-cones) and deuteranopes (missing medium-wavelength sensitive photoreceptors; M-cones). Previous simulations of dichromatic color perception suggest that protanopes and deuteranopes experience the spectrum within a range that trichromats would consider blues, grays, and yellows [8]. Although both have deficits in their red–green system, their percepts are not identical (e.g., protanopes perceive reds as darker-yellows and deuteranopes perceive reds as relatively lighter-yellows). Álvaro *et al.* found that, unlike trichromats who preferred blues most, dichromats maximally preferred saturated-yellow [7]. Surprisingly, deuteranope preferences were more similar to those of typical trichromatic males than to protanopes. This was partly because deuteranopes and trichromats strongly disliked dark-yellow relative to most other colors, whereas protanopes liked dark-yellow as much as reds, cyans, and even some blues.

Because color preferences of trichromats can be modeled by cone-contrast mechanisms in the visual system [9], Álvaro *et al.* predicted that dichromatic preferences could be modeled by modified predictors tailored to dichromats' altered cone-opponent mechanisms. The standard cone-contrast model explained significant variance (40%) in trichromat color preferences but it accounted for no significant variance in dichromat preferences for the full set of colors. However, the red–green system predicted deuteranope preferences for the subset of light colors, suggesting that the so-called 'red–green colorblind' individuals have some red–green discriminability. A modified cone-contrast predictor coding for perceived saturation (vividness) in protanopes strongly predicted their color preferences, which may be analogous to trichromats preferring more saturated colors [1,3,4]. This factor also predicted deuteranope preferences, but only for saturated colors. Accordingly, modified versions of the cone-contrast mechanisms are useful for characterizing some limited aspects of dichromatic color preferences.

Álvaro *et al.* emphasized physiological interpretations of their data, but their results can also be considered from an ecological perspective. The ecological valence theory (EVT) posits that color preferences are determined by preference for all objects or entities associated with the colors [1]. For example, trichromats like saturated-blue because it is mostly associated with positive objects (e.g., clear sky), and dislike dark-yellow because it is mostly associated with negative objects (e.g., biological wastes). Average color preferences are strongly predicted by an estimate of preference for all objects associated with each color (Weighted Affective Valence Estimate, or WAVE; 80% variance explained). WAVEs explained substantially more variance than a model based on the cone-contrasts (37%) [1]. Moreover, the EVT makes strong predictions about individual differences: Individuals should have different color preferences to the extent that they have different preferences for the same color-associated objects or associate different objects with the same