Title: Abstract neural representations of category membership beyond information coding stimulus or response

Authors: Robert M. Mok[1*] & Bradley C. Love[1,2*]

Affiliations:

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, United Kingdom.

[2]The Alan Turing Institute, United Kingdom

*Correspondence to: robert.mok@ucl.ac.uk, b.love@ucl.ac.uk

Abstract

How does the brain construct a mental representation appropriate for categorization? For decades, researchers have debated whether mental representations are symbolic or grounded in sensory inputs and motor programs. We evaluated these competing accounts with human participants using functional magnetic resonance imaging (fMRI). Participants completed a probabilistic concept learning task in which sensory, motor, and category variables were not perfectly coupled nor entirely independent. Our design made it possible to observe evidence for either account. Using a model to estimate participants' category rule and multivariate pattern analysis of fMRI data, we found left prefrontal cortex and MT coded category without information coding stimulus or response, despite category being based on the stimulus. Our results suggest that certain brain areas support categorization behaviour by constructing a concept representation in a representational format akin to a symbol that differs from stimulus-motor codes.

Introduction

Concepts organize our experiences into representations that can be applied across domains to support higher-order cognition. How does the brain organize sensory input into an appropriate representation for categorization? Are concepts simply a combination of sensory signals and motor plans, or does the brain construct a separate concept representation, abstracted away from sensory-motor codes? Despite much research on how people organize sensory information into a format suited for categorization (e.g. Kruschke, 1992; Love et al., 2004; Nosofsky, 1986) and its neural basis (e.g. Cromer et al., 2010; Davis et al., 2012b, 2012a; Folstein et al., 2013; Freedman & Assad, 2006; Mack et al., 2013, 2016; Seger & Miller, 2010; Sigala & Logothetis, 2002), few have explicitly examined whether category representations exist independently of sensory-motor information (Figure 1A).

Some concepts seem to be 'grounded' in sensory or motor experiences (Barsalou, 2008). For instance, the idea of 'pain' is based on experiences of pain, and the metaphorical use of the word is presumably linked to those bodily experiences. Certain aspects of concepts are more abstracted from first-hand experience and act more like symbols or pointers, which can support flexible cognition. For example, we know water can be used to clean the dishes, but when we are thirsty, we drink it. The same object can also appear entirely different in some contexts, such as a camouflaging stick insect appearing as a leaf, or when a caterpillar changes into a butterfly. In such cases where sensory information is unreliable or exhibits changes, an amodal symbol working as an abstract pointer may aid reasoning and understanding. Cognitive science and artificial intelligence researchers discuss the use of amodal symbols – abstracted away from specific input patterns – for solving complex tasks, arguing they provide a foundation to support higher cognition (Fodor, 1975; Marcus, 2001; Pylyshyn, 1984; also see Markman & Dietrich, 2000). In contrast, theories of grounded cognition suggest that all 'abstract' representations are grounded in, and therefore fully explained by, sensory-motor representations (Barsalou, 1999; Harnad, 1990). Indeed, sensory-motor variables and categories are often correlated in the real world and the brain may never need to represent 'category' in a way that can be disentangled from perception and action.

Here, we consider several competing accounts. Closely related to 'grounded cognition', some researchers emphasize a central role of action for cognition (e.g. Rizzolatti et al., 1987; Wolpert & Ghahramani, 2000; Wolpert & Witkowski, 2014), such that category representations could simply consist of the appropriate stimulus-motor representations and associations (Figure 1B). An alternative view holds that category-modulated stimulus representations are key for categorization, where stimulus information is transformed into a representation suited for categorization (as in cognitive models: e.g. Kruschke et al., 1992; Love et al., 2004; Nosofsky, 1986). In these models, an attention mechanism gives more weight to relevant features so that within-category stimuli become closer and across-category stimuli are pushed apart in representational space (Figure 1C). Finally, the brain may recruit an additional amodal, symbol-like concept representation (Fodor, 1975; Marcus, 2001; Newell, 1980; Pylyshyn, 1984) to explicitly code for category, separate from sensory-motor representations. For instance, sensory information is processed (e.g. modulated by category structure), then transformed into an abstract category representation before turning into a response (Figure 1D). This representation resembles an amodal symbol

in that it has its own representational format (e.g. orthogonal to sensory-motor codes), and act as a pointer between the relevant sensory signals (input) and motor responses (output). The advantage of such a representation is that it can play a role in solving the task and can persist across superficial changes in appearance and changes in motor commands. People's ability to reason and generalize in an abstract fashion suggests the brain is a type of symbol processor (Marcus, 2001).

Here, we aimed to test whether the brain constructs an abstract concept representation separate from stimulus and motor signals, if the 'category' code consists of category-modulated stimulus representations and motor codes, or if it simply consists of stimulus-motor mappings. We designed a probabilistic concept learning task where the stimulus, category, and motor variables were not perfectly coupled nor entirely independent, to allow participants to naturally form the mental representations required to solve the task, and used multivariate pattern analysis (MVPA) on fMRI data to examine how these variables were encoded across the brain. For evidence supporting the **amodal account (Figure 1D)**, some brain regions should encode category information but not the stimulus or response. For the **category-modulated sensory account (Figure 1C)**, regions should encode both stimulus and category information, with no regions that encode category without stimulus information. Finally, for the **sensory-motor account (Figure 1B)**, regions should code for category, stimulus, and motor response (separately or concurrently), with no regions encoding category without sensory or motor information.

