the computations may be conducted elsewhere and provided as input to the cerebellum. However, lesion results provide evidence that the cerebellum may be at least a necessary node for this computation [5]. When the nodulus and uvula are surgically removed, the VOR is no longer consistent with the state of the head, indicating that integration of rotation and acceleration signals to track head position does not occur. Therefore, the cerebellum appears to be a necessary structure to integrate the information from the otolith and canal afferents to provide state estimation, as reflected in the activity of the P cells.

Sensory illusions are a powerful method to test for the neural basis of forward models. For example, when people use a manipulandum to move a cursor on the screen, the geometric relationship between the motion of the hand and the motion of the cursor can be altered. After people learn the new relationship, they form an illusion regarding the motion of their own hand. Interestingly, people with cerebellar damage can also learn this relationship but do not form the illusion [6], providing further evidence that the cerebellum may be critical for the encoding of forward models.

In summary, Laurens *et al.* [1] provide new evidence that translation-selective neurons in lobules IX/X of the cerebellum estimate the state of the head using a computation that is consistent with a forward model.

References

1 Laurens, J. *et al.* (2013) Computation of linear acceleration through an internal model in the macaque cerebellum. *Nat. Neurosci.* 16, 1701–1708

2 Merfeld, D.M. *et al.* (1999) Humans use internal models to estimate gravity and linear acceleration. *Nature* 398, 615–618

3 Laurens, J. *et al.* (2010) Processing of angular motion and gravity information through an internal model. *J. Neurophysiol.* 104, 1370–1381

4 Wolpert, D.M. and Miall, R.C. (1996) Forward models for physiological motor control. *Neural Netw.* 9, 1265–1279

5 Angelaki, D.E. and Hess, B.J. (1995) Lesion of the nodulus and ventral uvula abolish steady-state off-vertical axis otolith response. *J. Neurophysiol.* 73, 1716–1720

6 Izawa, J. *et al.* (2012) Cerebellar contributions to reach adaptation and learning sensory consequences of action. *J. Neurosci.* 32, 4230–4239

An exemplar of model-based cognitive neuroscience

Thomas J. Palmeri

Department of Psychology, Vanderbilt University, Nashville, TN 37240, USA

Are categories learned by forming abstract prototypes or by remembering specific exemplars? Mack, Preston, and Love observed that patterns of functional MRI (fMRI) brain activity were more consistent with patterns of representations predicted by exemplar models than by prototype models. Their work represents the theoretical power of emerging approaches to model-based cognitive neuroscience.

A primary aim of cognitive science is to understand the mechanisms that give rise to faculties of mind like perception, learning, and decision making. One approach formalizes hypotheses about cognitive mechanisms in computational models. Cognitive models predict behavior, like the errors people make and the time it takes them to respond, and how behavior varies under different conditions, using different stimuli, with different amounts of learning. Another approach turns to the brain to identify neural mechanisms associated with different aspects of cognition, using techniques like neurophysiology, electrophysiology, and fMRI.

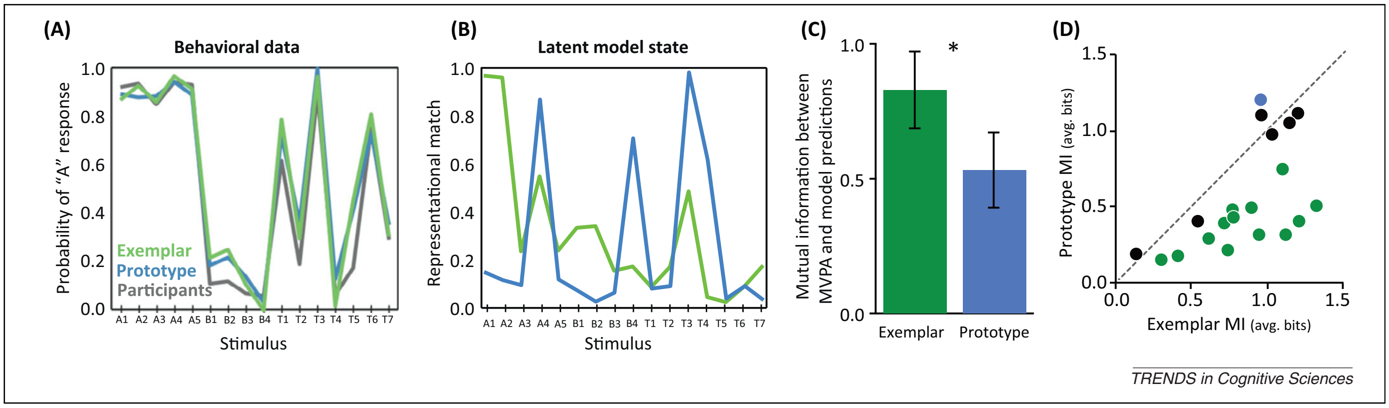
These two come together in a powerful new approach called model-based cognitive neuroscience [1]. Cognitive models decompose complex behavior into representations and processes and these latent model states are used to explain the modulation of brain states under different experimental conditions. Reciprocally, neural measures provide additional data that help constrain cognitive models and adjudicate between competing cognitive models that make similar predictions of behavior. For example, brain measures are related to cognitive model parameters fitted to individual participant data [2], measures of brain dynamics are related to measures of model dynamics [3,4], model parameters are constrained by neural measures [4], model parameters are used in statistical analyses of neural data [5], or neural data, behavioral data, and cognitive models are analyzed jointly within a hierarchical statistical framework [6].

