UNIVERSITY OF CALIFORNIA SAN DIEGO

Representations underlying efficient and selective processing in the visual system

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy

in

Neurosciences with a Specialization in Computational Neurosciences

by

Margaret Marie Henderson

Committee in charge:

Professor John Serences, Chair
Professor Eran Mukamel
Professor Tatyana Sharpee
Professor Viola Störmer
Professor Bradley Voytek

2021

The Dissertation of Margaret Marie Henderson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CNN: convolutional neural network

FI: Fisher information

FIB: Fisher information bias

SEM: standard error of the mean

FDR: false discovery rate

fMRI: functional magnetic resonance imaging

ROI: region of interest

HRF: hemodynamic response function

V1: primary visual cortex, or striate cortex

IT: inferotemporal cortex

LOC: lateral occipital complex

IPS: intraparietal sulcus

PFC: prefrontal cortex

MD: multiple-demand network

M1: primary motor cortex

S1: primary somatosensory cortex

PMc: premotor cortex

WM: working memory

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I want to thank my parents for their unconditional love and support. My parents always encouraged me to follow my passions and ambitions and I would not have been able to achieve this without them. I also want to thank my sister and my brother for their love and support through my life.

I also want to thank my advisor, John Serences, for being an amazing mentor and a constant source of inspiration. His ability to provide supportive guidance and clear insight, while also giving me the freedom to pursue several different research interests while in the lab, has been incredibly beneficial to me and has helped me learn and grow as a scientist. I also want to acknowledge all the current and former members of the Serences Lab whom I have had the fortune to work with. Chaipat Chunharas, Eddie Ester, Sirawaj Itthipuripat, Steph Nelli, Rosanne Rademaker, Nuttida Rungratsameetaweemana, Mary Smith, Tommy Sprague, and Vy Vo are all wonderful people and being in the lab with them made me feel inspired as well as entertained. My current lab-mates, including Kirsten Adam, Angus Chapman, Anna Shafer-Skelton, Tim Sheehan, Sunyoung Park, and Janna Wennberg, have been an incredibly supportive group and kept me motivated, especially during this last year. I couldn't have asked for a better advisor or better group of lab-mates.

I am eternally grateful to the UCSD Neurosciences program, for providing a wonderful and welcoming community during my time in graduate school. In particular, the 16 of us who started the program together in 2015 have developed a close bond, and it's been such a privilege to watch everyone grow and evolve during graduate school. I couldn't have made it through without this community, and will always be grateful for our friendship.

Finally, I want to thank my partner Chris Donahue for his unwavering encouragement and for making San Diego feel like home for me. I can't wait to see what's next for us.

Chapter 1 has been submitted for publication and is currently in revision. The author list and working title is Henderson, M. M., & Serences, J. T. (2021). Biased orientation representations can be explained by experience with non-uniform training set statistics. *In revision*. The dissertation author was the primary investigator and author of this paper.

Chapter 2 is a reprint of the material as it appears in: Henderson, M., & Serences, J. T. (2019). Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features. *Journal of Neurophysiology*, 121(4): 1410-1427. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is currently in preparation for publication. The author list and working title is Henderson, M. M., Rademaker, R. L., & Serences, J.T. (2021). Prospective response planning degrades spatial memory representations in human visual cortex. *In prep*. The dissertation author was the primary investigator and author of this paper.

Funding support was given by NIMH Training Grant in Cognitive Neuroscience (T32-MH020002), and by NEI R01-EY025872 to Dr. John Serences.

I also thank the San Diego Supercomputer Center for computing support, and the NVIDIA corporation for donation of a Quadro P6000 GPU that was used in the research described in Chapter 1.

VITA

2015	Bachelor of Science, Cornell University
2019	Master of Science, University of California San Diego
2021	Doctor of Philosophy, University of California San Diego
2021	Postdoctoral Fellow, Carnegie Mellon University

PUBLICATIONS

Henderson, M.M., & Serences, J.T. (2020). Biased orientation representations can be explained by experience with non-uniform training set statistics. *bioRxiv.*

Henderson, M.M.*, Vo, V.A.*, Chunharas, C., Sprague, T.C., Serences, J.T. (2019). Multivariate analysis of BOLD activation patterns recovers graded depth representations in human visual and parietal cortex. *eNeuro*.

Henderson, M.M. & Serences, J.T. (2019). Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features. *Journal of Neurophysiology.*

Henderson, M.M., Gardner, J., Raguso, R.A., Hoffman, M.P. (2017). *Trichogramma ostriniae* (Hymenoptera: Trichogrammatidae) response to relative humidity with and without host cues. *Biocontrol Science and Technology*.

* These authors made equal contributions.

ABSTRACT OF THE DISSERTATION

Representations underlying efficient and selective processing in the visual system

by

Margaret Marie Henderson

Doctor of Philosophy in Neurosciences with a Specialization in Computational Neurosciences

University of California San Diego, 2021

Professor John Serences, Chair

The visual system is tasked with processing massive amounts of sensory information, but its computational power is constrained by limited metabolic resources. This results in the need for selective filtering of inputs so that the most relevant information is highlighted for further processing. This selectivity can be implemented slowly over the course of visual system evolution and development, by adapting the tuning properties of sensory neurons to optimally represent the stimuli that are most likely to be encountered during natural behavior. It can also be implemented more rapidly during the behavior of an individual organism, as when the brain enhances representations of target objects that are known to be relevant in a given context. Finally, selectivity can be implemented in the memory system, by adaptively re-formatting

remembered information according to the demands of a particular task. In this dissertation, I will present three complementary experiments that exemplify the role of selective, efficient information processing in shaping visual system function.

INTRODUCTION

At every moment, the world presents an infinite amount of information that has potential relevance to an observer. The role of sensory systems is to interpret this information in a meaningful way, allowing the observer to more effectively interact with its environment. At the same time, biological constraints place a limit on the amount of processing that can be performed at once, encouraging the senses to be selective about the amount and type of information that is collected. As a result, rather than forming a faithful and complete representation of every encountered stimulus, the brain identifies stimuli that are likely to be relevant for an organism, based on factors like their frequency of appearance or their resemblance to a previously seen item, and highlights these for further processing. This selection process can occur over multiple timescales, including slow changes to the brain's architecture that occur over evolutionary time, faster changes that occur as individual organisms adapt to their environments during development, and rapid changes in neural response properties observed when switching between cognitive tasks. In this dissertation, I will provide several illustrations of how the need for selectivity shapes neural response properties in the visual system. In **Chapter 1,** I will examine the role of efficient coding of natural scene statistics in creating low-level orientation biases in the visual system. Next, in **Chapter 2,** I will investigate how the brain identifies target items based on relevant high-level object properties, while simultaneously ignoring irrelevant properties. Finally, in **Chapter 3,** I will show that selective information processing shapes the neural mechanisms of working memory (WM), by demonstrating that the brain can flexibly adjust the way it represents previously seen stimuli, in accordance with behavioral goals. The remainder of this section will provide additional background on these three experiments.

Due to metabolic resource constraints, the brain is under pressure to represent sensory stimuli in an efficient way, using the smallest possible number of energetically expensive events like action potentials (Barlow, 1961). This concept of efficient coding can be applied to a number of sensory domains, including the auditory and olfactory systems, but has been explored the most extensively in the visual system (Gervain & Geffen, 2019; Simoncelli & Olshausen, 2001; Tesileanu, Cocco, Monasson, & Balasubramanian, 2019). The visual properties of the natural world are distributed in a somewhat predictable way, and it is thought that the visual system exploits experience with this distribution by adapting its coding properties to these statistics. This adaptation has been demonstrated for neural populations coding relatively simple visual features such as luminance intensity, spatial frequency, motion, color, and orientation (Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001; Girshick, Landy, & Simoncelli, 2011; Simoncelli & Olshausen, 2001), and can even explain the appearance of biased perception and illusions related to these features (Girshick et al., 2011; Wei & Stocker, 2015; Weiss, Simoncelli, & Adelson, 2002). The organization of the primate ventral visual system is also thought to be driven in part by the statistics of high-level visual stimuli, including the mid-level features that tend to vary most among object categories, and the retinal locations at which features are most likely to appear (Bao, She, McGill, & Tsao, 2020; Hasson, Levy, Behrmann, Hendler, & Malach, 2002; Long, Yu, & Konkle, 2018). These findings suggest that the need to efficiently represent the sensory characteristics of the visual environment places a key constraint on the coding properties of the visual system.

Neural coding of orientation exemplifies the effects of this constraint. In natural images, including both natural and human-made environments, observers tend to encounter edges that are oriented vertically and horizontally (i.e. cardinal orientations) more often compared to edges that

are oriented diagonally (i.e. oblique orientations) (Coppola, Purves, McCoy, & Purves, 1998; Girshick et al., 2011). As predicted from the efficient coding principle, the distribution of preferred orientations for orientation-tuned neurons in visual cortex tends to mirror this distribution, with more neurons exhibiting preferred responses close to cardinal orientations, and fewer preferring oblique orientations (Li, Peterson, & Freeman, 2003; Mansfield, 1974; Shen et al., 2014). At the population level, cardinal orientations have also been shown to be represented in human visual cortex with higher precision than oblique orientations (van Bergen, Ma, Pratte, & Jehee, 2015).

In addition to these neural findings, the idea that the visual system devotes more resources to encoding cardinal orientations is supported by behavioral results. A large body of work suggests that humans, as well as many animal species including primates and cats, are more accurate at perceiving stimuli oriented vertically and horizontally relative to diagonally, a finding which has been termed the "oblique effect" (Appelle, 1972; Bauer, Owens, Thomas, & Held, 1979). The reported benefits include improvements in contrast sensitivity, spatial resolution, ability to discriminate small changes in orientation, and accuracy at reproducing recently seen orientations (Appelle, 1972; Girshick et al., 2011; van Bergen et al., 2015). Interestingly, in addition to these advantages for cardinal orientations, observers also exhibit biases in perception related to cardinal orientations. When presented with orientations that are tilted slightly relative to the cardinals, observers may perceive these stimuli as tilted either toward or away from the nearest cardinal orientation, depending on task and display parameters (Girshick et al., 2011; Wei & Stocker, 2015). These findings can be predicted by various models in which the brain encodes a prior probability distribution over orientation and uses Bayesian inference to combine this prior with noisy sensory evidence, leading to behavioral biases

(Girshick et al., 2011; Wei & Stocker, 2015, 2017). These seemingly sub-optimal behavioral effects support the idea that efficiently encoding highly probable stimulus features leads to a trade-off in performance at other tasks. Together with the neural findings described above, these findings suggest that orientation is represented by the brain in a non-uniform way which is consistent with the distribution of orientations in natural vision.

In **Chapter 1**, I will examine how the distribution of edge orientation in natural images contributes to non-uniformities in how the visual system represents orientation. I will utilize an artificial vision model called a convolutional neural network (CNN) as an experimental system, taking advantage of similarities in architecture and response properties between CNNs and the primate visual system which have been demonstrated previously (Cichy & Kaiser, 2019; Güçlü & van Gerven, 2015; Yamins et al., 2014). Specifically, I will train CNNs (Simonyan & Zisserman, 2014) to perform object categorization on images that are either upright or rotated by a fixed increment relative to upright, and examine how the response properties and orientation representations of the trained network relate to the distribution of orientations in the training set images. One advantage of using an artificial model for this experiment is that it is possible to measure the response properties of every unit in the model simultaneously, using a large number of experimental stimuli, which would not be possible with a biological organism using current neuroscience methods. Additionally, I will be able to directly control the visual stimuli that the network encounters during its training phase, so that the link between visual feature statistics and the response properties of the trained model can be accurately inferred. This approach is conceptually similar to classic experiments using visual deprivation in kittens (Blakemore & Cooper, 1970; Hirsch & Spinelli, 1970; Leventhal & Hirsch, 1975), with the additional advantages that the training images are more complex, and that the network is less likely to have

innate biases toward the cardinal orientations that may be present in newborn animals (Coppola & White, 2004; Hoy & Niell, 2015).

Chapter 1 will demonstrate that when a CNN is trained on upright images (i.e. having the same non-uniform distribution of edge orientations as seen in the natural world), it tends to over-represent the cardinal orientations, as is seen in the brains of primates and other animals. Next, I will demonstrate that when the training images for the CNN are rotated by a fixed increment relative to upright, the non-uniformity in the representations learned by the model shifts by a predictable amount. These findings will demonstrate that general visual experience with non-uniform orientation distributions is sufficient to induce the formation of anisotropic visual representations.

In addition to adaptation of the brain's circuity to the statistics of the environment over long timescales, prioritization of relevant information can also occur more rapidly within the behavior of individual animals. In a visual task, relevant information might be defined based on its ability to automatically capture attention (i.e. salience), its position in the visual field, or the presence of a relevant object feature such as color or shape. At the same time, relevant items might also be defined based on their relationship to previous experiences, such as the resemblance of a currently viewed item to an item that was recently viewed. For instance, when searching for a target item in a cluttered visual scene, the brain needs to simultaneously maintain a "template" representation of the target item while processing a stream of sensory inputs and comparing these to the search template. In **Chapter 2**, I will explore how the brain accomplishes recognition of target items based on relevant properties held in memory.

Identifying target objects based on search templates requires integrating visual input with information held in WM. One mechanism by which such a comparison might be implemented is through top-down projections from areas such as prefrontal cortex (PFC), involved in maintaining information about sought items, to sensory cortical areas involved in analysis of currently viewed items. These feedback projections could provide a biasing signal that selectively enhances representations of items sharing features with the search target (Bichot, 2005; Chelazzi, 2001; Miller, Erickson, & Desimone, 1996). The effect of feedback might also be to modify synaptic weights in sensory cortex, creating a filter that enhances the responses to sensory inputs that match the contents of WM (Mongillo, Barak, & Tsodyks, 2008; Sugase-Miyamoto, Liu, Wiener, Optican, & Richmond, 2008). Finally, to make target information accessible for behavior, representations of target status are likely re-formatted by cortical computations so that they can be read out independently of the identity of viewed and sought items (Pagan, Urban, Wohl, & Rust, 2013). These computations highlight the complex interactions between sensory processing and memory that are required for selection of target items.

At the same time, realistic target identification tasks present several distinct computational challenges. First, relevant objects may be defined based on high-level, abstract properties such as the identity of an object. Under natural viewing conditions, such properties need to be identified in a way that is invariant to low-level image properties, such as variability in illumination, the retinotopic position and size of an item, and the three-dimensional pose of an item. Building invariance to these incidental properties requires multiple stages of computation in the visual system (DiCarlo & Cox, 2007; Rust & DiCarlo, 2010). Second, the properties of an item that are relevant for determining its status as a target might depend on the context in which

it is sought. For instance, when crossing the street, one might be searching for a vehicle pointed in a particular direction but not interested in its identity (i.e. whether it is a bicycle, car, or truck). As a result, identifying items that are relevant targets within a given context also requires the ability to selectively attend to particular item properties, while ignoring irrelevant properties. It is not yet clear how the brain simultaneously solves all these computational challenges.

Areas of the frontal and parietal cortices of the brain are likely to play a role in such complex target recognition tasks. Previous work suggests frontal regions such as PFC are involved in selection of visual information by sending top-down modulatory projections to sensory cortical areas to selectively enhance particular inputs (Buffalo, Fries, Landman, Liang, & Desimone, 2010; Desimone & Duncan, 1995; Martinez-Trujillo & Treue, 2005; Moore & Fallah, 2004). In addition to their role in modulating other areas, neurons within PFC itself represent multiples types of information such as task context and abstract rules, in addition to sensory information, by mixing information in a high-dimensional task space (Rigotti et al., 2013; Wallis, Anderson, & Miller, 2001). These properties are thought to be key for allowing the brain to respond flexibly and adaptively to sensory inputs whose meaning may change depending on context (Duncan & Owen, 2000; Miller & Cohen, 2001). Beyond PFC, adaptive response properties are found within a network of regions in the frontal and parietal cortices of the primate brain which have been termed the multiple-demand (MD) network (Duncan, 2001, 2010; Fedorenko, Duncan, & Kanwisher, 2013). MD network regions have been shown to represent information such as task rules and stimulus response mappings, as well as selectively encode object properties that are relevant for a current task (Jackson, Rich, Williams, & Woolgar, 2017; Woolgar, Thompson, Bor, & Duncan, 2011). Further, MD regions have been shown to represent the behavioral meaning of objects in a way that is invariant to other aspects of their appearance

(Erez & Duncan, 2015; Freedman, Riesenhuber, Poggio, & Miller, 2003; Miller et al., 1996). Thus, these regions are likely to be involved in tasks that require integrating information about sensory inputs with memory and selectively encoding task-relevant variables.