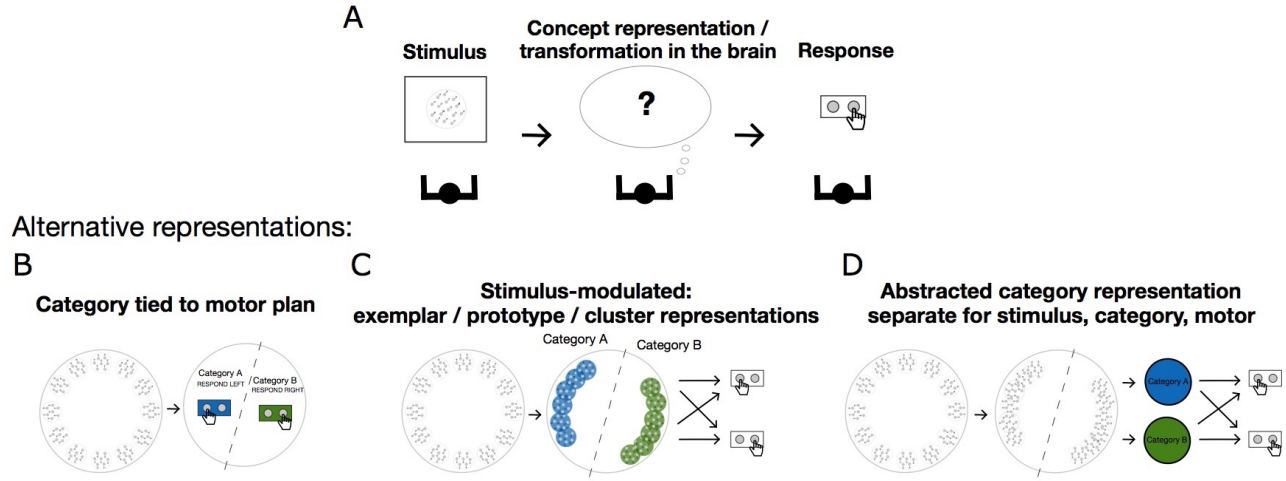

Figure 1. How the brain transforms stimulus into a concept representation for categorization. Stimuli are 12 motion dot patterns (100% coherent), from 0° to 330° in 30° steps. Blue and green colors denote the two categories. A) An observer must transform the percept into intermediate representations for accurate categorization behavior. B-D) Possible representations the brain might use for categorization. B) Each stimulus is associated with a motor response, where the category representation is grounded in sensory-motor codes. C) Stimulus-modulated representations as category representation. The stimulus representation is modulated by the category structure, which is turned into a motor representation for the response. D) The category-modulated stimulus representation is associated with an additional abstract representation of each category with a different representational format to the sensory motor codes (blue and green circles), is then turned into a response.

Results

We tested 33 human participants (23 female) on a probabilistic category learning task whilst they underwent a functional magnetic resonance imaging (fMRI) scan. On each trial, participants were presented with one of 12 dot-motion stimulus directions (100% coherent) and judged whether the stimulus direction belonged to category A or B (Figure 2). After each trial, participants were provided with probabilistic feedback which consisted of a face (category A) or building (category B) stimulus, which informed the participant which category the stimulus direction most likely belonged to. Naturalistic images were used to encourage task engagement and to produce a strong stimulus signal.

To test how stimulus, motor, and category are coded in the brain during category learning, we aimed to decouple these variables in the following ways. Probabilistic category feedback was used to encourage participants to form a strong internal category representation (i.e. forming and maintaining category knowledge despite incongruent feedback). Feedback was probabilistic such that the closer a stimulus was to the category bound, the more probabilistic the feedback was (Figure 2B). Furthermore, the response was flipped after each block (left/right button), to discourage associating each category with a single response across the experiment.

We used a model to estimate individual participants' subjective category bound (see Materials and Methods and Figure 3A-B). Briefly, the model assumes that participants form a mental decision boundary in the (circular) stimulus space to separate the categories, and there is some uncertainty of the placement of this bound. Formally, the model has three parameters: the first two determines bound placement (*b1* and *b2*), and the third is a standard deviation parameter (σ) that models the (normally distributed) noise in this bound. σ provides an estimate of how certain (lower σ) participants are of their boundary placement.

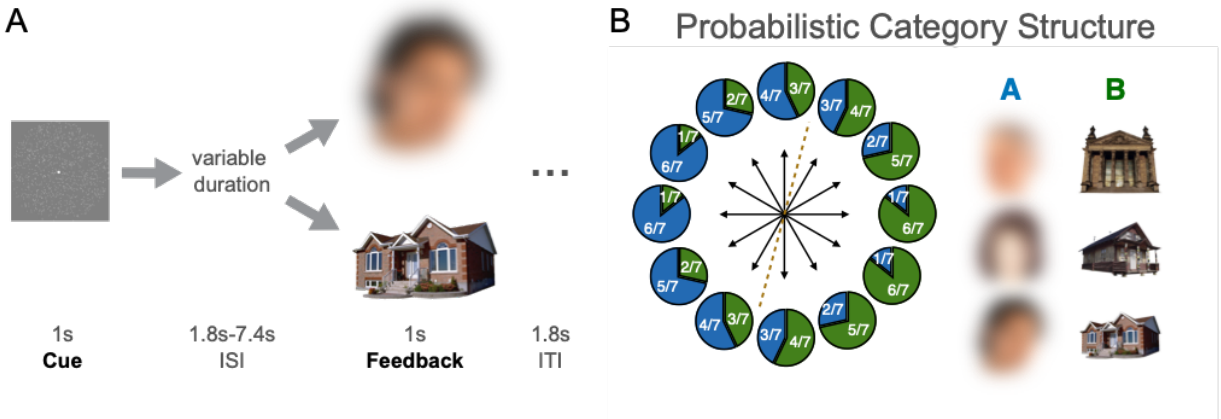

Figure 2. Behavioral task. A) On each trial, a dot-motion stimulus was presented and the participant judged whether it was in category A or B. At the end of each trial, probabilistic category feedback (a face or building stimulus) which informed the participant which category the motion stimulus most likely belonged to, allowing learning by trial and error. B) Probabilistic category structure. The closer the motion direction was to the category bound (dotted line), the more probabilistic the feedback. Note: faces are blurred for the preprint, but high quality images were used in the task.