Mack, Love, and Preston [7] adopted a model-based cognitive neuroscience approach to understand the mechanisms involved in category learning [8]. Consider everyday categories like dogs, cars, or chairs. Categories like these are abstractions in the sense that collections of visibly different objects are treated as the same kind of thing. But does that imply that the mental representations of categories are inherently abstract and that category learning involves creating abstractions? The earliest work on categorization assumed abstraction, either in the form of logical rules defining category membership or in the form of abstract prototypes capturing the family resemblance of category members. However, later work showed that cognitive models based on memory for experienced category exemplars could predict experimental results

Corresponding author: Palmeri, T.J. (thomas.j.palmeri@vanderbilt.edu).

1364-6613/$ – see front matter

http://dx.doi.org/10.1016/j.tics.2013.10.014



TRENDS in Cognitive Sciences

Figure 1. **(A)** Probability of a Category A response for each training stimulus (A1–A5, B1–B4) and transfer stimulus (T1–T7) observed from participants (gray), predicted by the exemplar model (green) and the prototype model (blue). **(B)** Latent model state (representational match) for each training stimulus (A1–A5, B1–B4) and transfer stimulus (T1–T7) predicted by the exemplar model (green) and the prototype model (blue). **(C)** Correspondence [mutual information (MI)] between patterns of brain activity revealed by multivoxel pattern analysis (MVPA) and representational match predicted by the exemplar model (green) and the prototype model (blue); higher MI means closer correspondence. **(D)** MI between MVPA and model predictions (representational match) for individual participants; correspondence for thirteen participants was significantly better for the exemplar than the prototype model (green), for six there was no significant difference (black), and for only one participant was it better for the prototype than the exemplar model (blue). Adapted, with permission, from [7].

that seemed to instead suggest abstraction. Although many argue that the evidence favors exemplar models, debate about exemplar models versus prototype models continues [8–10]. Could patterns of brain activity help adjudicate this theoretical controversy?

In [7], before scanning, participants learned to classify novel objects into one of two categories. Using a standard category-learning procedure [8] over several training blocks, participants viewed an object on each trial, categorized it as a member of Category A or B, and received corrective feedback. In the scanner, participants categorized training objects and new transfer objects as members of Category A or B without feedback (Figure 1A).

Mack and colleagues [7] used common mathematical formalizations of exemplar and prototype models, fitting them to the probability of categorizing objects as a member of each category for every participant (Figure 1A). The models make the same assumptions about how objects are represented, how similarities between objects and stored representations are computed, and how categorization decisions are made. Both models assume that categorization decisions are based on the relative similarity of an object to stored category representations. Naturally, they differ in the nature of those representations. For the exemplar model the evidence that an object is a member of Category A is based on the summed similarity of the object to stored exemplars of Category A divided by the summed similarity to stored exemplars of both categories, whereas for the prototype model the evidence is based on the similarity of the object to the prototype of Category A divided by the summed similarity to prototypes of both categories.

The summed similarity to the stored category representations – summed similarity to exemplars for the exemplar model versus summed similarity to prototypes for the prototype model – constitutes a latent model signature that Mack and colleagues called representational match. Although when fitted to behavioral data, the exemplar and prototype models make similar quantitative predictions about the probability that any given object is categorized as an A or a B, they differ considerably in the representational match for any given object that governs its predicted categorization (Figure 1B). Are the patterns of brain activity measured by fMRI while participants categorize each object more consistent with the representational match predicted by an exemplar model or a prototype model?

It is common to use multivoxel pattern analysis (MVPA) to identify patterns of brain activity that predict different kinds of stimuli, responses, or conditions. In [7], the goal was instead to use MVPA to identify patterns of brain activity that predict different values of representational match for different objects, where values of representational match came from fits of either the exemplar model or the prototype model to individual participant categorization data. A mutual information (MI) measure was used to quantify the relationship between brain states and latent model states, with higher MI reflecting greater consistency between patterns of voxel activity in the brain and patterns of representational match predicted by a model. The exemplar model was more consistent with brain measures than the prototype model, producing significantly greater MI measures (Figure 1C,D).

In [7], the exemplar and prototype models make nearly identical predictions about behavior. So comparing patterns of brain states with patterns of behavior, as might be traditionally done in cognitive neuroscience, would never uncover how the brain represents categories. Instead, by comparing how patterns of brain states compare with predicted latent model states we can begin to answer this fundamental question. Categories are learned by remembering exemplars not abstracting prototypes [2,8,9]. With its joint use of computational models of cognition with brain measures, this work well illustrates the growing sophistication and theoretical power of model-based cognitive neuroscience approaches [1–6].

Acknowledgments

This work was supported by NSF grant SMA 1041755 (Temporal Dynamics of Learning Center) and NEI grant R01 EY21833.

References

1 Forstmann, B.U. *et al.* (2011) Reciprocal relations between cognitive neuroscience and cognitive models: opposites attract? *Trends Cogn. Sci.* 6, 272–279

2 Nosofsky, R.M. *et al.* (2012) Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 333–338

3 Davis, T. *et al.* (2012) Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. *Cereb. Cortex* 22, 260–273

4 Purcell, B.A. *et al.* (2012) From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. *J. Neurosci.* 32, 3433–3446

5 White, C.N. *et al.* (2012) Perceptual criteria in the human brain. *J. Neurosci.* 32, 16716–16724

6 Turner, B.M. *et al.* (2013) A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage* 72, 193–206

7 Mack, M.L. *et al.* (2013) Decoding the brain's algorithm for categorization from its neural implementation. *Curr. Biol.* 23, 2023–2027

8 Richler, J.J. and Palmeri, T.J. (2013) Visual category learning. *Wiley Interdiscip. Rev. Cogn. Sci.* (in press)

9 Nosofsky, R.M. (2000) Exemplar representation without generalization? Comment on Smith and Minda's (2000) Thirty categorization results in search of a model. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1735–1743

10 Smith, D.J. and Minda, J.P. (2000) Thirty categorization results in search of a model. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 3