In **Chapter 2**, I will use functional magnetic resonance imaging (fMRI) and a novel object target recognition task to assess the role of MD network regions, as well as visual cortex, in identification of relevant target objects based on high level object properties. While in the fMRI scanner, human subjects will view three-dimensional novel object stimuli while performing a task that requires them to report the status of each item as a match to the previous item according to either its identity or its viewpoint. This task requires subjects to compare high-level properties of the viewed objects with properties held in memory, and to filter out information related to irrelevant object properties. I will show that regions of interest (ROIs) in the MD network encode representations of each item's status as a target in the relevant dimension only, and that these representations are strengthened on trials where the subject correctly identifies the object's target status. In contrast, visual cortex represents match status more weakly and shows no association with task performance. These results support the role of frontal and parietal cortex in selection of relevant objects from the environment.

In addition to prioritizing relevant objects during perception of visual scenes, the brain must also prioritize relevant information in memory. When sensory information must be maintained over short intervals, such as during guidance of eye movements to items in a visual scene, the brain relies on short-term information storage in WM. Storage in WM has a limited capacity and tends to decay over time, introducing an bottleneck that limits the amount of information that can be simultaneously maintained (Luck & Vogel, 2013). As a result, the neural mechanisms for representing information in WM are strongly constrained by the need for

efficient representations that prioritize important information while minimizing resource consumption (Barak & Tsodyks, 2014). This prioritization can take the form of emphasizing information about one visual feature of an item, such as its color or orientation, over another (Serences, Ester, Vogel, & Awh, 2009), or prioritizing information about all the features of a relevant item over a less relevant item (Souza & Oberauer, 2016). It can also influence the format of how items are represented in WM, with representations tending to take a format that will be most efficient and most useful for future behavioral goals.

One proposed mechanism for visual WM involves the utilization of cortical areas typically associated with sensory perception. According to the sensory recruitment model, neural populations in visual cortex that are selective for particular stimulus features during perception can be recruited through top-down feedback to represent the same type of feature information during memory (Awh & Jonides, 2001; Gazzaley & Nobre, 2012; Pasternak & Greenlee, 2005; Serences, 2016; Sreenivasan, Curtis, & D'Esposito, 2014). This mechanism is resource-efficient because it utilizes visual areas whose architecture is already optimized for encoding visual information, rather than forming an independent system for visual memory. Additionally, as described previously in this section, feedback to early sensory areas might involve modification of synaptic weights in these areas without modification of their spiking properties, creating an activity-silent representation that would be more efficient than an active spiking code (Mongillo et al., 2008; Stokes, 2015; Sugase-Miyamoto et al., 2008). In addition to being efficient, the format of representations afforded by a sensory recruitment mechanism would be adaptive for a variety of tasks that require memory for fine details of remembered items, such as discriminating an item from visually similar items.

However, recruitment of sensory populations for feature storage may not be a universally adaptive mechanism. For instance, a challenge introduced by realistic WM tasks is that visual input is constantly entering sensory cortex through feedforward connections, creating representations that may interfere with WM representations stored in these same areas. It has been proposed that information might be diverted to parietal areas such as the intraparietal sulcus (IPS) as a way to confer greater resistance to distracting visual inputs (Bettencourt & Xu, 2015). This problem might also be solved by representing memory inputs within the same retinotopic areas of cortex as sensory input, but within separate cortical layers (Van Kerkoerle, Self, & Roelfsema, 2017). Moreover, certain tasks may allow the contents of memory to be re-formatted in more dramatic ways so that they no longer resemble the initial sensory input. For instance, in a memory guided saccade task, subjects are required to remember the position of a briefly presented dot across a blank delay period, and execute an eye movement to the remembered position. In such a task, the required eye movement is known as soon as the target dot appears, making it possible to plan a response during the delay period. As a result, one strategy for solving this task is to re-map the spatial information about the dot position from its initial retinotopic reference frame into a more action-oriented format that reflects the direction of the saccade the subject plans to make after the delay period (Funahashi, Bruce, & Goldman-Rakic, 1989). Such a representation could be less complex than the initial sensory-like format, and might be more resistant to distraction and/or decay over time. WM representations that are transformed into this type of action-oriented or prospective format might rely more strongly on cortical areas such as PFC and motor cortex and less strongly on sensory cortex (Cisek & Kalaska, 2010; Goldman-Rakic, 1995; Serences, 2016). At the same time, this would likely result in a trade-off where the sensory details of items are represented less precisely.

In **Chapter 3,** I will assess the role of task constraints in shaping the neural mechanisms of WM. Specifically, the factors described above predict that when a subject has the opportunity to plan their motor response during the delay period of a WM task, memory representations will be re-mapped into an action-oriented format that is less reliant on sensory cortex. To evaluate this, I will perform fMRI in human subjects performing a spatial WM task that requires reporting which side of a spatial boundary a remembered position was presented on. To manipulate subjects' ability to plan their motor response, I will present subjects with a preview of the boundary early in the delay period which is either validly predictive or random with respect to the boundary they will ultimately be probed on. I will show that representations of the remembered spatial position in early visual and parietal cortex decline in quality when the subject is shown a valid preview, suggesting a reduced recruitment of sensory cortex for information storage. Furthermore, I will show that this decline in spatial representation strength is accompanied by the emergence of a response representation in primary motor, premotor, and somatosensory cortex prior to the actual execution of the response. This suggests that WM representations can be selectively re-structured according to the demands of particular tasks.

Together, the experiments described in this dissertation will demonstrate how the need for selective information processing shapes the architecture and function of the human visual system. Over slow timescales, adaptation to the orientation statistics of natural images leads to enhanced performance at processing the most common orientations, with a tradeoff in performance at more infrequently-encountered orientations. At more rapid timescales, the brain flexibly modifies how visual stimuli are represented, prioritizing information that is relevant for a given task. This can be seen in the identification of target objects based on integrating sensory inputs with templates held in memory, involving regions of frontoparietal cortex. Finally, storage

of information in memory introduces an additional information bottleneck. This constraint can lead the brain to re-map remembered information from a sensory-like code to a motor-like code, which may provide a means of reducing representational complexity. Taken together, these three experiments provide complementary illustrations of how the visual system's limited processing capacity leads to selective information processing and efficient representations. These properties are essential for the visual system's ability to support coherent perception and flexible decision-making.

References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11239812

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 1–6. https://doi.org/10.1038/s41586-020-2350-5

Barak, O., & Tsodyks, M. (2014, April 1). Working models of working memory. *Current Opinion in Neurobiology*, Vol. 25, pp. 20–24. https://doi.org/10.1016/j.conb.2013.10.008

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication* (pp. 217–234). https://doi.org/10.7551/mitpress/9780262518420.003.0013

Bauer, J. A., Owens, D. A., Thomas, J., & Held, R. (1979). Monkeys Show an Oblique Effect. *Perception*, *8*(3), 247–253. https://doi.org/10.1068/p080247

Bettencourt, K. C., & Xu, Y. (2015). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience*, *19*(1), 150–157. https://doi.org/10.1038/nn.4174

Bichot, N. P. (2005). Parallel and Serial Neural Mechanisms for Visual Search in Macaque Area V4. *Science*, *308*(5721), 529–534. https://doi.org/10.1126/science.1109676

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, *228*(5270), 477–478. https://doi.org/10.1038/228477a0

Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 361–365. https://doi.org/10.1073/pnas.0907658106

Chelazzi, L. (2001). Responses of Neurons in Macaque Area V4 During Memory-guided Visual Search. *Cerebral Cortex*, *11*(8), 761–772. https://doi.org/10.1093/cercor/11.8.761

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, Vol. 33, pp. 269–298. https://doi.org/10.1146/annurev.neuro.051508.135409

Coppola, D. M., Purves, H. R., McCoy, A. N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 4002–4006. https://doi.org/10.1073/pnas.95.7.4002

Coppola, D. M., & White, L. E. (2004). Visual experience promotes the isotropic representation of orientation preference. *Visual Neuroscience*, *21*(1), 39–51. https://doi.org/10.1017/s0952523804041045

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *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

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475–483. https://doi.org/10.1016/S0166-2236(00)01633-7

Erez, Y., & Duncan, J. (2015). Discrimination of Visual Categories Based on Behavioral Relevance in Widespread Regions of Frontoparietal Cortex. *Journal of Neuroscience*, *35*(36), 12383–12393. https://doi.org/10.1523/JNEUROSCI.1134-15.2015

Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*(23), 787–792. Retrieved from papers3://publication/uuid/5CE57789-3A77-4C34-8D94-3FDCD2C095C4

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*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(12), 5235–5246. https://doi.org/23/12/5235 [pii]

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1016/j.neuron.2012.12.039

Gazzaley, A., & Nobre, A. C. (2012, February 1). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, Vol. 16, pp. 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Gervain, J., & Geffen, M. N. (2019, January 1). Efficient Neural Coding in Auditory and Speech Perception. *Trends in Neurosciences*, Vol. 42, pp. 56–65. https://doi.org/10.1016/j.tins.2018.09.004

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

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

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, *34*(3), 479–490. https://doi.org/10.1016/S0896-6273(02)00662-1

Hirsch, H. V. B., & Spinelli, D. N. (1970). Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science*, *168*(3933), 869–871. https://doi.org/10.1126/science.168.3933.869

Hoy, J. L., & Niell, C. M. (2015). Layer-specific refinement of visual cortex function after eye opening in the awake mouse. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(8), 3370–3383. https://doi.org/10.1523/JNEUROSCI.3174-14.2015

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

Leventhal, A. G., & Hirsch, H. V. B. (1975). Cortical effect of early selective exposure to diagonal lines. *Science*, *190*(4217), 902–904. https://doi.org/10.1126/science.1188371

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Luck, S. J., & Vogel, E. K. (2013, August 1). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, Vol. 17, pp. 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Mansfield, R. J. W. (1974). Neural basis of orientation perception in primate vision. *Science*, *186*(4169), 1133–1135. https://doi.org/10.1126/science.186.4169.1133

Martinez-Trujillo, J. C., & Treue, S. (2005). The feature similarity gain model of attention: Unifying multiplicative effects of spatial and feature-based attention. *Neurobiology of Attention*, 300–304. https://doi.org/10.1016/B978-012375731-9/50053-7

Miller, E., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, E., Erickson, C. a, & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(16), 5154–5167. https://doi.org/10.1.1.41.2959

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science (New York, N.Y.)*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Moore, T., & Fallah, M. (2004). Microstimulation of the Frontal Eye Field and Its Effects on Covert Spatial Attention. *Journal of Neurophysiology*, *91*(1), 152–162. https://doi.org/10.1152/jn.00741.2002

Pagan, M., Urban, L. S., Wohl, M. P., & Rust, N. C. (2013). Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature Neuroscience*, *16*(8), 1132–1139. https://doi.org/10.1038/nn.3433

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013).

The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. https://doi.org/10.1523/JNEUROSCI.0179-10.2010

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Shen, G., Tao, X., Zhang, B., Smith, E. L., Chino, Y. M., & Chino, Y. M. (2014). Oblique effect in visual area 2 of macaque monkeys. *Journal of Vision*, *14*(2). https://doi.org/10.1167/14.2.3

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, Vol. 24, pp. 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Retrieved from http://arxiv.org/abs/1409.1556

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, and Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Computational Biology*, *4*(5). https://doi.org/10.1371/journal.pcbi.1000073

Tesileanu, T., Cocco, S., Monasson, R., & Balasubramanian, V. (2019). Adaptation of olfactory receptor abundances for efficient coding. *ELife*, *8*. https://doi.org/10.7554/eLife.39279

van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

Van Kerkoerle, T., Self, M. W., & Roelfsema, P. R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature Communications*, *8*(1), 1–14. https://doi.org/10.1038/ncomms13804

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953–956. https://doi.org/10.1038/35082081

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 201619153. https://doi.org/10.1073/pnas.1619153114

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn858

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752. https://doi.org/10.1016/j.neuroimage.2010.04.035

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

CHAPTER 1: Biased orientation representations can be explained by experience with non-uniform training set statistics

Abstract

Visual acuity is better for vertical and horizontal compared to other orientations. This cross-species phenomenon is often explained by "efficient coding", whereby more neurons show sharper tuning for the orientations most common in natural vision. However, it is unclear if experience alone can account for such biases. Here, we measured orientation representations in a convolutional neural network, VGG-16, trained on modified versions of ImageNet (rotated by 0°, 22.5°, or 45° counter-clockwise of upright). Discriminability for each model was highest near the orientations that were most common in the network's training set. Furthermore, there was an over-representation of narrowly tuned units selective for the most common orientations. These effects emerged in middle layers and increased with depth in the network. Our results suggest that biased orientation representations can emerge through experience with a non-uniform distribution of orientations, supporting the efficient coding hypothesis.

Introduction

Contrary to common intuition, visual perception is not perfectly uniform across orientation space. One example of this principle is the "oblique effect", which has been demonstrated in humans and a wide range of animal species, including cats, octopi and goldfish, among others. This effect describes the finding that observers' ability to discriminate small changes in orientation, as well as other forms of acuity, tend to be worst for stimuli that have edges oriented diagonally (oblique orientations) and better for stimuli with edges oriented vertically or horizontally (cardinal orientations) (Appelle, 1972; Bauer et al., 1979). In visual cortex, this finding has been linked to a larger number of orientation tuned neurons with a preference for cardinal orientations, as has been shown in cats (Li et al., 2003), and macaques (Mansfield, 1974; Shen et al., 2014), among other species. Some evidence also suggests that

cardinally-tuned neurons may have narrower tuning than other orientations, which may also contribute to higher acuity (Kreile et al., 2011; Li et al., 2003).

One compelling explanation for the origin of the oblique effect is the efficient coding hypothesis, which suggests that because the brain operates with limited resources, coding resources should be preferentially allocated to stimuli that are highly probable during natural vision (Barlow, 1961; Girshick et al., 2011). On this view, biased orientation perception may reflect an adaptation to the statistics of natural images, in which vertical and horizontal orientations are most common (Coppola et al., 1998; Girshick et al., 2011). Support for an experience-driven account of the oblique effect includes evidence that in primates, the over-representation of cardinal orientations in visual cortex increases with age (Shen et al., 2014). Additionally, exposing developing kittens or mice to an environment with contours of only one orientation can induce changes in the distribution of cortical orientation tuning, suggesting some degree of plasticity (Blakemore & Cooper, 1970; Hirsch & Spinelli, 1970; Kreile et al., 2011; Leventhal & Hirsch, 1975).

In addition, innate factors may also contribute to the efficient coding of cardinal orientation. For instance, while it is possible to significantly modify the distribution of orientation tuning preferences in visual cortex through experience, exposing an animal to only diagonal lines during development does not entirely obliterate tuning for cardinal orientations (Kreile et al., 2011; Leventhal & Hirsch, 1975). Similarly, rearing animals in complete darkness can result in a more extreme over-representation of cardinal-tuned units (Leventhal & Hirsch, 1980). In both mice and ferrets, it has been suggested that innate factors result in a strong oblique effect early in development, while visual experience tends to make orientation tuning more uniform over time (Coppola & White, 2004; Hoy & Niell, 2015). These observations are

consistent with the efficient coding account if we assume that the visual system can adapt to environmental regularities over the course of evolution, resulting in feature biases that are encoded in the genome.