To evaluate the three main accounts of how the brain organizes information for categorization, we performed MVPA across visual, parietal, and prefrontal regions of interest (ROIs) hypothesized to be involved in the task (see Materials and Methods). Specifically, we trained linear support vector machines (SVMs; c.f. Kamitani & Tong, 2005) to assess which brain regions contained information about category (A or B), stimulus (opposite directions), and response (left or right). For a strict test of a category signal unrelated to stimulus features, we trained a classifier to discriminate between motion directions in category A versus directions in category B, and subtracted this from a control classifier trained to discriminate between directions in category A rotated 90° versus directions in category B rotated 90°. This ensured that the classifier was not simply picking up information discriminating opposite stimulus directions.

Our findings most strongly align with the hypothesis that the brain constructs an amodal symbol for representing category, independent of sensory-motor variables. Specifically, we found a category signal over and above stimulus information in the middle portion of the left middle frontal gyrus (mMFG, area 8: $p=0.0025$, $q(FDR)=0.029$; Figure 3C) and left motion-sensitive area MT ($p=0.0086$, $q(FDR)=0.048$; Figure 3D). This is particularly striking since the category is based on the stimulus direction, and there was no hint of a direction signal in these regions (p's>0.41). Consistent with the idea that abstract category representations can aid performance, we found that the strength of category decoding was positively correlated with categorization performance in left MT (robust regression; $beta=0.74$, $p<0.05$; Figure 3E) with a similar trend for left mMFG ($beta=0.73$, $p=0.067$). We also confirmed that the category signal was stronger than the motor code in left mMFG by subtracting the classifier trained to discriminate between motion directions across categories from the motor classifier ($p=0.015$).

As expected, we found information coding motor response in motor cortex (right: $p=0.006$; Bonferonni-corrected for hemisphere $p=0.011$; left $p=0.095$, Bonferonni-corrected $p=0.19$), but no information about category or direction (p's>0.42), as the motor response for category was flipped across blocks.

Notably, category coding was only present for the participant-specific subjective category structure ('objective' category bound classifiers across all ROIs: p's>0.06). Furthermore, we found no evidence of category coding in the fusiform face area or in the parahippocampal place area (p's>0.31).

Although we did not find category and stimulus representations intertwined, this was not because stimulus representations were not decodable in our data. We trained a classifier on orientation in the early visual cortex and found activity coding orientation ($p<0.05$, Bonferroni-corrected for hemisphere). We also trained a 12-way classifier in order to assess if there was any information about the stimulus that would not be found simply by examining orientation or direction responses, and found that right MT encoded information the stimulus ($p=0.005$, $q(FDR)=0.03$), and a trend for right EVC ($p=0.06$, $q(FDR)=0.18$). Notably, there was no evidence for this in the left mMFG or MT, which encoded category (p's>0.74).

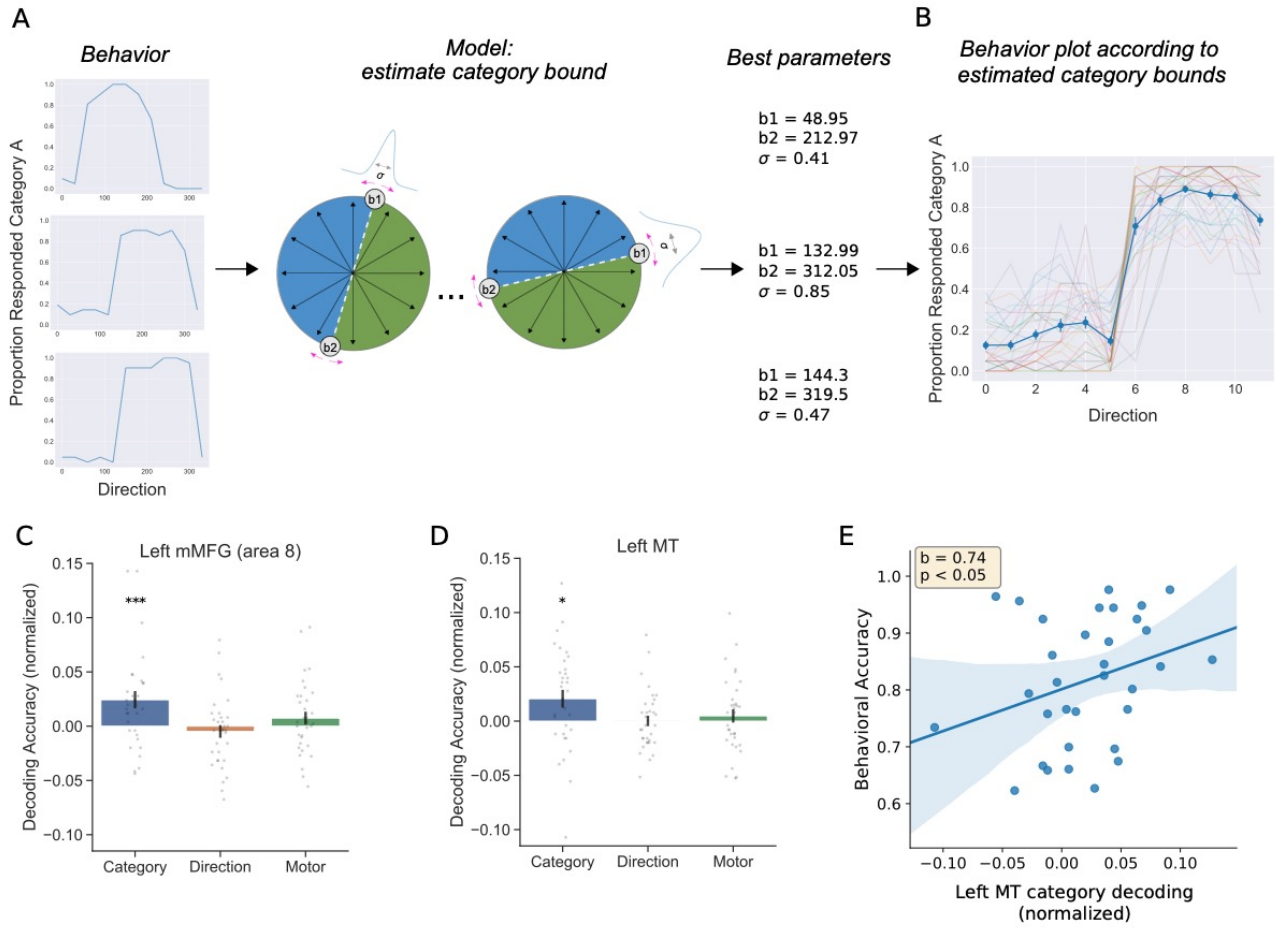

*Figure 3. Task model and main results. A) The model takes individual participant behaviour as input and estimates their subjective category bound (b1 and b2) and standard deviation (σ). B) Categorization behavior. Proportion category A responses plotted as a function of motion directions ordered by individual participants' estimated category boundary. Blue curve represents the mean, and error bars represent standard error of the mean (SEM). Translucent lines represent individual participants. C-E) Abstract category coding in left prefrontal cortex and area MT. Multivariate pattern analysis shows significant category coding in left mMFG (C) and left MT (D), with no evidence of stimulus and motor coding. Grey dots are individual participants. Error bars represent SEM. C) The strength of this category coding in MT was correlated with categorization behavior. The beta coefficient is from a robust regression analysis, and the shaded area represents 95% confidence intervals for the slope. *** p=0.0025, * p<0.01.*

Discussion

We found that the brain constructs an abstract category signal with a different representational format to sensory and motor codes for categorization. Specifically, left prefrontal cortex (PFC) and MT encoded category in absence of stimulus information, despite category structure being based on those stimulus features. Furthermore, the strength of this representation was correlated with categorization performance based on participants' subjective category bound estimated by our model.

Although some representations may be grounded in bodily sensations, for tasks that require flexibility and representations to support abstract operations, an amodal symbol of a different representational format to that of sensory-motor representations may prove useful (Fodor, 1975; Marcus, 2001; Newell, 1980; Pylyshyn, 1984). Indeed, a category representation tied to a motor plan or stimulus feature would facilitate stimulus-motor representations effectively in specific circumstances, but become unusable given slight changes in context. In this study, it was possible to solve the task using multiple ways, such as a combination of the sensory-motor variables, using a category-modulated sensory representation, or additionally recruiting an amodal representation. Although our task did not encourage participants to solve the task in one of these ways, we found that the brain produces an additional abstract representation for categorization.

Previous studies have found strong stimulus coding and category-related modulation stimulus representations during concept or perceptual learning (e.g. Freedman & Assad, 2006; Kourtzi et al., 2005; Watanabe et al., 1998). This is consistent with our view that some brain regions that play a role in categorization are grounded by sensory-motor representations (Barsalou, 1999), whereas other regions such as the PFC transforms these signals into a different representational format. An abstract, amodal symbol can act as a pointer to the relevant sensory and motor signals, functioning as an intermediary between input and output. By using a task where the variables were more uncoupled but not entirely independent, our task allowed us to observe whether the brain mainly relies of sensory-motor signals or in fact performs an additional step to construct an abstract representation of category.

In real-world scenarios, there are often no explicit rules and reliable feedback is rare. Building an abstract representation that can be mapped onto different contexts can be useful in real-world tasks, where the meaning of a situation can remain constant whilst the contextually appropriate stimulus or response changes. As we find here, the brain constructs an amodal, abstract representation with a different representational format separate from sensory-motor codes, well-suited for flexible cognition in a complex world.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*(14). https://doi.org/10.3389/fninf.2014.00014

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex*, *9*(5), 415–430. https://doi.org/10.1093/cercor/9.5.415

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, *12*(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Avants, B. B., Tustison, N., & Song, G. (2009). Advanced Normalization Tools (ANTS). *Insight Journal*, *2*, 1–35.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Corbetta, M., Miezin, F. M., Shulman, G. L., & Petersen, S. E. (1993). A PET study of visuospatial attention. *J Neurosci*, *13*(3), 1202–1226. http://www.ncbi.nlm.nih.gov/pubmed/8441008

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5), 171–78. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of Multiple, Independent Categories in the Primate Prefrontal Cortex. *Neuron*, *66*, 796–807. https://doi.org/10.1016/j.neuron.2010.05.005

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(4), 821–839. https://doi.org/10.1037/a0027865

Dubner, R., & Zeki, S. M. (1971). Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Research*, *35*(2), 528-32. https://doi.org/10.1016/0006-8993(71)90494-X

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci*, *2*(11), 820–829. https://doi.org/10.1038/35097575

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual

environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Erez, Y., & Duncan, J. (2015). Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. *Journal of Neuroscience*, *35*(36), 12383e12393. https://doi.org/10.1523/JNEUROSCI.1134-15.2015

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A.I. et al. 2018. "FMRIPrep 1.2.3." *Software*. Zenodo. https://doi.org/10.5281/zenodo.852659.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1315235110

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Fonov, V., Evans, A., McKinstry, R., Almli, C., & Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, *47 (Supp.*, S102. https://doi.org/10.1016/s1053-8119(09)70884-5

Freedman, D. J., & Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, *443*, 85–88. https://doi.org/10.1038/nature05078

Freedman, D. J., & Assad, J. A. (2016). Neuronal Mechanisms of Visual Categorization: An Abstract View on Decision Making. *Annual Review of Neuroscience*, *100*, 1407–1419. https://doi.org/10.1146/annurev-neuro-071714-033919

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–24. https://doi.org/10.1016/j.neuroimage.2013.04.127

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. In *Neuron* (Vol. 14, Issue 3, pp. 477–485). https://doi.org/10.1016/0896-6273(95)90304-6

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python. *Frontiers in Neuroinformatics*, *5*(13). https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., et al. 2018. "Nipype." *Software*. Zenodo. https://doi.org/10.5281/zenodo.596855.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*, 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Jackson, J., Rich, A. N., Williams, M. A., & Woolgar, A. (2017). Feature-selective attention in frontoparietal cortex: Multivoxel codes adjust to prioritize task-relevant information. *Journal of Cognitive Neuroscience*, *29*(2), 310–321. https://doi.org/10.1162/jocn_a_01039

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. https://doi.org/10.1016/S1361-8415(01)00036-6

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1098/Rstb.2006.1934

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci*, *23*, 315–341. https://doi.org/10.1146/annurev.neuro.23.1.315

Kourtzi, Z., Betts, L. R., Sarkheil, P., & Welchman, A. E. (2005). Distributed neural plasticity for shape learning in the human visual cortex. *PLoS Biology*, *3*(7), e204. https://doi.org/10.1371/journal.pbio.0030204

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Lanczos, C. (1964). Evaluation of Noisy Data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*. https://doi.org/10.1137/0701007

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press.