However, factors that are independent of visual input statistics may also separately contribute to the presence of cardinal orientation biases in animals. For example, some anatomical properties of the visual system naturally give a privileged status to the cardinal axes, such as the horizontal raphe of the retina, the role of the horizontal axis in vestibular and oculomotor system organization, and the distinction between processing of vertical and horizontal disparity (Westheimer, 2003). Such properties need not be related to the orientation content of natural images, but may instead reflect general physical and/or developmental constraints. It is plausible that the presence of these architectural factors leads to cardinal biases, independent from the statistics of natural images. Thus, whether the efficient coding mechanism alone can account for the emergence of the oblique effect has not been clearly established.

Here, we addressed this question by examining whether a convolutional neural network (CNN) exhibits biased orientation representations. We focus on the popular VGG-16 model, a standard feedforward network that achieves high performance at classifying objects in natural images (Simonyan & Zisserman, 2014). We first test whether a pre-trained VGG-16 model exhibits the classical oblique effect, assessed using the Fisher information measured at entire layers of the network, and the distribution of single-unit tuning properties. In addition to a test of the efficient coding hypothesis, measuring orientation bias in this pre-trained model will provide an assessment of whether existing CNNs, often used as models of the primate visual system (Cichy & Kaiser, 2019; Kell & McDermott, 2019), exhibit this defining characteristic of biological vision.

We next trained VGG-16 models on modified versions of the ImageNet database (Deng et al., 2009) that had been rotated by 0°, 22.5° or 45° relative to upright. This allowed us to determine whether a bias centered around other axes can be induced equally as well as a cardinal bias, and whether the biases observed in the pre-trained network were simply artifacts of some intrinsic property of the CNN (e.g. a square pixel grid that results in a cardinal reference frame). We demonstrate that, contrary to this alternative, networks trained on rotated images exhibited rotated biases that were consistent with the networks' training set statistics. These results suggest that general visual experience with a non-uniform orientation distribution is sufficient to promote the formation of biased orientation representations. Further, our findings highlight how biased training data can fundamentally impact visual information processing in neural network models.

Materials and Methods

Training stimuli

During training, each model was presented with a modified version of the ILSVRC-2012-CLS training image set, a set of ~1.3 million colored images with substantial variability in layout and background, each including an object in one of 1,000 categories (Deng et al., 2009; Russakovsky et al., 2015). Three modified versions of this image set were generated, corresponding to rotations of 0°, 22.5°, and 45° counter-clockwise relative to vertical. The purpose of generating a 0° (no rotation) version of the image set was to provide a control to isolate the effect of image rotation from any other properties of our modified image set.

To generate each version of the image set, we loaded each image from the original ILSVRC image set, rotated it by the specified amount, and cropped the image centrally by a specified amount that was the same for all rotations. Images were then scaled to a size of [224 x

224] pixels, and multiplied by a smoothed circular mask. The smoothed mask set to background all pixels at a radius of more than 100 pixels from the center, retained all pixels at a radius of less than 50 pixels from the center, and applied a cosine function to fade out the intermediate pixels. Finally, the background pixels were adjusted to a grey color that closely matches the mean RGB value of the training ImageNet images (Simonyan & Zisserman, 2014). All image processing for training set images was done in Python 3.6 (Python Software Foundation, Wilmington DE) using the Python Imaging Library. For each training set, a corresponding validation set was generated using the same procedure, and this validation set was used to evaluate performance during training. When preprocessing the images for training and validation, we modified the procedure from Simonyan and Zisserman's paper by skipping the random rescaling and random left-right flipping steps. The purpose of this was to preserve the original spatial frequency and orientation content of the images as closely as possible.

Evaluation stimuli

Networks were evaluated using sets of images that had known orientation content. To generate these image sets, we randomly sampled images from the ILSRVC-2012-CLS image set and filtered them to have a particular orientation content. Before filtering each image, we first rotated it by a randomly chosen value in the range of 0-179 degrees, then cropped it centrally to a square and scaled to a size of [224 x 224] as described above. This was done to prevent any dependencies between orientation and other low-level image properties, such as spatial frequency content and luminance contrast, in the final filtered images. After this step, we converted to grayscale, z-scored the resulting luminance values, and masked the image with the smoothed circular mask described above. The image was then padded with zeros to a size of [1012 x 1012] pixels, and transformed into the frequency domain. We multiplied the frequency-

domain representation by an orientation filter consisting of a circular Gaussian (Von Mises) function centered at the desired orientation (k=35) and a bandpass spatial frequency filter with Gaussian smoothed edges (0.02 to 0.25 cycles/pixel, SD=0.005 cycles/pixel). We then replaced the image's phase with random values uniformly sampled between -pi to +pi, and transformed back into the spatial domain. Next, we cropped the image back to its original size of [224 x 224], multiplied again by the smoothed circular mask, and converted the image into a 3-channel RGB format. Finally, the luminance in each color channel was normalized to have a mean equal to the mean of that color channel in the training ImageNet images and a standard deviation of 12 units. All image processing for the evaluation image sets was done using Matlab R2018b (MathWorks, Natick MA).

Using the above procedures, we generated four evaluation image sets, each starting with a different random set of ImageNet images. Images in each evaluation set had orientations that varied between 0° and 179°, in steps of 1°, resulting in 180 discrete orientation values. Throughout this paper, we use the convention of 0° for vertical and 90° for horizontal orientations, with positive rotations referring to the clockwise direction, and negative rotations referring to the counter-clockwise direction. Each evaluation set included 48 examples of each orientation, for a total of 8640 images per set.

Measuring image set statistics

To verify that the modified versions of the ImageNet images had the anisotropic orientation statistics that we expected, we measured the orientation content of each training image using a Gabor filter bank. The filter bank included filters at orientations from 0° to 175° in 5° steps, at spatial frequencies of 0.0200, 0.0431, 0.0928, and 0.200 cycles per pixel. The filter bank was generated using the *gabor* function in Matlab R2018b (MathWorks, Natick MA).

Before filtering each image, we converted each image to grayscale, and subtracted its background color so that the background was equal to zero. Filtering was done in the frequency domain. After converting back to the spatial domain, we averaged the magnitude of the filtered image across all pixel values to obtain a single value for each filter orientation and spatial frequency. To visualize the distribution of orientation content across all images, we z-scored the magnitude values across the orientation dimension, averaged over spatial frequency, and divided the resulting values by their sum to estimate the probability distribution over orientation. This analysis was done on the training set images only, which included ~1300 images in each of 1000 categories, for a total of ~1.3 million images.

Network training and evaluation

We trained VGG-16 networks (Simonyan & Zisserman, 2014) on three different modified versions of the ImageNet dataset (see *Training stimuli* for details). For each of the three image sets, we initialized and trained four VGG-16 networks (replicates), giving a total of 12 models. All models were trained using Tensorflow 1.12.0 (Abadi et al., 2016), using the TF-slim model library (Silberman & Guadarrama, 2016) and Python 3.6 (Python Software Foundation, Wilmington DE). All models were trained using the RMSProp algorithm with momentum of 0.80 and decay of 0.90. The learning rate was 0.005 with an exponential decay factor of 0.94, and the weight decay parameter was 0.0005. Networks were trained until performance on the validation set (categorization accuracy and top-5 recall) began to plateau, which generally occurred after around 350K-400K steps. The validation images used to evaluate performance were always rotated in an identical manner to the training set images. Training was performed on an NVIDIA Quadro P6000 GPU (NVIDIA, Santa Clara CA). All evaluation was performed using the first checkpoint saved after reaching 400K steps. As noted above, we did not perform

data augmentation steps during image pre-processing for training. Removing these procedures may have contributed to the relatively low final classification performance that we observed (top-5 recall accuracy ~60%).

To measure activations from each trained network, we split the evaluation image sets (consisting of 8640 images each) into 96 batches of 90 each. We then passed each batch through each trained network and measured the resulting activations of each unit as the output of the activation function (a rectified linear operation). We saved the activations for each unit in each layer for all images, which were then submitted to further analysis. We performed this evaluation procedure on a total of 17 networks: the 12 models trained on modified ImageNet images, a pre-trained VGG-16 network from the TF-slim model library (Silberman & Guadarrama, 2016), and four randomly initialized VGG-16 models that served as a control. All subsequent analyses were performed using Python 3.6 (Python Software Foundation, Wilmington DE).

Computing Fisher information (FI)

To measure the ability of each network layer to discriminate small changes in orientation, we estimated Fisher information (FI) as a function of orientation. To estimate FI for each network layer, we first computed FI for each unit in that layer, then combined information across units. FI for each unit was computed based on the slope and variance of that unit's activation at each point in orientation space, according to the following relation:

$$FI_i(\theta) = \frac{\left(\frac{\partial f_i(\theta)}{d\theta}\right)^2}{v_i(\theta)}$$

Where $f_i(\theta)$ is the unit's measured orientation tuning curve, and $v_i(\theta)$ is the variance of the unit's responses to the specified orientation. We estimated the slope of the unit's tuning curve at

θ based on the difference in its mean response (μ_i) to sets of images that were Δ=4° apart (using different values of Δ did not substantially change the results).

$$\left(\frac{\partial f_i(\theta)}{d\theta}\right) \cong \frac{\mu_i(\theta_1) - \mu_i(\theta_2)}{\Delta}$$

Where

$$\theta_1 = \theta - \frac{\Delta}{2}$$

$$\theta_2 = \theta + \frac{\Delta}{2}$$

We presented an equal number of images (48) at each orientation, so the pooled variance was calculated as:

$$v_i(\theta) = \frac{v_i(\theta_1) + v_i(\theta_2)}{2}$$

Finally, we summed this measure across units of each layer to obtain a population level estimate of FI.

$$FI_{pop}(\theta) = \sum_{i=0}^{nUnits} FI_i(\theta)$$

Where $nUnits$ is the number of units in the layer. We computed $FI_{pop}(\theta)$ for theta values between 0° and 179°, in steps of 1°. When plotting FI, to aid comparison of this measure across

layers with different numbers of units, we divided FI_{pop} by the total number of units in the layer, to capture the average FI per unit.

Fisher information bias (FIB)

To quantify the amount of bias (non-uniformity) in Fisher information at each layer of the network, we computed a measure which we refer to as the Fisher information bias (FIB). For the pre-trained model and the networks trained on upright images, we expected the network to over-represent cardinal orientations, showing peaks in FI around vertical and horizontal. However, the models trained on rotated images were expected to show peaks rotated by a specified amount relative to the cardinal orientations. To account for these different types of bias, we computed three versions of the FIB: one that measures the height of peaks in FI around the cardinal orientations (FIB-0), one that measures the height of peaks in FI that are 22.5° counter-clockwise of the cardinals (FIB-22), and one that measures the height of peaks in FI that are 45° counter-clockwise of the cardinals (FIB-45), relative to a baseline. The equation for each FIB measure is as follows:

$$FIB = \frac{FI_{peaks} - FI_{baseline}}{FI_{peaks} + FI_{baseline}}$$

Where FI_{peaks} is the sum of the FI values in a range $\pm$10° around the orientations of interest (0° and 90° for FIB-0, 67.5° and 157.5° for FIB-22, and 45° and 135° for FIB-45), and $FI_{baseline}$ is the sum of the FI values in a range $\pm$10° around the orientation chosen as a baseline (22.5° and 112.5°). Since FI is necessarily positive, each of these FIB measures can take a value between +1 and -1, with positive values indicating more information near the orientations of interest relative

to the baseline (peaks in FI), and negative values indicating less information near the orientations of interest relative to baseline (dips in FI).

To test whether FIB differed significantly between trained models and the randomly initialized (not trained) models, we performed t-tests between FIB values corresponding to each training set and the random models. Specifically, we tested the hypothesis that the primary form of bias measured in models corresponding to each training set (e.g. FIB-0 for the models trained on upright images, FIB-22 for the models trained on 22.5° rotated images, FIB-45 for the models trained on 45° rotated images) was significantly higher for the models trained on that image set than for the random (not trained) models. Since we generated four replicate models for each training image set, and evaluated each model on four evaluation image sets, there were 16 total FIB values at each layer corresponding to each training set. All tests were implemented as one-tailed t-tests using SciPy (version 1.1.0), assuming unequal variance. The p-values were FDR corrected across model layers at q=0.01 (Benjamini & Yekutieli, 2001). The same procedure was used to test for differences in FIB-0 between the pre-trained model and the control model (note that there was only one replicate for the pre-trained model, so this test included only 4 data points).

Single-unit tuning analysis

To measure the extent to which training set statistics impacted the orientation tuning of individual units in each network, we measured tuning functions based on each unit's responses to the evaluation image set, and we performed curve fitting to quantify tuning properties. First, we measured an orientation tuning function for each unit at each layer of the model by averaging its responses to all evaluation set images that had the same orientation (in each image set, there were 48 images at each of 180 orientations). Any units that had a constant response across all

images or a zero response to all images were removed at this stage (this included mainly units whose spatial selectivity was outside the range stimulated by the circular image aperture, around 35% of units per layer at the earliest layers). We computed and saved an orientation tuning curve for each unit in response to each of the four evaluation image sets. We then averaged over these four evaluation sets before fitting.

To characterize the tuning curves, we fit each with a circular Gaussian (Von Mises) function, having the basic form:

$$v(\theta) = e^{(k*\cos(\theta-u)-1)}$$

Where u is a parameter that describes the center of the unit's tuning function, and k is a concentration parameter that is inversely related to the width of the tuning function. In this formulation, the k parameter modifies both the height and the width of the tuning function. To make it possible to modify the curve's height and width independently, we normalized the Von Mises function to have a height of 1 and a baseline of 0, and then added parameters for the amplitude and baseline, as follows:

$$f(\theta) = b + a * v_n(\theta)$$

Where $v_n(\theta)$ denotes the Von Mises function after normalization. This resulted in a curve with four total parameters: center, size, amplitude, and baseline.

We fit a curve of this form to each unit's average tuning function using linear least-squares regression, implemented with the optimization library in SciPy (version 1.1.0). To

initialize the fitting procedure, we used the argmax of the tuning function as an estimate of its mean, the minimum value as an estimate of its baseline, and the range as an estimate of its amplitude. The concentration parameter k was always initialized at 1. Values for the center were constrained to lie within the range of [-0.0001, 180.0001], k was constrained to positive values $>10^{-15}$, and amplitude and baseline were allowed to vary freely. To prevent any bias in the center estimates due to the edges of the allowed parameter range, we circularly shifted each curve by a random amount before fitting.

After fitting was complete, we assessed the goodness of the fit using R^2. To assess the consistency of tuning across different versions of the evaluation image set, we used R^2 to assess the fit between the single best-fit Von Mises function (computed using the tuning function averaged over all evaluation image sets) and each individual tuning curve (there were four individual tuning curves, each from one version of the evaluation image set). We then averaged these four R^2 values to get a single value. We used a threshold of average $R^2 > 0.40$ to determine which units were sufficiently well-fit by the Von Mises function, and retained the parameters of those fits for further analysis.

Results

We measured the activation responses of several trained VGG-16 networks (Figure 1.1A) (Simonyan & Zisserman, 2014) presented with oriented images (Figure 1.1B) to evaluate whether each network showed non-uniformities in its orientation representations across feature space. First, we tested whether a pre-trained VGG-16 model (Silberman & Guadarrama, 2016) exhibits the classical oblique effect. Next, we evaluated whether this bias changed in a

predictable way when networks with the same architecture were trained on modified versions of the ImageNet database (Figure 1.4A).