Markman, A. B., & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences*, *4*(12), 470–475. https://doi.org/10.1016/S1364-6613(00)01559-X

Mesulam, M. M. (1981). A cortical network for directed attention and unilateral neglect. *Ann Neurol*, *10*(4), 309–325. https://doi.org/10.1002/ana.410100402

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.90248.2008

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.115.1.39

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perktold, J., & Seabold, S. (2010). Statsmodels: Econometric and Statistical Modeling with Python Quantitative histology of aorta View project Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Rizzolatti, G., Riggio, L., Dascola, I., & Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1A), 31–40. http://www.ncbi.nlm.nih.gov/pubmed/3574648

Roy, J. E., Riesenhuber, M., Poggio, T., & Miller, E. K. (2010). Prefrontal cortex activity during flexible categorization. *Journal of Neuroscience*, *30*(25), 8519–8528. https://doi.org/10.1523/JNEUROSCI.4837-09.2010

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annual Review of Neuroscience*, *33*, 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*(17), 318–320. https://doi.org/10.1038/415318a

Watanabe, T., Harner, A. M., Miyauchi, S., Sasaki, Y., Nielsen, M., Palomo, D., & Mukai, I. (1998). Task-dependent influences of attention on the activation of human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(19), 11489–11492. https://doi.org/10.1073/pnas.95.19.11489

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*, 1212–1217. https://doi.org/10.1038/81497

Wolpert, D. M., & Witkowski, J. (2014). A Conversation with Daniel Wolpert. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 297–298. https://doi.org/10.1101/sqb.2014.79.19

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.

https://doi.org/10.1109/42.906424

Materials and Methods

Participant Recruitment and Behavioral Session

We recruited participants to partake in a behavioral session to assess their ability to learning the probabilistic concept learning task. Participants that achieved greater than 60% accuracy over two blocks of the behavioral task were invited to participate in the fMRI study. We tested 131 participants, 39 of which passed the behavioral criterion, were MRI compatible, and agreed to participate in the fMRI study.

Participants (fMRI study)

39 participants participated in the fMRI study. Six participants were excluded due to lower than chance performance, misunderstanding the task, or falling asleep during the experiment. The remaining 33 participants (23 female) were aged 19-34 (mean: 24.04 ± 0.61 standard error of the mean; SEM). The study was approved by the UCL Research Ethics Committee (Reference: 1825/003).

Stimuli and Apparatus

Stimuli consisted of coherently moving dots moving produced in PsychoPy (Peirce et al., 2019), images of faces and buildings (main task), and images of flowers and cars (practice). In each dot-motion stimulus there were 1000 dots, dots were 2 pixels in size, and moved at a velocity of ~0.8° per second. The dot-motion stimuli and images were 12° in diameter (or on longest axis). The fixation point was a black circle with 0.2° diameter. A grey circle (1° diameter) was placed in front of the dot stimulus but behind the fixation point to discourage smooth pursuit. The natural images were provided by members of Cognitive Brain Mapping Lab at RIKEN BSI. The task was programmed and run in PsychoPy in Python 2.7. The task was presented on an LCD projector (1024 x 768 resolution) which was viewed through a tilted mirror in the fMRI scanner. We monitored fixation with an eye tracker (Eyelink 1000 Plus, SR Research, Ottawa, ON, Canada), and reminded subjects to maintain fixation between runs as necessary.

Behavioral Task

To examine how the brain constructs an appropriate mental representation for categorization, we designed a concept learning task to be performed in the fMRI scanner. Specifically, we set out to test whether any brain regions coded for an abstract category signal separate from stimulus and motor signals, if the category signal mainly consisted of category-modulated sensory signals, or if the category signal was simply a combination or co-existence of sensory-motor signals. To this end, we designed a probabilistic concept learning task where the task variables (category, stimulus, and motor response) were not perfectly coupled, nor entirely independent.

On each trial, participants were presented with a set of moving dots moving coherently in one direction and were required to judge whether it belonged to one category (‘Face’) or another (‘Building’) with a corresponding left or right button press. The

motion stimulus was presented for 1 second (s), followed by an inter-stimulus interval ranging from 1.8-7.4s (jittered), then the category feedback for 1s. The inter-trial interval was 1.8s.

The moving-dot stimuli spanned 12 directions from 0° to 330° in 30° steps, with half the motion directions assigned to one of two categories determined by a category bound. For half the participants, the objective category bound was placed at 15°, so that directions from 30° to 180° were in one category, and directions from 210° to 330° and 0° were in the other category. For the other half of the participants, the objective category bound was placed at 105°, so that directions from 120° to 270° were in one category, and directions from 0° to 90° and 300° to 330° were in the other category.

The corrective category feedback consisted of a face or building stimulus, which informed the participant which category the motion stimulus was most likely part of. The feedback was probabilistic such that the closer to the bound a stimulus was the more probabilistic the feedback was (see Figure 1A).

The category-response association flipped after each block (e.g. left button press for category A in the first block, right button press for category A in the second block), to discourage participants simply associating each category with a motor response across the experiment.

In sum, the task required participants to learn the category that each motion-dot stimulus belonged to by its probabilistic association to an unrelated stimulus (face or building as category feedback), whereby the probabilistic feedback encouraged participants to actively learn and form a strong internal category representation. Furthermore, the category-motor association was flipped across blocks. Together, the category, stimulus, and motor variables were weakly coupled, allowing us to assess whether there are brain regions that code for these variables together or independently of one another.

Participants completed three blocks or four block runs (four participants performed an extra block due to low performance on early block runs). In each block participants completed seven trials per direction condition, giving 84 trials per block.

Localizer Tasks

After the main experiment, participants completed two localizer runs. To localize the face-selective fusiform face area (FFA; Allison et al., 1999; Kanwisher et al., 1997) and place-sensitive parahippocampal place area (PPA; Epstein & Kanwisher, 1998) in individuals, participants completed an event-related localizer scan where they were presented with faces and buildings, and made a response when they saw an image repeat (1-back task), which was followed by feedback (the fixation point changed to green for correct and red for incorrect). On each trial, an image of a face or building was presented for 1s with inter-stimulus intervals between stimulus and feedback (green/red fixation color change) ranging from 1.8s to 7.4s (jittered), with an inter-trial interval of 1.8s. A total of 42 faces and 42 buildings were presented in a random order. Participants also completed a motion localizer run that was not used here.

Experimental Procedure: Behavioral Session

The task was the same as described above except that the images used for feedback were pictures of flowers and cars. To ensure participants understood the task, they were given four practice runs with versions of the task gradually increasing in task complexity. Prior to each practice run, the experimenter explained the task to the participant. In the first run, a moving dot stimulus was presented, and participants had to judge whether it belonged to one category ('Flower') or another ('Car'). If they responded correctly, the fixation point turned green, if they responded incorrectly, it turned red, and if they responded too slowly, it turned yellow. The category ('Flower' or 'Car') feedback was deterministic. In this first run, the category boundary was at 90° (up-down rule), and motion directions were presented in sequential order around the circle. In the second run, the task was the same except the motion directions were presented in a random order. In the third run, the category feedback was probabilistic as in the main experiment. In the fourth run, it was a practice run of the main task with face and building images, and a new category boundary (15° or 105°) that will be used in the main experiment. Once participants completed the practice runs and were comfortable with the task, they proceeded to complete three experimental runs.

Experimental Procedure: fMRI Session

Participants were given one practice block run as a reminder of the task, then proceeded to complete the main experiment in the scanner. Participants completed three blocks of the probabilistic category learning task, then a motion localizer block and a face-scene localizer block (block order for localizer runs were counterbalanced across participants). After the scan session, participants completed a post scan questionnaire to assess their understanding of the task and to report their subjective category rule.

Behavioral Model and Data Analysis

The probabilistic nature of the feedback meant that participants did not perform exactly according to the objective category rule determined by the experimenter, and inspection of behavioral performance curves suggested that most participants formed a category rule slightly different to the objective rule. To get a handle on the category rule participants formed, we applied a behavioral model to estimate each participants' subjective category boundary from their responses.

The model contains a decision bound defined by two points, b_1 and b_2, on a circle (0° to 359°). Category A proceeds clockwise from point b_1 whereas Category B proceeds clockwise from b_2. Therefore, the positions of b_1 and b_2 define the deterministic category boundary between categories A and B. To illustrate, if b_1 = 15° and b_2 = 175°, stimulus directions from 15° to 175° would be in one category, and stimulus directions from 175° to 359° and from 0° to 15° would be in the other category. Note that the number of stimulus directions are not constrained to be equal across categories, as illustrated in this example (five and seven directions in each category). Despite this, there were six stimulus directions in each category for most participants.

The only source of noise in this model are the positions of b_1 and b_2 which are normally distributed as $\mathcal{N}(0, \sigma)$. As the σ parameter – the standard deviation of the positions of b_1 and b_2 – increases, the position of the boundary for a given trial becomes noisier and therefore it becomes more likely that an item may be classified contrary to the position of the boundary. In practice, no matter the value of σ, it is always more likely that an item will be classified according to the positions of b_1 and b_2. The standard deviation parameter provides as estimate of how uncertain participants were of the category boundary. If a participant responded perfectly consistently according to a set of bounds (deterministically), σ would be low, whereas if the participant was more uncertain of the bound locations and responded more probabilistically, σ would be higher.

The probability a stimulus x is an A or B is calculated according to whichever boundary b_1 or b_2 is closer. This is a numerical simplification as it is possible for the further boundary to come into play and even for boundary noise to lead to b_1 or b_2 to traverse the entire circle. However, for the values of σ we consider, both of these possibilities are highly unlikely. The probability that stimulus x is labelled according to the mean positions of b_1 and b_2 is:

$$1 - p\left(z > \frac{|x - b_x|}{\sigma}\right)$$

where z is distributed according to the standard normal distribution and b_x is b_1 or b_2, whichever is closer to x. Intuitively, the further the item is from the boundary position, the more likely it is to classified according to the boundary position as noise (i.e., σ) is unlikely to lead to sufficient boundary movement that trial. The probability an item is labeled in the alternative category (i.e., "incorrect" responses against the bound defined by the mean positions of b_1 and b_2) is simply 1 minus the above quantity.

In other words, the probability stimulus is in a certain category is a Gaussian function of the distance to the closest bound, where the further away the stimulus is from the bounds, the more likely it is part of that category (see Figure 3A for an illustration of the model).

Maximum-likelihood estimation was used to obtain estimates for each participant (using the optimize function in SciPy). Model estimates of the subjective category bound fit participant behaviour as expected. Specifically, there was high accuracy (concordance) with respect to the estimated subjective category bound (mean proportion correct: 0.82±0.01 SEM; see Figure 3B).

Modeling and analyses was performed in Python 3.7.