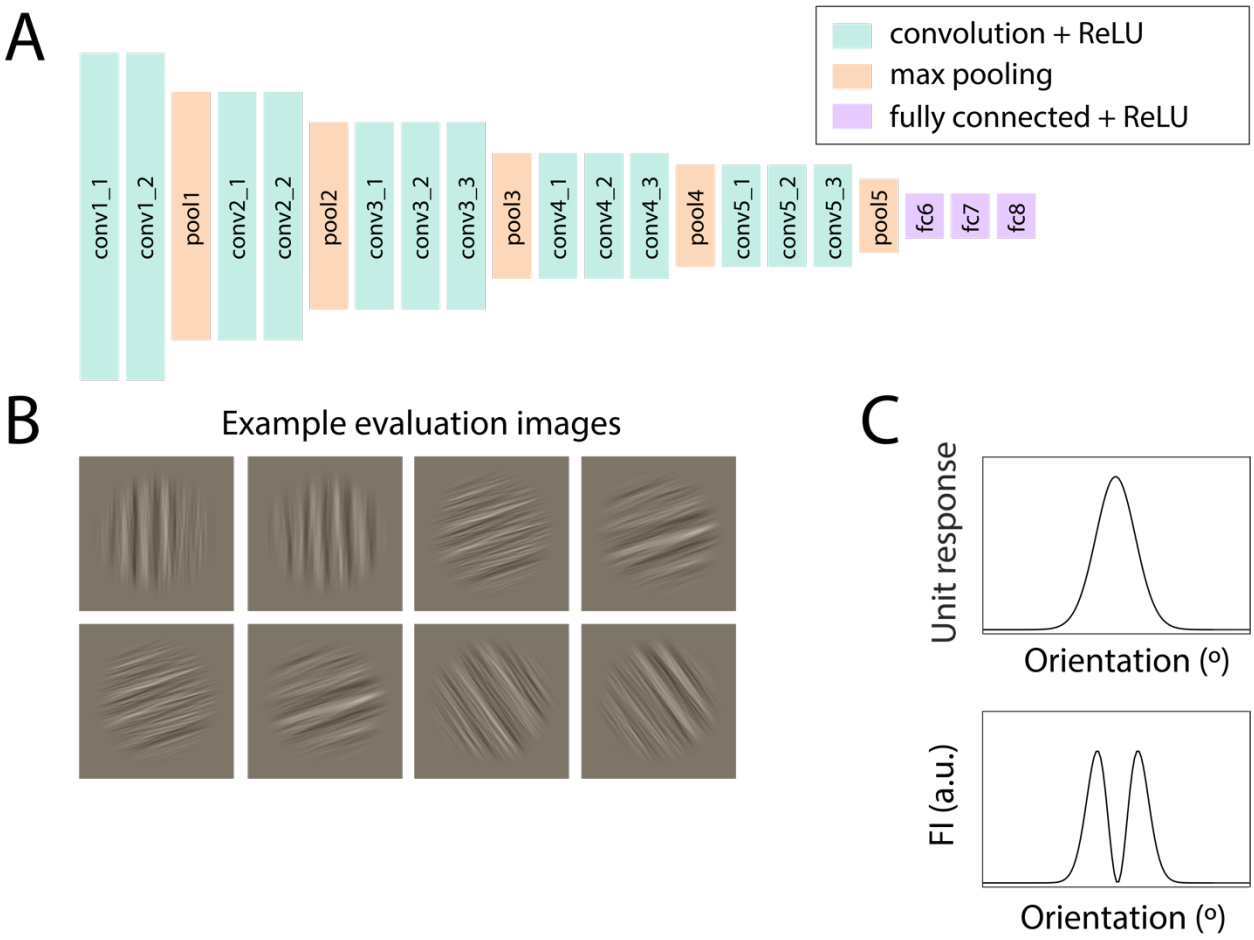


Figure 1.1. Evaluating orientation discriminability in a trained neural network model. **(A)** Schematic of the VGG-16 network architecture, with layers arranged from shallowest (left) to deepest. **(B)** Examples of oriented images used to measure orientation representations in the pre-trained network. Images were generated by filtering ImageNet images within a narrow orientation range, preserving their broadband spatial frequency content. Orientations varied between 0-179°, in steps of 1° (see *Methods, Evaluation stimuli)*. **(C)** Cartoon depiction of the approximate relationship between an example single unit tuning function and the Fisher information (FI) measured from that unit as a function of orientation.

Measuring cardinal biases in a pre-trained VGG-16 model

We first evaluated non-uniformity at the level of each pre-trained network layer by computing the layer-wise Fisher information (FI), which reflects how well each layer's activations can distinguish small changes in orientation (see *Methods, Computing Fisher information*). Briefly, the contribution of each network unit to the layer-wise FI is the squared slope of a unit's tuning function at each orientation normalized by the variance of the response at that orientation. Thus, the steep part of a unit's tuning function will carry more information because two similar orientations will evoke different responses (Figure 1.1C). However, the flat parts of a unit's tuning curve (i.e. at the peak or in the tails) will not carry very much information because the unit will respond about the same to two similar orientations.

For a pre-trained VGG-16 model, plotting FI as a function of orientation reveals noticeable deviations from uniformity, particularly at deep layers of the network (navy blue curves in Figure 1.2A). While the first layer of the model (conv1_1), gives a relatively flat profile of FI with respect to orientation, by layer 7 (conv3_1), peaks in FI are apparent around the cardinal orientations, 0°/180° and 90°. At later layers of the model, the peaks in FI are more pronounced and begin to take on a characteristic double-peaked shape, where FI is maximal just a few degrees to the left and right of the cardinal orientations, with a dip at the exact position of the cardinal orientations (this shape is discussed in more detail in the next section after we report statistics about the preferred orientation and width of single unit tuning functions). In contrast, when the same analysis is done on a randomly initialized VGG-16 model (no training performed), FI is flat with respect to orientation at all layers, suggesting that a randomly initialized model does not exhibit this same cardinal bias (gray curves in Figure 1.2A).

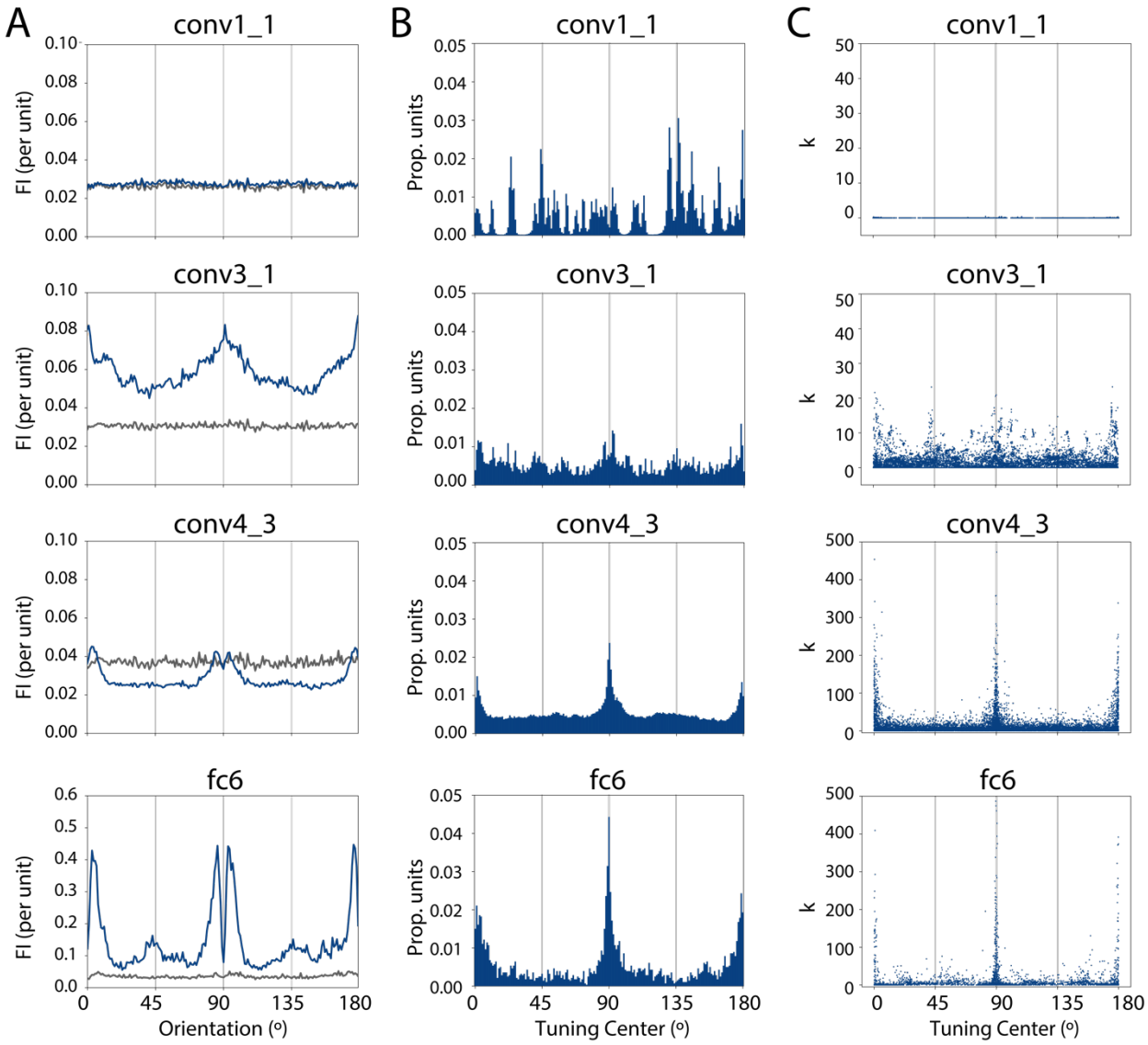


Figure 1.2. Pre-trained VGG-16 shows maximum orientation information just off of cardinal orientations, and non-uniformity in the distribution of single unit tuning properties. **(A)** FI is plotted as a function of orientation for several example layers of the pre-trained model (navy blue) and a randomly initialized model (gray). See *Methods, Computing Fisher information* for details. **(B)** Distribution of the tuning centers of pre-trained network units that were well-fit by a Von Mises function. See Supplementary Figure 1.1 for the proportion of well-fit units per layer, and the distribution of centers for the randomly initialized model. **(C)** Concentration parameter (k) versus center for individual units in the pre-trained model (data in the top three panels of C have been down-sampled to a maximum of 10,000 points for visualization purposes).

To quantify this effect at each layer, we computed a metric which we term the Fisher Information Bias (FIB), which captures the relative height of the peaks in FI compared to a baseline (see *Methods, Fisher information bias*). We defined three versions of this metric, the FIB-0, FIB-22, and FIB-45, which denote the height of peaks in FI around the cardinal orientations, around orientations 22.5° counter-clockwise of cardinals, and around orientations 45° counter-clockwise of cardinals, respectively. For example, to get the FIB-0, we take the mean FI in 20° bins around 0° and 90°, subtract the mean FI in in a baseline orientation range, and divide by the sum of these two means. Because the pre-trained model showed peaks in FI around cardinals only, we focus on the FIB-0 in this section; the FIB-22 and FIB-45 are discussed in the following section (*Training networks on rotated images*). We found that for the pre-trained model, the FIB-0 increased with depth in the network, showing values close to zero for the first four layers, then showing positive values that increase continuously at each layer (navy blue line in Figure 1.3). In contrast, we found less evidence for a cardinal bias in the randomly initialized model, shown by smaller values of the FIB-0 at all layers (gray line in Figure 1.3). The difference in FIB-0 between the pre-trained and randomly initialized models was significant starting at the fifth layer (conv2_2), and at all layers deeper than conv2_2 (one-tailed t-test, FDR corrected q=0.01). However, there was a small increase in the FIB-0 at the later layers of the randomly initialized model, reflecting a weak cardinal bias (at the deepest layer, the FIB-0 was still more than 5x as large for the pre-trained model as for the random model). We return to this issue for more consideration in the *Discussion*.

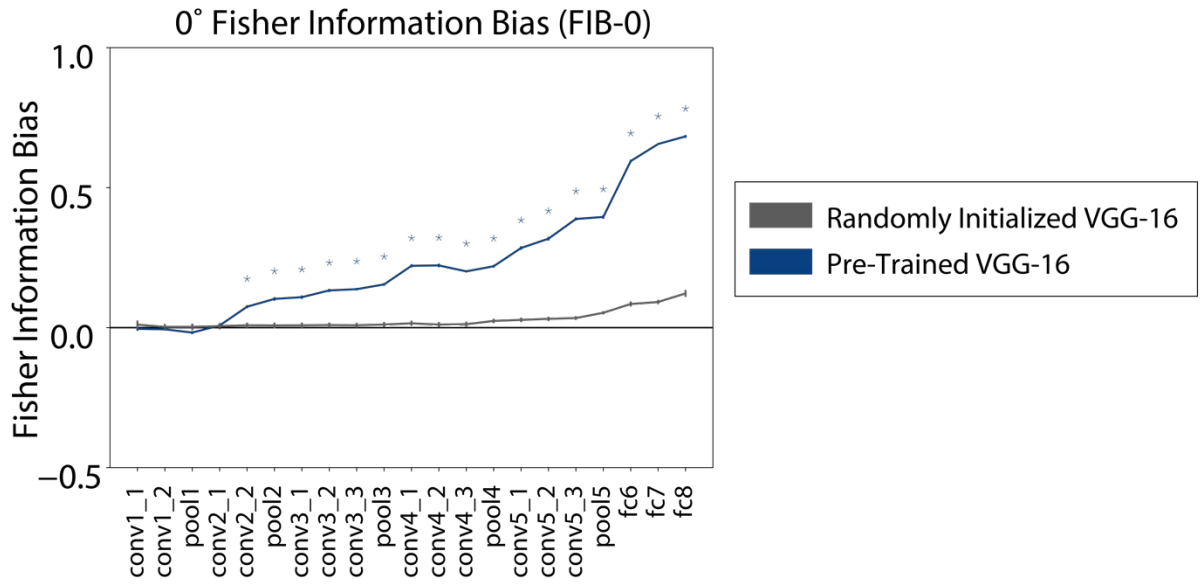


Figure 1.3. Cardinal bias in a pre-trained VGG-16 model increases with depth. FIB-0, a measure of cardinal information bias (see *Methods, Fisher information bias)*, plotted for a pre-trained model (navy blue) and a randomly initialized control model (gray), with asterisks indicating layers for which the pre-trained model had significantly higher FIB-0 than the random model (one-tailed t-test, FDR corrected q=0.01). Error bars reflect standard deviation across four evaluation image sets.

Having demonstrated that a pre-trained CNN exhibits an advantage for discriminating cardinal versus other orientations, we were next interested in whether this bias was linked to the distribution of tuning properties across single units at each layer, as has been observed in the brains of animals such as cats and macaques (Li et al., 2003; Shen et al., 2014; Vogels & Orban, 1994). To investigate this, we computed the average orientation tuning profiles for individual units in response to stimuli of all orientations and fit these profiles with Von Mises functions to estimate their center and concentration parameter (or width, denoted *k*). Units that were not well-fit by a Von Mises were not considered further (approximately 30% of all units, see *Methods, Single-unit tuning analysis* and Supplementary Figure 1.1. Figure 1.2B shows the distribution of fit centers for all units in four example layers of the pre-trained model that were well-fit by a

Von Mises function. These distributions show peaks at random locations for the first layer of the network, but exhibit narrow peaks around the cardinal orientations for the deeper conv4_3 and fc6 layers. In contrast, the randomly initialized model did not show an over-representation of cardinal-tuned units (Supplementary Figure 1.1). In addition, plotting the concentration parameter for each unit versus the center (Figure 1.2C) shows that for the deepest three layers shown, the most narrowly-tuned units (high k) generally have centers close to the cardinal orientations. Together, these findings indicate that middle and deep layers of the pre-trained network have a large proportion of units tuned to cardinal orientations, and that many of these units are narrowly tuned.

These findings may provide an explanation for the double-peaked shape of the FI curves for the pre-trained model at deep layers (Figure 1.2A). Since FI is related to the slope of a unit's tuning function, it is expected to take its maximum value on the flanks of a tuning curve, where slope is highest, and take a value of zero at the tuning curve peak (Figure 1.1C). Thus, having a large number of narrowly-tuned units with their peaks precisely at 0° and 90° could result in layer-wise FI having local maxima at the orientations just off of the cardinals.

Training networks on rotated images

Having demonstrated that a pre-trained VGG-16 network exhibits a much stronger cardinal orientation bias compared to a randomly initialized network, we next tested whether training a model on rotated images would result in rotated biases. This test is needed to demonstrate that the frequently-observed cardinal bias is not the only possible orientation bias that can be induced in a visual system through exposure to a set of images with non-uniform statistics. We trained networks on three modified versions of the ImageNet dataset (Deng et al., 2009), consisting of images that were rotated by either 0°, 22.5°, or 45° in a clockwise direction

relative to the upright position (Figure 1.4A). Separately, we also verified that the image statistics of each of the modified sets exhibited the expected distribution, such that vertical and horizontal orientations were most common in the upright training set, orientations 22.5° counter-clockwise of cardinals were most common in the -22.5° rotated set, and orientations 45° counter-clockwise of cardinals were most common in the -45° rotated set (Figure 1.4B).