MRI data acquisition

Functional and structural MRI data were acquired on a 3T TrioTim scanner (Siemens, Erlangen, Germany) using a 32-channel head coil at the Wellcome Trust Centre for Neuroimaging at UCL. An EPI-BOLD contrast image with 40 slices was acquired in 3-mm^3 voxel size, repetition time (TR) = 2800 ms, echo time (TE) = 30 ms, and the flip angle was set to 90°. A whole brain fieldmap with 3-mm^3 voxel size was obtained with

a first TE = 10 ms, second TE = 12.46ms, TR = 1020 ms, and the flip angle was set to 90°. A T1-weighted structural image was acquired with 1-mm^3 voxel size, TR = 2.2 ms, TE = 2.2 ms, and the flip angle was set to 13°.

fMRI pre-processing

Results included in this manuscript come from preprocessing performed using fMRIprep 1.2.3 (Esteban, et al., 2018; Esteban et al., 2019; RRID:SCR_016216), which is based on Nipype 1.1.6-dev (Gorgolewski et al., 2011; Gorgolewski et al., 2018; RRID:SCR_002502).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using N4BiasFieldCorrection (Tustison et al. 2010, ANTs 2.2.0), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using antsBrainExtraction.sh (ANTs 2.2.0), using OASIS as target template. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009; RRID:SCR_008796) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0, RRID:SCR_004757, Avants et al., 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001).

Functional data preprocessing

For each of the five or six BOLD runs found per subject (across all tasks and sessions; three or four task runs two localizer runs), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A deformation field to correct for susceptibility distortions was estimated based on a field map that was co-registered to the BOLD reference, using a custom workflow of fMRIPrep derived from D. Greve's epidewarp.fsl script and further improvements of HCP Pipelines (Glasser et al., 2013). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9, Jenkinson & Smith, 2001) with the boundary-based registration (Greve & Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox & Hyde, 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled to MNI152NLin2009cAsym standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-

stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with 128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.4.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

fMRI general linear model

We used the general linear model (GLM) in FMRI Expert Analysis Tool (FEAT; Woolrich et al., 2001; FMRIB Software Library (FSL) version 6.00; https://fsl.fmrib.ox.ac.uk/fsl/) to obtain estimates of the task-evoked brain signals for each stimulus, which was used for subsequent multivariate pattern analyses (MVPA).

For the main GLM, we included one explanatory variable (EV) to model each motion stimulus trial (estimating trial-wise betas for subsequent MVPA) and an EV for each category feedback stimulus linked to each motion stimulus condition (12 EVs, not used in subsequent analyses; see trial-wise GLM examining the feedback response below). No spatial smoothing was applied. Stimulus EVs were 1s with inter-stimulus intervals between stimulus and feedback ranging from 1.8s to 7.4s (jittered), and the inter-trial interval was 1.8s. Each block run was modelled separately for leave-one-run-out cross-validation for (MVPA).

To examine motor-related brain responses, we performed an additional GLM using the same number of EVs except the EVs were time-locked to the response rather than the motion stimulus (stimulus time plus reaction time) and modelled as an event lasting

0.5s. For trials without a response, the stimulus was modelled from stimulus onset as done in the main GLM, then excluded in subsequent motor-related analyses. Feedback stimuli were modelled with a single EV as above.

To localize the face-selective FFA and place-sensitive PPA, we performed an additional GLM in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). We applied spatial smoothing using a Gaussian kernel of full width at half maximum (FWHM) 6 mm, and included one EV for faces and one EV for building stimuli, and polynomials of degrees 0:6 to model drift in the data. For the GLM, one EV was used to model the face and one EV for the building stimuli. Stimulus EVs were 1s with inter-stimulus intervals between stimulus and feedback (green/red fixation point color change) ranging from 1.8s to 7.4s (jittered), with an inter-trial interval of 1.8s.

In order to localize the FFA and PPA, we included the contrast Faces > Buildings, Buildings > Faces, and overall Visual Activation (Faces & Buildings). To define individual participant ROIs, we used minimum statistic conjunctions with visual activations. Specifically, to localize the FFA, the conjunction was: (Face > Building) & Visual. For the PPA, the conjunction was: (Building > Face) & Visual. The rationale behind this conjunction is that functional ROIs should not only simply be selective but also visually responsive (all voxels that were de-activated by visual stimulation were not included). The conjunction was thresholded liberally at $p<0.01$ uncorrected. The peaks for each functional ROI was detected visually in the SPM results viewer, and we extracted the top 100 contiguous voxels around that peak.

For some participants, we were unable to obtain a clear peak in the expected brain region for the relevant conjunction (left or right fusiform or parahippocampal gyrus). Those participants were excluded from analyses involving those ROIs (see next section).

To examine information during the category feedback, performed an additional GLM modelling the same events as the main GLM (locked to motion stimuli and feedback stimuli), except that one EV was used to model each feedback trial (estimating trial-wise betas for subsequent MVPA), and one EV for each motion stimulus condition (12 EVs). This additional GLM was used to estimate the trial-wise feedback mainly for practical reasons. If modelling all cue and feedback trials, it becomes a substantially larger model for FSL. By modelling the cue period trials in a separate GLM to the feedback trials, we were able to reduce the number of EVs per model (96 rather than 168).

We also had one motion localizer run that was not used in this study.

Regions of interest

To study how the brain represented category, stimulus, and response variables in the probabilistic concept learning task, we focused on a set of visual, parietal, and prefrontal brain regions of interest (ROIs) that were hypothesized to be involved in coding these variables.

We selected anatomical masks from Wang et al. (2014; https://scholar.princeton.edu/napl/resources) to examine areas involved in early visual

processing, motion processing, and attention, including early visual cortex (EVC; V1, V2, and V3 merged), motion-sensitive area MT/V5 (Dubner & Zeki, 1971) and the intraparietal sulcus (IPS). We included EVC to assess stimulus-related representations including orientation and direction. The IPS is implicated in both attention (Corbetta et al., 1993; Kastner & Ungerleider, 2000; Mesulam, 1981) and category learning (Freedman & Assad, 2016; Seger & Miller, 2010). However, we did not have strong reasons to focus on specific parts of the IPS, so we merged IPS1 to IPS5 to make a large IPS ROI.