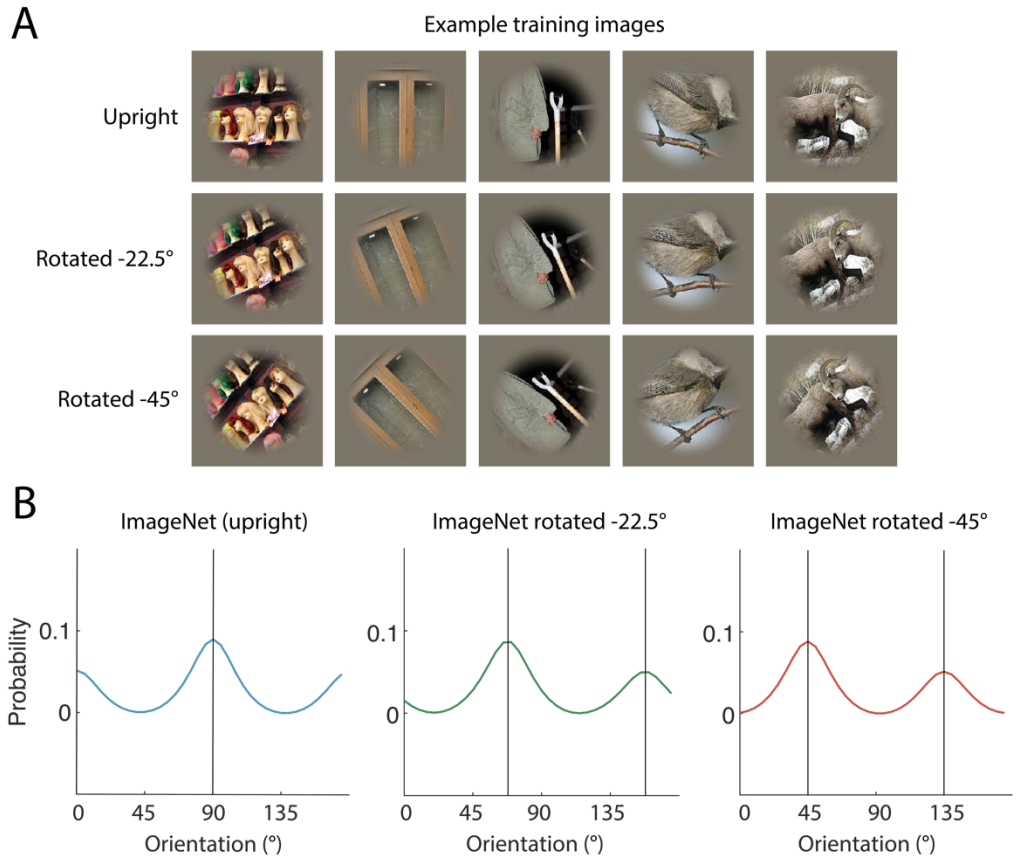


Figure 1.4. Rotated images used to train VGG-16 networks. **(A)** Separate networks were trained on either upright or rotated versions of the ImageNet image set, with a smoothed circular mask applied to remove vertical and horizontal image edges. **(B)** Orientation content from images in each of the training sets in (A) was measured using a Gabor filter bank (see *Methods, Measuring image set statistics*).

Our results indicate that training on rotated images shifted the orientation bias by a predictable amount. FI for the models that were trained on upright images shows a relatively similar shape to the pre-trained model, with peaks appearing at a few degrees to the left and right of the cardinal orientations (Figure 1.5A). This demonstrates that though our training procedure and image set were not identical to those used for the pre-trained model, they resulted in the formation of similar orientation biases. In contrast, the models trained on rotated images each showed a FI curve that was similar in shape but shifted relative to the curve from the model

trained on upright images, such that the peaks in FI were always near the orientations that were most common in the training set images (Figure 1.5D,1.5G).

The distribution of single-unit tuning properties also shifted with training set statistics. In the upright-trained model, the highest proportion of units had their tuning near the cardinals, while the networks trained on 22.5° and 45° rotated images had more units with tuning at either 22.5° or 45° counter-clockwise relative to the cardinal orientations, respectively (Figure 1.5B,1.5E,1.5H). Additionally, for all models, the most narrowly-tuned units tended to be those that were tuned to the orientations most common in the training set (Figure 1.5C,1.5F,1.5I). As described above, the high number of narrowly-tuned units with their centers close to these most common orientations may underly the double-peaked shape seen in FI.

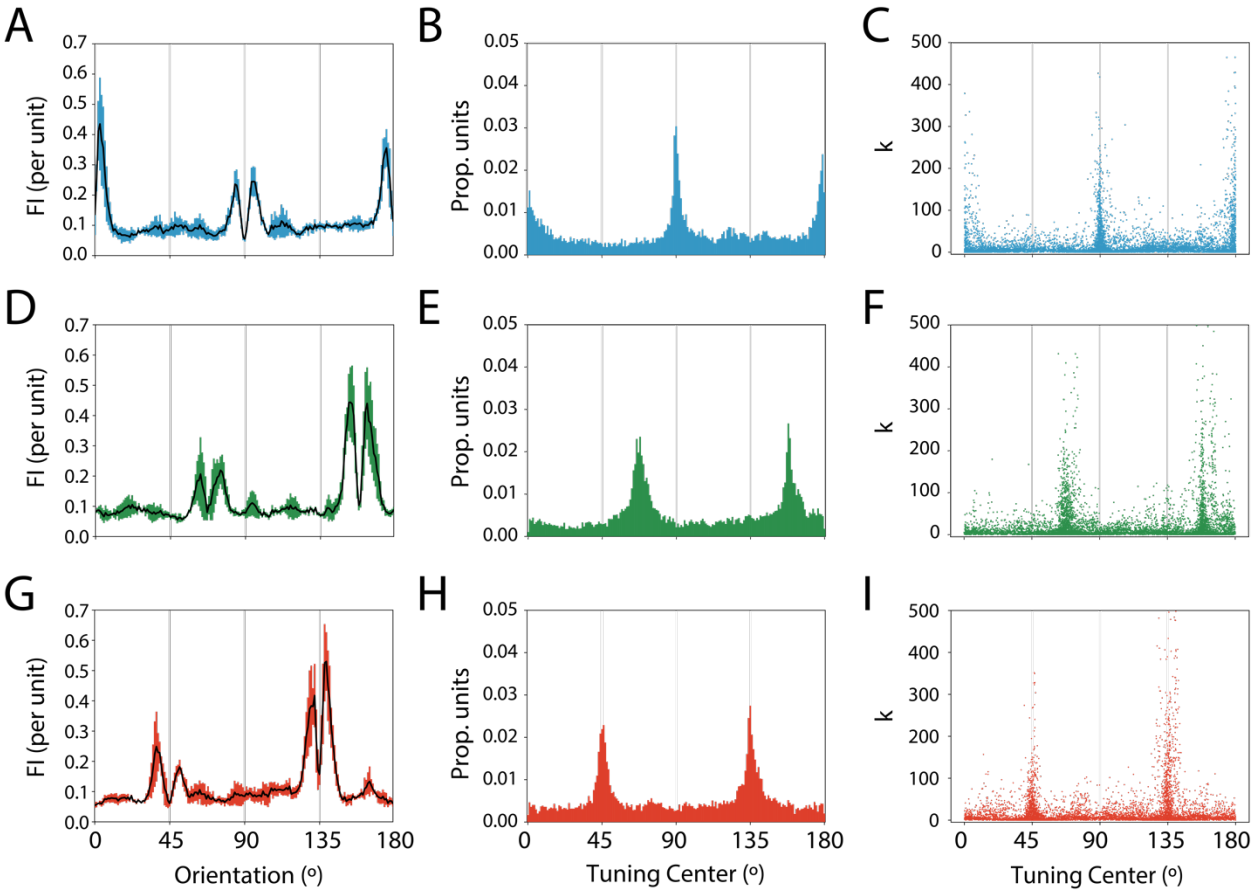


Figure 1.5. When networks are trained on rotated images, both population-level information and single unit tuning distributions reflect modified training set statistics. **(A-C)** show data from one example layer (fc6) of four separately initialized networks trained on upright images, **(D-F)** show data for fc6 of networks trained on images rotated 22.5° counter-clockwise of upright, **(G-I)** show data for fc6 of networks trained on images rotated 45° counter-clockwise of upright. For each group of networks, panels **(A,D,G)** show FI plotted as a function of orientation, with error bars reflecting standard deviation across four networks with the same training image set **(B,E,H)** show distribution of fc6 unit tuning centers, combining data across networks **(C,F,I)** show concentration parameter (k) versus center for individual units.

Calculating the FIB for each of these models further demonstrated how these effects emerged across the processing hierarchy. Like the pre-trained model, the models trained on upright images showed high values of the FIB-0 at middle and deep layers: models showed significantly higher FIB-0 than the randomly initialized models for pool1, conv3_1, and all layers deeper than conv3_1 (one-tailed t-test, FDR corrected q=0.01) (Figure 1.6A). In contrast,

the models trained on images rotated by 22.5° and 45° showed higher values for the FIB-22 and FIB-45, respectively (Figure 1.6B,1.6C). In models trained on images rotated by 22.5°, the FIB-22 significantly exceeded that of the random models at pool2 and all layers deeper than pool2, with the exception of conv3_3 (one-tailed t-test, FDR corrected q=0.01). For the models trained on 45° rotated images, the FIB-45 significantly exceeded that of the random models for conv3_1 and all layers deeper than conv3_1 (one-tailed t-test, FDR corrected q=0.01).

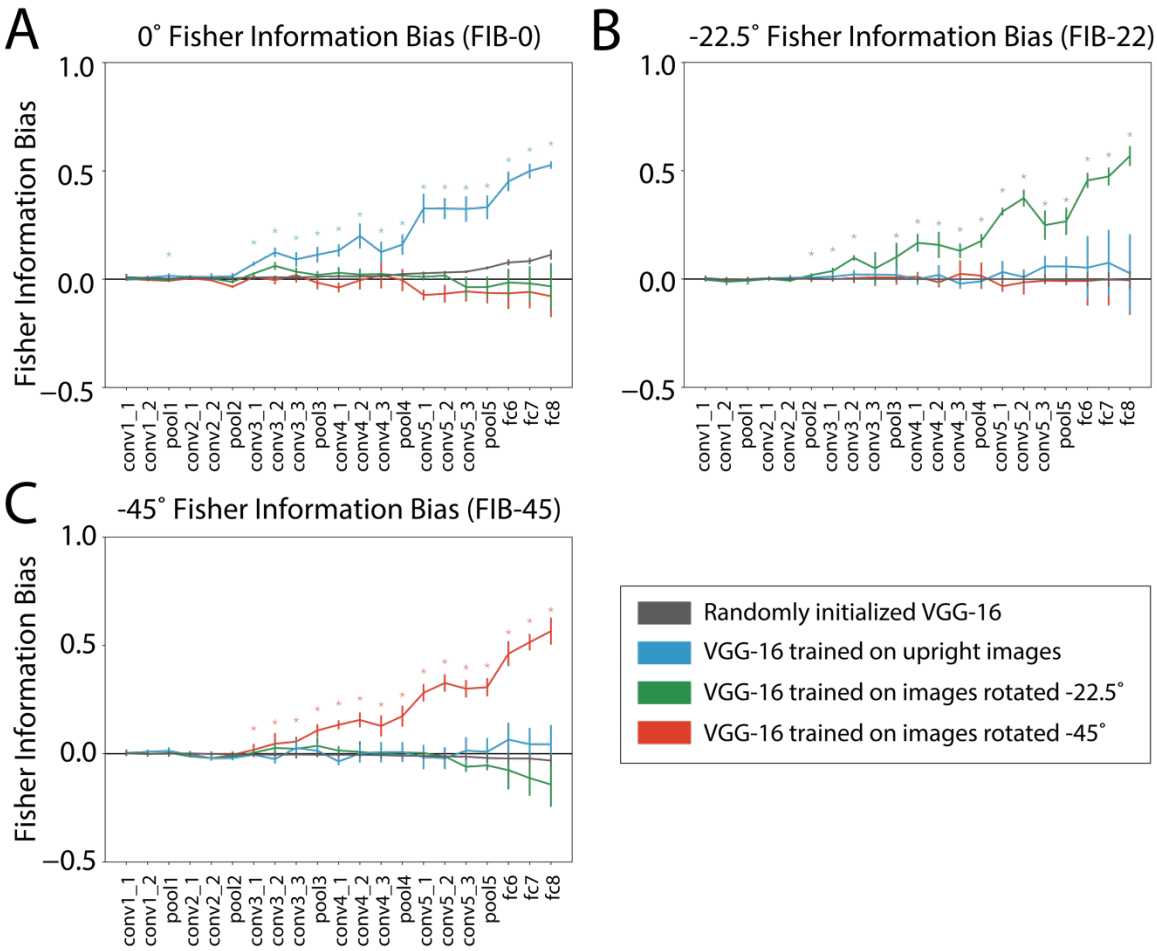


Figure 1.6. Networks shows biases in orientation discriminability that are consistent with training set statistics. FIB-0, FIB-22, and FIB-45 represent the relative value of FI at cardinal orientations, 22.5° counter-clockwise of cardinals, and 45° counter-clockwise of cardinals, respectively, relative to a baseline (see *Methods, Fisher information bias)*. Panels show **(A)** FIB-0, **(B)** FIB-22, and **(C)** FIB-45 for models trained on each rotated version of ImageNet (colored), and randomly initialized models (gray). Colored asterisks indicate layers for which the models corresponding to that color had significantly higher FIB than the random models (one-tailed t-test, FDR corrected q=0.01). Error bars represent the standard deviation of the FIB over four initializations of each model and four evaluation image sets.

Discussion

We investigated whether CNNs trained to perform object classification exhibit biased orientation representations that reflect non-uniformities in the statistics of the training set images. We found that middle and deep layers of a pre-trained VGG-16 network (Silberman & Guadarrama, 2016; Simonyan & Zisserman, 2014) represented orientation with higher discriminability near the cardinal (vertical and horizontal) orientations, with relatively lower discriminability around oblique (diagonal) orientations. Bias was also seen in the tuning properties of single units in the network: there was an over-representation of units that preferred the cardinal orientations, and units tuned to the cardinal orientations had narrower tuning profiles. Furthermore, when we trained models with the same architecture on rotated versions of ImageNet, each of these effects shifted by a predictable amount, such that discriminability was highest near whichever orientations were most common in the network's training set. These results demonstrate that general visual experience with non-uniform image statistics is sufficient to produce the biases that are observed for low-level feature representations in a wide array of biological visual systems.

In general, the strength of the biases we measured tended to increase with depth in each network, showing little or no bias in the first 4-6 layers (Figure 1.3, Figure 1.6). In primates, neural correlates of the oblique effect, reflected by an over-representation of cardinal-tuned neurons, have been shown in V1 (Celebrini et al., 1993; De Valois et al., 1982; Mansfield, 1974), V2 (Shen et al., 2014), and IT cortex (Vogels & Orban, 1994). To relate these physiology findings to our results, we can consider a recent finding that for a similar network, VGG-19, the ability of network activations to explain data from primate V1 was best at an intermediate layer, conv3_1, suggesting that earlier layers of the model might be more analogous to processing in

the retina and/or lateral geniculate nucleus (Cadena et al., 2019). Therefore, our observation that bias did not emerge until the middle layers of the VGG-16 model is roughly consistent with a cortical origin for the oblique effect. The finding that the bias continues to increase with depth in the network is also consistent with some behavioral and physiological results suggesting that the primate oblique effect may be dependent on higher-order processing beyond V1 (Shen et al., 2014; Westheimer, 2003).

Another property of the biases we observed was that the FI measured in deep layers of each network tended to peak just a few degrees off of the orientations that were most common in the training set, with a dip at the precise locations of the most common orientations. As discussed above, this double-peaked shape follows from the fact that FI is highest on the flanks of tuning curves, and many narrowly-tuned units in deep layers tended to have their centers around the most common orientations. However, this finding is not generally reflected in human psychophysics, in which the ability to make small orientation discriminations tends to show a single maximum around each of the cardinal orientations (Appelle, 1972; Girshick et al., 2011). One potential reason for this apparent discrepancy is that in this experiment, we were able to present a relatively large number of images (8640 per image set) to the CNN, with images finely spaced by 1° steps in orientation, whereas psychophysics experiments typically present fewer images at more coarsely spaced orientations (Caelli et al., 1983; Girshick et al., 2011; Westheimer, 2003). Additionally, we were measuring directly from every unit without any additional sources of downstream noise or interference, which may have made the double-peaked shape of Fisher information more apparent than it would be when estimating orientation thresholds from behavior (Butts & Goldman, 2006). It is also possible that this qualitative difference between the FI curves we measured and the shape of human discriminability functions

represents an actual difference between visual processing in CNNs and primates. More extensive behavioral experiments may be needed to resolve this.

Finally, we also observed weak evidence for a cardinal bias in FI measured from the deep layers of a random network with no training (Figure 1.3, Figure 1.6A). This may indicate that some aspect of the model's architecture, such as its use of a square image grid, square convolutional kernels, and pooling operations over square image regions, introduced an intrinsic cardinal reference frame. However, the possible presence of such a reference frame cannot account for the effects we observed for several reasons. First, the magnitude of the FIB-0 was 5x lower for the deepest layer of the random models as compared to the trained-upright models, and the random models did not show an over-representation of cardinal-tuned units, while the upright-trained models did (Figure 1.2B, Figure 1.5B, Supplementary Figure 1.1). This suggests that the network response properties underlying any intrinsic cardinal FI bias were different than those underlying the experience-driven biases we observed. Second, the magnitude of the shifted biases we measured in models trained on rotated images were of similar magnitude to the cardinal biases we measured in models trained on upright images (Figure 1.6), which demonstrates that having an intrinsic reference frame that matches the orientation distribution of training images is not required for a substantial bias to emerge. These results suggest that training may be able to override some intrinsic response properties of CNNs. However, they also highlight the general importance of examining the biases inherent in CNNs before making analogies to the visual system.

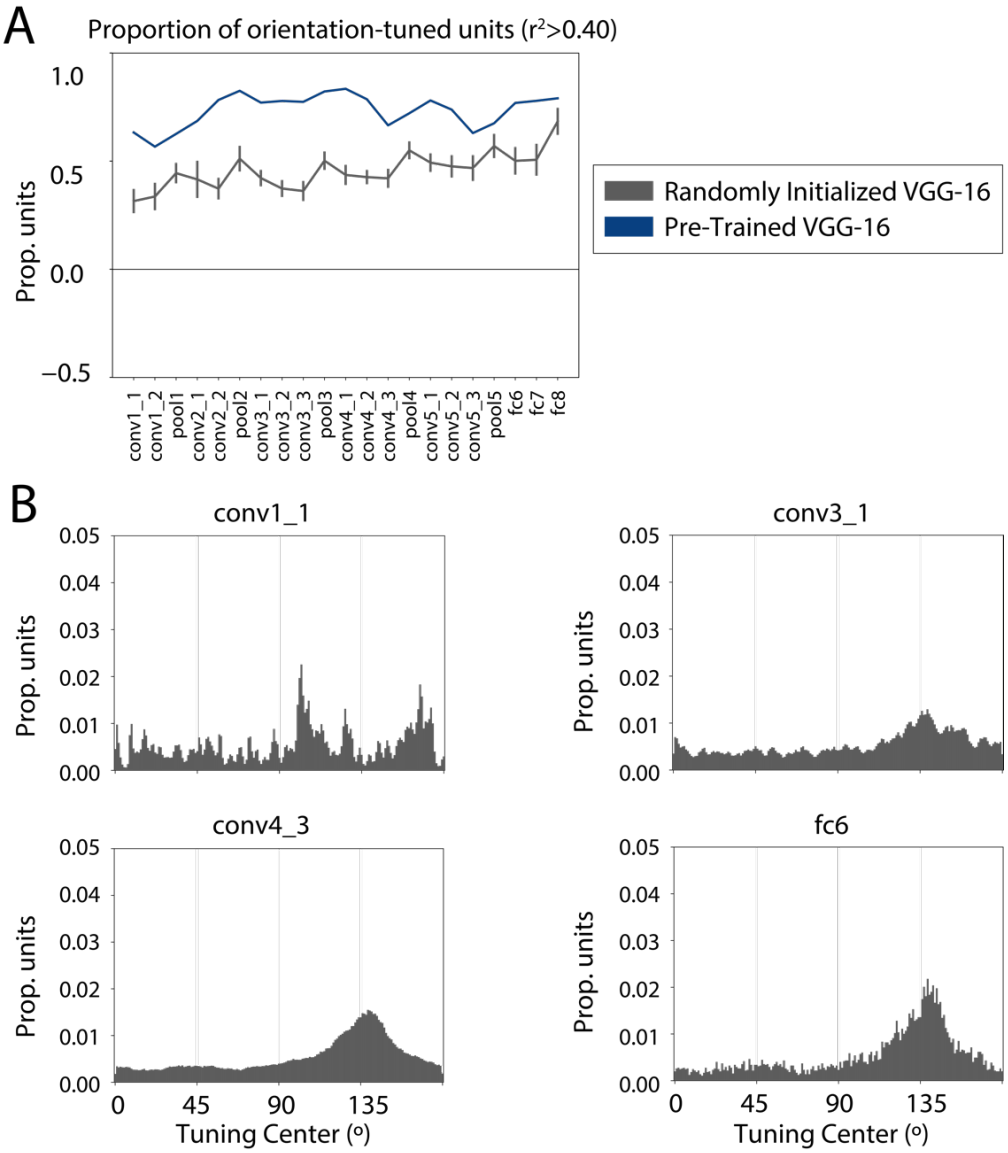
These findings also have general relevance for the use of CNNs in vision research. First, our results show that a popular CNN model exhibits a form of the classical oblique effect, suggesting that this key aspect of low-level primate vision is reproduced by the model. This adds

to a growing body of work demonstrating similarities between deep neural networks and the brains and behavior of primates (Kubilius et al., 2016; Pospisil et al., 2018; Rideaux & Welchman, 2020; Ward, 2019; Yamins et al., 2014). Second, we have demonstrated that non-uniformities in the statistics of training set images can dramatically influence the feature representations that are learned by a CNN. Specifically, image features that are over-represented during training are likely to be more discriminable by the trained network, which may lead to a performance advantage for processing certain stimuli over others. Accounting for such influences is critical for avoiding unwanted algorithmic biases, particularly in modeling high-level visual functions such as face recognition (Cavazos et al., 2019; Klare et al., 2012).

Overall, our results suggest that the classical oblique effect is reproduced in a CNN trained to perform object recognition on an image set containing an over-representation of cardinal orientations. Furthermore, a rotated version of this bias can be induced by training a CNN on rotated versions of these same images. These results indicate that general visual experience, without the presence of an innate bias that matches the viewed orientation distribution, is sufficient to induce the formation of orientation biases, providing support for an experience-driven account of the oblique effect.

Chapter 1 has been submitted for publication and is currently in revision. The author list and working title is Henderson, M. M., & Serences, J. T. (2021). Biased orientation representations can be explained by experience with non-uniform training set statistics. *In revision*. The dissertation author was the primary investigator and author of this paper.

Supplementary Figures



Supplementary Figure 1.1. (A) Proportion of units in each layer that were well-fit by a Von Mises function (see *Methods, Single-unit tuning analysis*), for a pre-trained VGG-16 model (navy blue) and randomly initialized models with no training (gray). Error bars on the gray line reflect standard deviation across four different random initializations of the model. **(B)** Distribution of pre-trained network unit tuning centers for the randomly initialized models (distributions are combined across four different random initializations of the model).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *ArXiv*. http://arxiv.org/abs/1605.08695

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication* (pp. 217–234). https://doi.org/10.7551/mitpress/9780262518420.003.0013

Bauer, J. A., Owens, D. A., Thomas, J., & Held, R. (1979). Monkeys Show an Oblique Effect. *Perception*, *8*(3), 247–253. https://doi.org/10.1068/p080247

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. In *Annals of Statistics* (Vol. 29, Issue 4, pp. 1165–1188). Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1013699998

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, *228*(5270), 477–478. https://doi.org/10.1038/228477a0

Butts, D. A., & Goldman, M. S. (2006). Tuning Curves, Neuronal Variability, and Sensory Coding. *PLoS Biology*, *4*(4), e92. https://doi.org/10.1371/journal.pbio.0040092

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897. https://doi.org/10.1371/journal.pcbi.1006897

Caelli, T., Brettel, H., Rentschler, I., & Hilz, R. (1983). Discrimination thresholds in the two-dimensional spatial frequency domain. *Vision Research*, *23*(2), 129–133. https://doi.org/10.1016/0042-6989(83)90135-9

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2019). *Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?* http://arxiv.org/abs/1912.07398

Celebrini, S., Thorpe, S., Trotter, Y., & Imbert, M. (1993). Dynamics of orientation coding in area VI of the awake primate. *Visual Neuroscience*, *10*(5), 811–825. https://doi.org/10.1017/S0952523800006052

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Coppola, D. M., Purves, H. R., McCoy, A. N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 4002–4006. https://doi.org/10.1073/pnas.95.7.4002

Coppola, D. M., & White, L. E. (2004). Visual experience promotes the isotropic representation of orientation preference. *Visual Neuroscience*, *21*(1), 39–51. https://doi.org/10.1017/s0952523804041045

De Valois, R. L., William Yund, E., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 531–544. https://doi.org/10.1016/0042-6989(82)90112-2

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPRW.2009.5206848

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Hirsch, H. V. B., & Spinelli, D. N. (1970). Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science*, *168*(3933), 869–871. https://doi.org/10.1126/science.168.3933.869

Hoy, J. L., & Niell, C. M. (2015). Layer-specific refinement of visual cortex function after eye opening in the awake mouse. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(8), 3370–3383. https://doi.org/10.1523/JNEUROSCI.3174-14.2015

Kell, A. J., & McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. In *Current Opinion in Neurobiology* (Vol. 55, pp. 121–132). Elsevier Ltd. https://doi.org/10.1016/j.conb.2019.02.003

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Kreile, A. K., Bonhoeffer, T., & Hübener, M. (2011). Altered visual experience induces instructive changes of orientation preference in mouse visual cortex. *Journal of Neuroscience*, *31*(39), 13911–13920. https://doi.org/10.1523/JNEUROSCI.2143-11.2011

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology*, *12*(4), e1004896. https://doi.org/10.1371/journal.pcbi.1004896

Leventhal, A. G., & Hirsch, H. V. B. (1975). Cortical effect of early selective exposure to diagonal lines. *Science*, *190*(4217), 902–904. https://doi.org/10.1126/science.1188371

Leventhal, A. G., & Hirsch, H. V. B. (1980). Receptive-field properties of different classes of neurons in visual cortex of normal and dark-reared cats. *Journal of Neurophysiology*, *43*(4), 1111–1132. https://doi.org/10.1152/jn.1980.43.4.1111

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Mansfield, R. J. W. (1974). Neural basis of orientation perception in primate vision. *Science*, *186*(4169), 1133–1135. https://doi.org/10.1126/science.186.4169.1133

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *ELife*, *7*. https://doi.org/10.7554/eLife.38242

Rideaux, R., & Welchman, A. E. (2020). But still it moves: Static image statistics underlie how we see motion. *Journal of Neuroscience*, *40*(12), 2538–2552. https://doi.org/10.1523/JNEUROSCI.2760-19.2020

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Shen, G., Tao, X., Zhang, B., Smith, E. L., Chino, Y. M., & Chino, Y. M. (2014). Oblique effect in visual area 2 of macaque monkeys. *Journal of Vision*, *14*(2). https://doi.org/10.1167/14.2.3

Silberman, N., & Guadarrama, S. (2016). *TensorFlow-Slim image classification model library*.

Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. http://arxiv.org/abs/1409.1556

Vogels, R., & Orban, G. A. (1994). Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *Journal of Neurophysiology*, *71*(4), 1428–1451. https://doi.org/10.1152/jn.1994.71.4.1428

Ward, E. J. (2019). Exploring perceptual illusions in deep neural networks. *Journal of Vision*, *19*(10), 34b. https://doi.org/10.1167/19.10.34b

Westheimer, G. (2003). Meridional anisotropy in visual processing: Implications for the neural site of the oblique effect. *Vision Research*, *43*(22), 2281–2289. https://doi.org/10.1016/S0042-6989(03)00360-2

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

CHAPTER 2: Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features

J Neurophysiol 121: 1410–1427, 2019.
First published February 13, 2019; doi:10.1152/jn.00015.2019.

RESEARCH ARTICLE | *Higher Neural Functions and Behavior*

Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features

Margaret Henderson[1] and John T. Serences[1,2,3]

[1]*Neurosciences Graduate Program, University of California, San Diego, La Jolla, California;* [2]*Department of Psychology, University of California, San Diego, La Jolla, California; and* [3]*Kavli Foundation for the Brain and Mind, University of California, San Diego, La Jolla, California*

Submitted 14 January 2019; accepted in final form 5 February 2019

Henderson M, Serences JT. Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features. *J Neurophysiol* 121: 1410–1427, 2019. First published February 13, 2019; doi:10.1152/jn.00015.2019.—Searching for items that are useful given current goals, or "target" recognition, requires observers to flexibly attend to certain object properties at the expense of others. This could involve focusing on the identity of an object while ignoring identity-preserving transformations such as changes in viewpoint or focusing on its current viewpoint while ignoring its identity. To effectively filter out variation due to the irrelevant dimension, performing either type of task is likely to require high-level, abstract search templates. Past work has found target recognition signals in areas of ventral visual cortex and in subregions of parietal and frontal cortex. However, target status in these tasks is typically associated with the identity of an object, rather than identity-orthogonal properties such as object viewpoint. In this study, we used a task that required subjects to identify novel object stimuli as targets according to either identity or viewpoint, each of which was not predictable from low-level properties such as shape. We performed functional MRI in human subjects of both sexes and measured the strength of target-match signals in areas of visual, parietal, and frontal cortex. Our multivariate analyses suggest that the multiple-demand (MD) network, including subregions of parietal and frontal cortex, encodes information about an object's status as a target in the relevant dimension only, across changes in the irrelevant dimension. Furthermore, there was more target-related information in MD regions on correct compared with incorrect trials, suggesting a strong link between MD target signals and behavior.

NEW & NOTEWORTHY Real-world target detection tasks, such as searching for a car in a crowded parking lot, require both flexibility and abstraction. We investigated the neural basis of these abilities using a task that required invariant representations of either object identity or viewpoint. Multivariate decoding analyses of our whole brain functional MRI data reveal that invariant target representations are most pronounced in frontal and parietal regions, and the strength of these representations is associated with behavioral performance.

fMRI; frontoparietal; invariance; object; target recognition

INTRODUCTION

To flexibly guide behavior, humans can choose to hold in mind information about the identity of sought-after items or about the current state of those items. For example, when searching the parking lot at the end of a long day, you might search for the presence of your blue sedan, but when crossing the street, you might search for any car moving quickly in the rightward direction. The former task is challenging because the retinal projection of your car can have considerable variability due to changes in pose, position, and environmental conditions, whereas the latter task is challenging because relevant cars may be a variety of makes, models, sizes, and colors (DiCarlo and Cox 2007; Ito et al. 1995; Lueschow et al. 1994; Marr and Nishihara 1978; Tanaka 1993). To overcome both types of challenges, recognizing relevant targets under realistic viewing conditions is likely to require high-level, abstract search templates (Biederman 2000; Freiwald and Tsao 2010; Riesenhuber and Poggio 2000; Tarr et al. 1998).