Since these masks are provided in T1 structural MRI space (1-mm^3), when they were transformed into individual participant functional space (3-mm^3), several masks did not cover grey matter accurately (too conservative, thereby excluding some grey matter voxels). Therefore, we applied a small amount of smoothing to the mask (with a Gaussian kernel of 0.25 mm, using fslmaths) for a more liberal inclusion of neighboring voxels, before transforming it to individual-participant space. In addition, several potential ROIs were too small to be mapped onto our functional scans. Specifically, there were several participants with zero voxels in those masks after transforming to functional space, even with smoothing. This included the motion-sensitive area MST and the superior parietal lobule (SPL1), which were not included.

Prefrontal cortex is strongly implicated representing abstract task variables (Duncan, 2001; Miller & Cohen, 2001) and task-relevant sensory signals (e.g. Erez & Duncan, 2015; Goldman-Rakic, 1995; Jackson et al., 2017; Meyers et al., 2008; Roy et al., 2010). We selected prefrontal regions implicated in cognitive control and contain task-related representations from the multiple-demand system (Duncan, 2010; Fedorenko et al., 2013)http://imaging.mrc-cbu.cam.ac.uk/imaging/MDsystem) including the posterior, middle (approximately area 8), and anterior (approximately area 9) portion of the middle frontal gyrus (MFG).

Primary motor cortex was selected to examine representations related to the motor response and to test for any stimulus or category signals. Primary motor cortex masks were taken from the Harvard-Oxford atlas.

Finally, we used functionally localized FFA and PPA ROIs for each individual participant to test for a category signal during the motion stimulus and during feedback in these regions. It is worth noting we are interested in assessing the information coding the *learnt* category (category A versus B), not the probabilistically presented face versus building feedback stimulus.

There were four participants for which we could not find clear peaks for the left FFA (participant 01, 15, 19, 24), five participants for the right FFA (participant 02, 10, 12, 15, 17, 20), seven participants for the left PPA (participant 05, 07, 09, 10, 18, 24, 25), and six participants for the right PPA (participant 05, 09, 23, 24, 28, 33). Since we were unable to reliably localize these areas for all participants in both hemispheres, we used unilateral ROIs for participants with unilateral FFA/PPA ROIs, and excluded participants for that ROI if they did not have either a left or right FFA or PPA ROI. In summary, when testing the FFA, we excluded two participants (no left or right FFA in participant 08, 15), and when testing the PPA, we excluded four participants (no left or right PPA in participant 05, 08, 09, 24).

Apart from the FFA and PPA, we included both left and right ROIs. Masks were transformed from standard MNI space to each participant's native space using Advanced Normalization Tools (ANTs; Avants et al., 2009).

Multivariate pattern analysis (MVPA)

To examine brain representations of category, stimulus, and motor response, we used MVPA across our selected ROIs. Specifically, we trained linear support vector machines (c.f. Kamitani & Tong, 2005) to assess which brain regions contained information about the category ('Face' or 'Building'), stimulus (direction, orientation, and 12-way classifier), and motor response (left or right button press).

Decoding analyses were performed using linear support vector classifiers (SVC; C = 0.1) using Scikit-learn Python package (Pedregosa et al., 2011) with a leave-one-run-out cross-validation procedure.

To test for category coding, we first trained a classifier to discriminate between motion directions belonging to the two categories for each participant's subjective category bound. To ensure that this was a pure category signal unrelated to stimulus differences (e.g. simply decoding opposite motion directions), we trained a classifier based on the participant's subjective category bound, rotated 90°. For a stricter test of a category signal, we subtracted the classification accuracy of the first classifier from accuracy of the second classifier. To ensure this category signal was not related to motor preparation or the response, we also subtracted the former category classifier accuracy from a motor response classifier accuracy (discriminating between left versus right button presses).

For stimulus direction coding, we trained classifiers to discriminate between all six pairs of opposite motion directions (0° versus 180°, 30° versus 210°, etc.) and averaged across the classification accuracies.

To examine motor response coding, we trained classifiers to discriminate between left and right button presses on the GLM where we locked the EVs to the motor responses (reaction times).

As a control analysis, we tested whether a classifier trained on the objective category structure (i.e. defined by the experimenter) produced similar results to the subjective category analysis. The procedure was the same as the category classifier above, except that the directions in each category were determined by the experimenter. In another set of control analyses, we assessed if there was any information about stimulus. We tested for orientation coding by training a classifier on all 12 pairs of orthogonal orientations irrespective of the motion direction (0° versus 90° and 0° versus 270°, 30° versus 120°, 30° versus 300°), and averaged across the classification accuracies. Finally, for a more general measure of stimulus coding, we trained a 12-way classifier to assess stimulus coding for each motion direction.

We used one-sample t-tests (one-tailed) against chance-level performance of the classifier (using SciPy; Jones et al., 2001). Multiple comparisons across ROIs were corrected by controlling the expected false-discovery rate (FDR) at 0.05 (Perktold & Seabold, 2010). For decoding category, we corrected across 12 ROIs (all apart from

bilateral EVC and bilateral motor cortex), and for the 12-way classifier, we corrected across 14 ROIs (excluding bilateral motor cortex). Bonferonni correction was used for tests with two ROIs (correcting for visual and motor hemispheres for orientation and motor decoding, respectively). For others, we report the uncorrected p-values since none survived even without correction. MVPA and statistical analyses were performed in Python 3.7.

Brain-behavior correlations

To assess whether the brain's representation of the abstract category signal contributed to categorization performance, we performed robust regression (Seabold & Perktold, 2010) to assess the relationship between categorization performance (concordance to the estimated subjective category structure) with classifier accuracy for the category for the ROIs with greater than chance classification accuracy for category.

Acknowledgements

We thank Johan Carlin for his help on experimental design and data collection, and Amna Ali for her help on data collection. We thank the Love Lab for the helpful discussions on the project. We are grateful to members of Cognitive Brain Mapping Lab at RIKEN BSI for sharing natural images used in this study.

Competing Interests

Authors declare no competing interests.