Such abstract search templates have been found in multiple brain regions. In inferotemporal (IT) cortex, neurons encode object identity across identity-preserving transformations, even during passive viewing (Anzellotti et al. 2014; Erez et al. 2016; Freiwald and Tsao 2010; Tanaka 1996). Neurons in IT and entorhinal cortex (ERC) also signal the target status of objects, both when targets can be identified on the basis of an exact match of retinal input and when targets have to be identified across changes in size and position (Lueschow et al. 1994; Miller and Desimone 1994; Pagan et al. 2013; Roth and Rust 2018; Woloszyn and Sheinberg 2009). In addition to these ventral regions, neurons in prefrontal cortex (PFC) have been shown to signal the target status of objects, both for superordinate and subordinate identification tasks (Freedman et al. 2003; Kadohisa et al. 2013; McKee et al. 2014; Miller et al. 1996).

In agreement with the single-unit modulations in prefrontal cortex, recent studies in humans suggest that a set of frontal and parietal regions, collectively referred to as the multiple-demand (MD) network, may play a role in target selection by representing objects according to their task-relevant properties (Bracci et al. 2017; Duncan 2010; Fedorenko et al. 2013; Jackson et al. 2017; Vaziri-Pashkam and Xu 2017). Accordingly, MD representations have also been found to differentiate images on the basis of their status as a target object or category, and these representations exhibit invariance across changes in low-level image properties (Erez and Duncan 2015; Guo et al. 2012). Target representations in frontoparietal regions are also

Address for reprint requests and other correspondence: M. Henderson, Neurosciences Graduate Program, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA, 92093 (e-mail: mmhender@ucsd.edu).

associated with decision confidence and task difficulty, suggesting that they play a role in shaping decisions (Guo et al. 2012). However, in all past studies, sought targets were defined according to the identity or category of the object, dimensions whose representation in the visual system has been extensively characterized (Conway 2018; DiCarlo and Cox 2007; Grill-Spector 2003; Tanaka 1996). Less well studied is how the visual system computes matches in dimensions orthogonal to object identity, such as object pose or viewpoint. Past work has demonstrated sensitivity to object viewpoint in multiple regions of the human and primate brain, including IT cortex and the intraparietal sulcus (IPS), but it is not yet clear how viewpoint representations are involved in identification of relevant targets (Andresen et al. 2009; Grill-Spector et al. 1999; Hong et al. 2016; Tanaka 1993; Valyear et al. 2006; Ward et al. 2018). Thus, although we predict that representations of target status based on object viewpoint will be found in the same frontoparietal regions known to encode target status based on identity and category of objects, this has yet to be shown.

In the present study we tested the hypothesis that regions of the MD network support performance during a task where target status is defined on the basis of either object identity or object viewpoint. We generated a novel object stimulus set (Fig. 1), in which three-dimensional (3D) objects of multiple identities were rendered at multiple viewpoints. Subjects determined the target status of objects according to either their identity (identity task) or viewpoint (viewpoint task) while ignoring the other dimension. Critically, this paradigm required subjects to form viewpoint-invariant representations of identity and identity-invariant representations of viewpoint, both of which were defined so as not to be predictable from the retinotopic shape of an object. We used multivariate pattern analyses (MVPA) on single-trial voxel activation patterns to decode the status of each image as a target in both the task-relevant and the task-irrelevant dimensions. Our findings suggest that whereas ventral visual cortex exhibits some sensitivity to an object's status as a target, regions of the MD network encode robust, abstract target representations that are sensitive to changes in task demands and that are selectively linked with behavioral performance.

MATERIALS AND METHODS

Participants. Ten subjects (3 men) between the ages of 20 and 34 yr were recruited from the University of California, San Diego (UCSD) community (mean age 24.7 ± 4.7 yr), having normal or

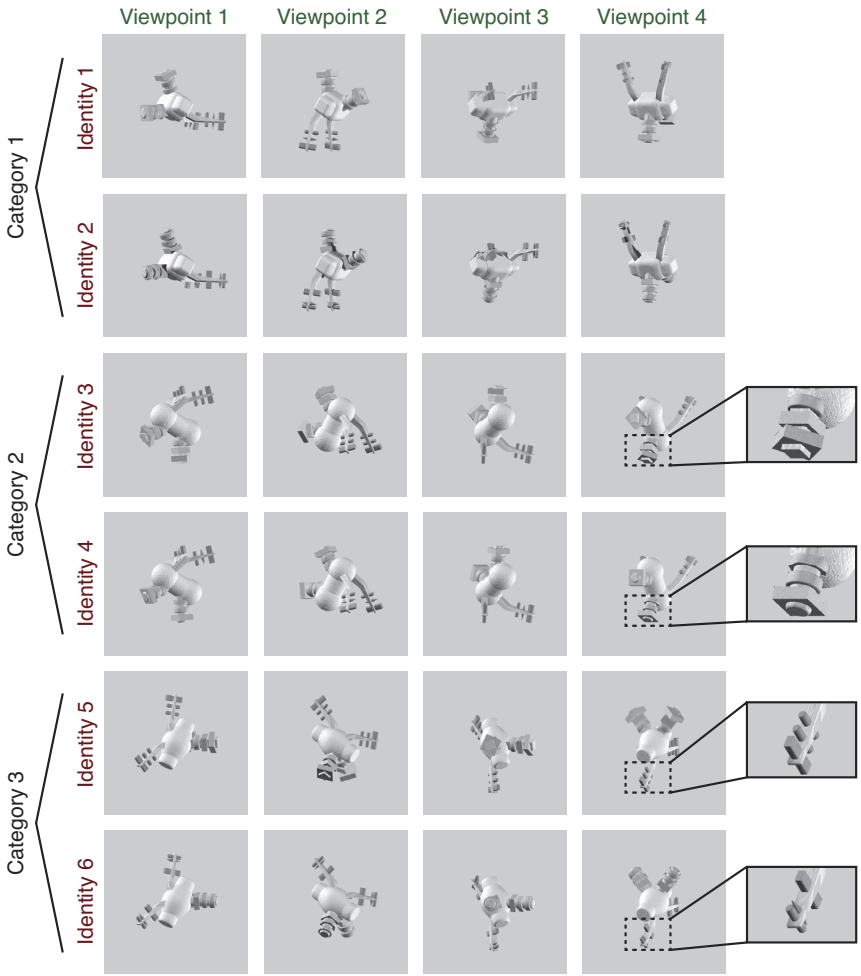


Fig. 1. Example set of images shown to a subject during scanning, consisting of 6 unique object identities, each rendered at 4 viewpoints. Subjects were instructed either to match the exact identity of the object irrespective of viewpoint (shown in rows of the matrix) or to match the viewpoint of the object irrespective of identity (columns of the matrix). The 6 identities comprised 2 exemplars in each of 3 categories, with categories defined by overall body shape and exemplars defined by details of the peripheral features (see *insets* for examples of differentiating features). Object viewpoint was generated in an arbitrarily defined coordinate system so that low-level visual features had a minimal contribution to the viewpoint matching task (see METHODS for details). Two complete sets of novel objects were generated, with half the subjects (5/10) viewing *set A* and half viewing *set B*. The images shown are from object *set B*; see Fig. 2 for examples of object *set A*.

corrected-to-normal vision. The study protocol was submitted to and approved by the Institutional Review Board at UCSD, and all participants provided written informed consent. Each subject performed a behavioral training session lasting ~1 h, followed by one or two scan sessions, each lasting ~2 h. Participants were compensated at a rate of $10/h for behavioral training and $20/h for the scanning sessions.

Novel object sets and one-back tasks. All objects were generated and rendered using Strata 3D CX software (version 7.6; Santa Clara, UT). To ensure a variety of stimuli, we generated two unique sets of objects and assigned half of our subjects (5 of 10) to object *set A* and half to object *set B* (because we did not observe a difference in performance between the 2 stimulus sets for either task, we combined our analyses across all subjects). Each stimulus set comprised 3 categories of objects, each with 36 total exemplars. Objects within a category shared a common body plan, including the shape of the main body and the configuration of peripheral features around the body (see Figs. 1 and 2). Exemplars in each category were differentiated by small variations in the details of peripheral features, such as the size or shape of a spike. These peripheral features always appeared in pairs that were attached symmetrically to the body, making the overall objects bilaterally symmetric. Feature details were always matched within each peripheral feature pair, ensuring that even when one feature in a pair was occluded at a particular viewpoint, the details could always be discerned from the other feature in the pair. During scanning, each subject viewed two exemplars in each category (selected on an individual subject basis from the full set of 36 exemplars; see below for details), giving a total of six object "identities" (2 exemplars in each of 3 categories). Each of the six object identities was rendered from four different viewpoints, for a total of 24 unique images shown to each subject.

While in the scanner, each subject performed two different one-back tasks (identity task and viewpoint task). In the identity task, subjects responded to each image on the basis of whether it matched the identity of the immediately preceding image. Identity matches had to be the same exemplar from the same category but did not have to match in viewpoint. In the viewpoint task, subjects responded on the basis of whether the current image matched the viewpoint of the immediately preceding image, whereas both category and exemplar status were irrelevant. In order for subjects to identify matches in viewpoint between objects with different identities, they were trained to recognize an arbitrary viewpoint of each object as its "frontal" viewpoint, or a rotation of [0°, 0°] about the *Y* (vertical)- and *Z* (front-back)-axes. All other viewpoints were then defined relative to this reference point. The four viewpoints used, in coordinates of [*Z* rotation, *Y* rotation], were [30°, 300°], [120°, 240°], [210°, 30°], and [300°, 150°]. We defined viewpoint in this way to ensure that subjects formed a representation of viewpoint that was largely invariant to 2D shape. Importantly, because the frontal viewpoint was chosen arbitrarily, it was not systematically predictable from the overall axis of elongation of the main body. As a result, images of the objects in

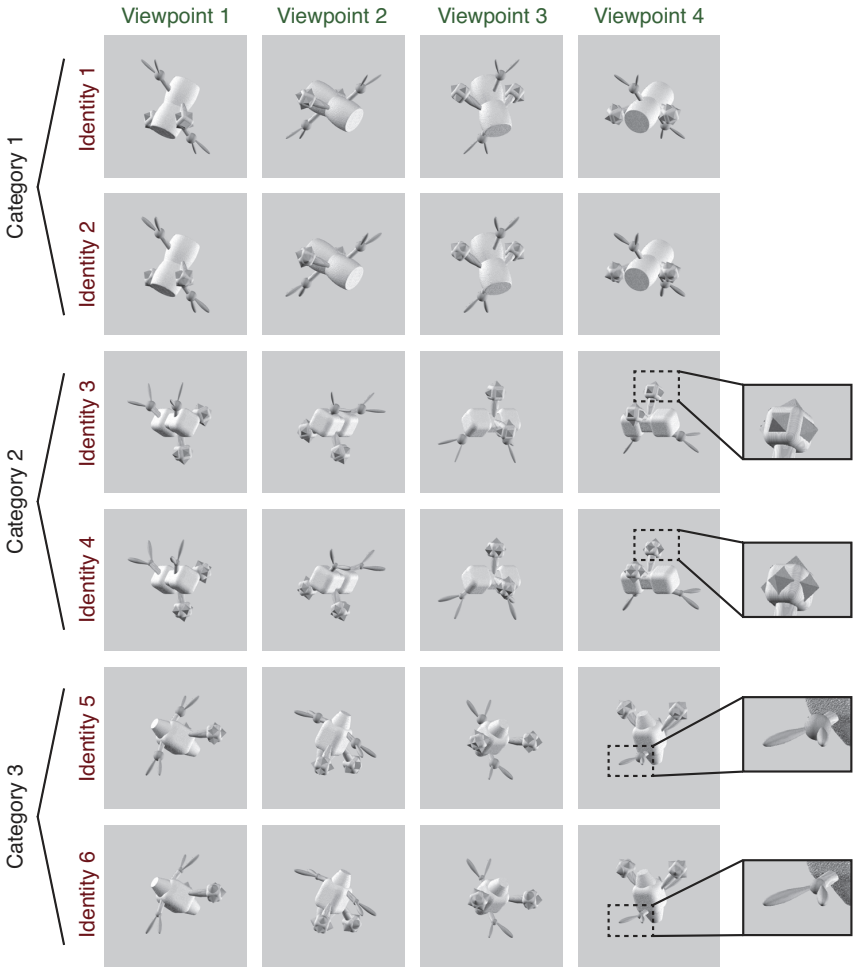


Fig. 2. Example stimulus set from object *set A* (see Fig. 1 and METHODS for details).

different categories at the same viewpoint differed dramatically in the shape they produced when projected onto a 2D plane.

Prescan training. Subjects were familiarized with the novel object viewpoints during a self-guided session performed before scanning. During the first half of the training session, subjects viewed a 3D model of each of the three novel object categories and were able to rotate the model around two axes using the arrow keys on a keyboard (each key press gave a rotation of 30° about either the *Y*- or *Z*-axis). During this entire exercise, the angular position of the object, expressed using the format "*X* rotation = *m* degrees, *Y* rotation = *n* degrees," was displayed at the top of the screen. Subjects were encouraged to use the angular coordinates to learn how each view of the object was defined relative to the arbitrarily chosen frontal ([0°, 0°]) position. During the second half of this training, subjects were presented with images of all three categories simultaneously, at matching viewpoints, and encouraged to study how the same viewpoint was defined across categories. Subjects performed both parts of this training at least once and were allowed to return to it as many times as they wished. On average, subjects spent ~20 min on the self-guided training. In addition to the self-guided training, subjects performed several practice runs of the viewpoint task. During these runs, the four object viewpoints that were used during scanning were never used so that even though subjects were familiarized with different viewpoints of each object category, they were not overexposed to the target viewpoints. After each viewpoint task practice run, subjects could return to the self-guided viewpoint training, and they repeated as many iterations of self-guided training and practice runs as were necessary to reach 70% performance (between 4 and 10 runs across subjects).

The six object identities viewed by each subject were selected on the basis of a behavioral thresholding experiment performed before scanning. This allowed us to control the difficulty of the identity task by manipulating the similarity between the exemplars in each category. The task used for thresholding was identical to the one-back identity task used during scanning, but it used only objects from a single category, presented at four random viewpoints. Subjects performed six runs of this task, with two runs for each object category. On the basis of their performance, we selected two exemplars in each object category that were confusable ~70% of the time. Following this thresholding procedure, subjects performed two practice runs of the identity task, using the final set of exemplars that they would view during scanning.

Immediately before each scanning session, subjects performed another short self-guided training exercise (~5 min), in which they were shown examples of the exact images they would see during scanning. First, they were presented with the two exemplars in each object category, side by side, and allowed to freely rotate the objects using the arrow keys to compare the appearance of the two exemplars from many viewpoints. Next, they were presented with images from each category, side by side, at each of the four viewpoints they would see during the task and encouraged to use this information to prepare for the viewpoint task.

Behavioral task in the scanner. During scanning, subjects performed between 8 and 11 runs each of the identity and viewpoint tasks. Task runs always occurred in pairs with the identity task followed by the viewpoint task. An identical sequence of visual stimuli was presented on both runs in each pair so that visual stimulation was perfectly matched between conditions. Each 6-min run consisted of 48 trials, and each trial consisted of a single image presentation for 1,500 ms, followed by a jittered intertrial interval ranging from 2,000 to 6,000 ms. Each of the 24 images was shown twice per run. The sequence of image presentations was pseudorandomly generated, with the constraint that there was a 0.50 probability that the current stimulus was from the same category as the previous stimulus (within-category trials). This constraint was adopted to more closely equate the difficulty of the two tasks, because the viewpoint task was inherently more difficult to solve on across-category trials, and the identity task was more difficult to solve on within-category trials. This resulted in a probability of 0.23 of any trial being a match in either viewpoint or identity, and a probability of 0.04 of a match in both dimensions.

In both tasks, subjects responded to every image by pressing a button using either their index finger ("1") or their middle finger ("2"), depending on the current response mapping rule. Response mapping rules were counterbalanced within each subject so that on half of the runs the subject responded with 1 for "match" and 2 for "non-match," and on the other half of runs they responded with 1 for "non-match" and 2 for "match." The purpose of these different response mapping rules was to ensure that match-related information was not confounded with motor responses.

Magnetic resonance imaging. All MRI scanning was performed on a General Electric (GE) Discovery MR750 3.0-T research-dedicated scanner at the UC San Diego Keck Center for Functional Magnetic Resonance Imaging. Functional echo-planar imaging (EPI) data were acquired using a Nova Medical 32-channel head coil (NMSC075-32-3GE-MR750) and the Stanford Simultaneous Multi-Slice (SMS) EPI sequence (MUX EPI), with a multiband factor of 8 and 9 axial slices per band (total slices = 72; 2-mm^3 isotropic; 0-mm gap; matrix = 104 × 104; field of view = 20.8 cm; TR/TE = 800/35 ms; flip angle = 52°; in-plane acceleration = 1). Image reconstruction procedures and unaliasing procedures were performed on local servers using reconstruction code from CNI (Center for Neural Imaging at Stanford). The initial 16 repetition times (TRs) collected at sequence onset served as reference images required for the transformation from k-space to the image space. Two short (17 s) "topup" data sets were collected during each session, using forward and reverse phase-encoding directions. These images were used to estimate susceptibility-induced off-resonance fields (Andersson et al. 2003) and to correct signal distortion in EPI sequences using FSL (FMRIB Software Library) topup functionality (Jenkinson et al. 2012).

During each functional session, we also acquired an accelerated anatomical scan using parallel imaging [GE ASSET on a fast spoiled gradient-echo (FSPGR) T1-weighted sequence; 1 × 1 × 1-mm^3 voxel size; TR = 8,136 ms; TE = 3,172 ms; flip angle = 8°; 172 slices; 1-mm slice gap; 256 × 192-cm matrix size] using the same 32-channel head coil. We also acquired one additional high-resolution anatomical scan for each subject (1 × 1 × 1-mm^3 voxel size; TR = 8,136 ms; TE = 3,172 ms; flip angle = 8°; 172 slices; 1-mm slice gap; 256 × 192-cm matrix size) during a separate retinotopic mapping session using an Invivo eight-channel head coil. This scan produced higher quality contrast between gray and white matter and was used for segmentation, flattening, and visualizing retinotopic mapping data.

In addition to the multiband scan protocol described above, five subjects participated in retinotopic mapping experiments using a different scan protocol, previously reported (Sprague and Serences 2013). The remaining five subjects participated in retinotopic mapping runs using the multiband protocol described above. Where possible, the data used to generate retinotopic maps (see *Retinotopic mapping stimulus protocol*) were combined across these sessions.

Preprocessing. First, the structural scan from each session was processed in BrainVoyager 2.6.1 to align the anatomical and the functional data sets. Automatic algorithms were used to adjust the structural image intensity to correct for inhomogeneities, as well as to remove the head and skull tissue. Structural scans were then aligned to the anterior commissure-posterior commissure (AC-PC) plane using manual landmark identification. Finally, an automatic registration algorithm was used to align the structural scan to the high-definition structural scan collected during each subject's retinotopic mapping session. This high-definition structural scan was transformed into Talairach space, and the parameters of this transformation were used to transform all other scans for this subject into Talairach space.

Next, each functional run was aligned to the same-session structural scan. We then used BrainVoyager 2.6.1 to perform slice-time correction, affine motion correction, and temporal high-pass filtering

to remove first-, second-, and third-order signal drifts over the course of each functional run. These data were spatially transformed into Talairach space to align with the anatomical images. Finally, the blood oxygen level-dependent (BOLD) signal in each voxel was *z*-transformed within each run.

General linear model to estimate trial-by-trial responses. After preprocessing, single-trial activation estimates (beta weights), which were used for subsequent MVPA, were obtained using a general linear model (GLM) with a design matrix created by convolving the trial sequence with the canonical two-gamma hemodynamic response function (HRF) as implemented in BrainVoyager (peak at 5 s, undershoot peak at 15 s, response undershoot ratio 6, response dispersion 1, undershoot dispersion 1). Throughout this study, the same HRF parameters were used for all GLM analyses.

Retinotopic mapping stimulus protocol. We followed previously published retinotopic mapping protocols to define the visual areas V1, V2, V3, V3AB, V4, IPS0–1, and IPS2–3 (Engel et al. 1997; Jerde and Curtis 2013; Sereno et al. 1995; Swisher et al. 2007; Wandell et al. 2007; Winawer and Witthoft 2015). Subjects performed mapping runs in which they viewed a contrast-reversing checkerboard stimulus (4 Hz) configured as a rotating wedge (10 cycles, 36 s/cycle), an expanding ring (10 cycles, 36 s/cycle), or a bowtie (8 cycles, 40 s/cycle). To increase the quality of data from parietal regions, subjects performed a covert attention task on the rotating wedge stimulus, which required them to detect contrast-dimming events that occurred occasionally (on average, 1 event occurred every 7.5 s) in a row of the checkerboard (mean accuracy = 61.8 ± 13.9%). This stimulus was limited to a 22° × 22° field of view.

Multiple-demand localizer. To define regions of interest (ROIs) in the MD network, we used an independent functional localizer to identify voxels whose BOLD response was significantly modulated by the load of a spatial working memory task (Duncan 2010; Fedorenko et al. 2013). Subjects performed one or two runs of this task during each functional scanning session. During each trial of this task, subjects were first presented with an empty rectangular grid comprising either 8 or 16 squares. Half of the squares in the grid were then highlighted one at a time, and subjects were required to remember the locations of the highlighted squares. Subjects were then shown a probe grid and asked to report whether the highlighted squares matched the remembered locations. Runs were divided into blocks with either high or low load. Performance was significantly poorer on high-load blocks (mean d' for low load = 2.48 ± 0.28, mean d' for high load = 1.04 ± 0.21, $P < 0.001$; paired 2-tailed *t*-test).

We used the data from these runs to generate a statistical parametric map for each subject, which expressed the degree to which each voxel showed elevated BOLD signal for high-load vs. low-load working memory blocks. We defined a GLM with a regressor for each block type and solved for the β-coefficients corresponding to each load condition. Coefficients were then entered into a one-tailed, repeated-measures *t*-test against a distribution with a mean of 0 [false discovery rate (FDR)-corrected $q = 0.05$]. This resulted in a single mask of load-selective voxels for each subject.

To subdivide this mask into the typical MD ROIs, we used a group-level parcellation from a previously published data set (Fedorenko et al. 2013). We used this parcellation to generate masks for five ROIs of interest: the intraparietal sulcus (IPS), the superior precentral sulcus (sPCS), the inferior precentral sulcus (iPCS), the anterior insula/frontal operculum (AI/FO), and the inferior frontal sulcus (IFS). Because we had already defined two posterior subregions of the IPS, IPS0–1 and IPS2–3 (see *Retinotopic mapping stimulus protocol*), we removed all voxels belonging to these retinotopic regions from the larger IPS mask and used the remaining voxels to define a region that we refer to as the superior IPS (sIPS). This was done within each subject separately. We intersected each subject's mask of load-selective voxels with the mask for each ROI to generate the final MD ROI definitions.

For one subject, this procedure failed to yield any voxels in the sPCS ROI. Therefore, when performing group-level ANOVA of decoding performance, we performed linear interpolation (Roth 1994) based on sPCS responses in the remaining nine subjects to generate an estimate of the missing value. We did this by calculating a *t*-score comparing the missing subject's d' score with those of the other nine subjects in each ROI and condition where it was defined and using the mean of these *t*-scores to estimate d' in sPCS for each condition. As an alternative to this interpolation method, we also ran the repeated-measures ANOVA with all values for the missing subject removed: we observed similar results (see Table 2). For all ANOVA results reported in this paper, Mauchly's test revealed that the data did not violate the assumption of sphericity, so we report uncorrected *P* values.

Lateral occipital complex localizer. We identified two subregions of the lateral occipital complex, LO and pFus, using a functional localizer developed by the Stanford Vision and Perception Laboratory (Stigliani et al. 2015) to identify voxels that showed enhanced responses to intact objects (cars and guitars) vs. phase-scrambled versions of the same images. Between two and four runs of this task were performed during functional scanning sessions, with each run lasting 5 min and 16 s. During each run, subjects viewed blocks of sequentially presented images in a particular category (cars, guitars, faces, houses, body parts, scrambled objects) and performed a one-back repeat detection task (mean $d' = 3.09 \pm 0.49$). We used a GLM to define voxels that showed significantly higher BOLD responses during car/guitar blocks vs. scrambled blocks (FDR-corrected $q = 0.05$). We then projected this mask onto a computationally inflated mesh of the gray matter-white matter boundary in each subject and defined LO and pFus on this mesh, based on the mask in conjunction with anatomical landmarks (Vinberg and Grill-Spector 2008).

Object localizer task. After the ROIs described above were defined, voxels within each ROI were thresholded on the basis of their visual responsiveness during performance on an independent novel object matching task. Subjects performed two to three runs of this task during each scanning session. This task was identical to the identity task described above, except that the alternate object set was used (e.g., if the subject viewed *set A* during the main one-back task runs, they viewed *set B* during the localizer). The object exemplars shown during this task were randomly selected for each run. Performance on this task was consistently lower than performance on the main one-back tasks, due to the fact that subjects had not been trained on this stimulus set (mean $d' = -0.23 \pm 0.14$).

For each subject, we combined data from all object localizer runs to generate a statistical parametric map of voxel responsiveness, based on a GLM in which all image presentations were modeled as a single predictor. We then selected only the voxels whose BOLD signal was significantly modulated by image presentation events (FDR-corrected $q = 0.05$). This limited the voxels selected in each ROI to those that were responsive to object stimuli that were visually similar (but not identical) to those presented during the main task. For one ROI in the MD network (AI/FO), this thresholding procedure yielded fewer than 10 voxels for several subjects, so for this ROI we chose to analyze all the voxels that were defined by the MD localizer. The final definitions of each ROI (centers and sizes) following this thresholding procedure are summarized in Table 1.

MVPA decoding. The goal of our MVPA analysis was to estimate the amount of linearly decodable information about object "match" status in each task dimension (identity and viewpoint) that was represented in each ROI during each task. Because match status depended on the relation of each object to the previous one in the sequence, each object had an equal probability of appearing as a match or a nonmatch in each dimension, so this decoding was orthogonal to the visual properties of the objects. To evaluate the behavioral relevance of information about match/nonmatch status in each ROI, we also assessed how match decoding was affected by the

Table 1. *Centers and sizes of the final ROIs defined for each subject, following functional localization and additional thresholding with a novel object localizer*

Subject	Center LH	Center RH	No. of Voxels LH	No. of Voxels RH
		V1		
1	[−16, −94, −12]	[12, −92, −17]	331	368
2	[−9, −82, 8]	[6, −82, 6]	797	444
3	[−15, −97, 7]	[11, −97, 5]	84	192
4	[−12, −97, −8]	[14, −94, 4]	249	186
5	[−6, −95, −4]	[16, −95, 1]	273	164
6	[−8, −90, −12]	[13, −89, −1]	269	434
7	[−9, −88, −5]	[11, −92, −5]	406	393
8	[−11, −92, −8]	[10, −91, −4]	348	254
9	[−12, −93, −8]	[15, −96, −5]	477	324
10	[−16, −89, −18]	[12, −92, −9]	299	239
		V2		
1	[−23, −92, −11]	[14, −90, −15]	325	311
2	[−14, −86, 7]	[19, −82, 5]	495	602
3	[−17, −92, 4]	[15, −93, 3]	271	226
4	[−13, −93, −7]	[18, −92, 4]	354	212
5	[−10, −94, −8]	[19, −93, 5]	179	167
6	[−13, −85, −13]	[18, −89, −1]	272	388
7	[−14, −93, −6]	[15, −93, −5]	363	419
8	[−14, −87, −12]	[16, −89, −8]	416	308
9	[−13, −92, −6]	[14, −90, −6]	277	300
10	[−20, −88, −20]	[17, −90, −7]	370	364
		V3		
1	[−25, −85, −8]	[22, −88, −9]	331	360
2	[−22, −83, 1]	[12, −88, 6]	500	515
3	[−19, −87, 6]	[21, −89, 3]	247	320
4	[−20, −90, −3]	[22, −85, 6]	359	216
5	[−20, −88, −7]	[29, −85, 1]	382	292
6	[−23, −85, −9]	[27, −84, −0]	259	474
7	[−24, −90, −3]	[23, −88, −5]	646	703
8	[−22, −83, −16]	[25, −88, −7]	432	348
9	[−18, −89, −3]	[22, −86, −6]	442	318
10	[−23, −87, −16]	[24, −84, −8]	366	630
		V3AB		
1	[−26, −84, 7]	[28, −83, 7]	136	200
2	[−28, −85, 19]	[23, −80, 30]	425	253
3	[−25, −84, 21]	[28, −85, 22]	269	214
4	[−26, −83, 11]	[26, −77, 19]	141	167
5	[−26, −87, 17]	[34, −79, 19]	281	396
6	[−31, −86, 12]	[31, −78, 20]	120	137
7	[−29, −83, 12]	[24, −78, 15]	357	306
8	[−36, −84, 5]	[36, −83, 1]	350	474
9	[−24, −89, 18]	[30, −83, 10]	341	329
10	[−25, −91, −2]	[28, −78, 18]	450	311
		V4		
1	[−27, −75, −16]	[26, −75, −14]	116	245
2	[−29, −76, −8]	[25, −76, −8]	336	299
3	[−29, −78, −7]	[22, −78, −9]	303	154
4	[−22, −80, −13]	[25, −75, −9]	75	158
5	[−27, −79, −16]	[32, −75, −12]	243	139
6	[−29, −75, −17]	[32, −72, −14]	115	272
7	[−29, −76, −14]	[29, −76, −14]	406	380
8	[−31, −75, −23]	[30, −83, −18]	261	164
9	[−24, −76, −18]	[26, −77, −16]	304	362
10	[−32, −76, −22]	[24, −75, −17]	243	329
		LO		
1	[−43, −76, −13]	[37, −78, −10]	617	798
2	[−40, −77, 3]	[37, −78, 5]	911	1113
3	[−38, −81, 5]	[35, −81, 1]	253	314
4	[−33, −87, 1]	[34, −80, 0]	163	177
5	[−35, −84, 4]	[38, −81, 5]	207	132
6	[−40,−78, −7]	[40,−73, −3]	933	724
7	[−35,−88, −4]	[32,−82, −1]	372	449
8	[−41,−73,−13]	[38,−76,−14]	812	347
9	[−34,−80, −5]	[37,−78, −7]	792	634
10	[−47,−76,−14]	[35,−81, 2]	119	77
		pFus		
1	[−34,−75,−16]	[34,−51,−15]	215	54
2	[−37,−63,−10]	[36,−62, −9]	354	415
3	[−35,−69, −8]	[34,−70, −9]	233	98
4	[−37,−80,−10]	[36,−67, −8]	252	341
5	[−43,−66, −8]	[40,−64,−10]	201	127
6	[−37,−63,−16]	[33,−55,−15]	234	268
7	[−38,−68,−15]	[34,−61,−17]	87	184
8	[−37,−54,−21]	[32,−62,−20]	391	656
9	[−33,−63,−12]	[35,−61,−15]	96	447
10	[−45,−67,−12]	[39,−61,−15]	207	117
		IPS0–1		
1	[−26,−80, 15]	[25,−79, 21]	162	236
2	[−31,−78, 27]	[22,−61, 40]	369	960
3	[−22,−70, 32]	[24,−75, 33]	331	280
4	[−22,−70, 28]	[25,−67, 33]	185	175
5	[−25,−70, 26]	[30,−67, 30]	524	397
6	[−30,−67, 26]	[22,−64, 40]	30	269
7	[−25,−76, 25]	[24,−73, 28]	478	586
8	[−34,−74, 15]	[32,−71, 13]	401	375
9	[−24,−75, 24]	[30,−74, 16]	675	482
10	[−27,−88, 9]	[26,−67, 33]	473	499
		IPS2–3		
1	[−29,−66, 32]	[25,−72, 39]	140	126
2	[−25,−63, 35]	[24,−46, 51]	488	491
3	[−25,−54, 46]	[22,−61, 44]	283	203
4	[−24,−62, 45]	[24,−60, 39]	135	47
5	[−27,−63, 43]	[22,−63, 43]	368	270
6	[−26,−62, 35]	[24,−54, 45]	221	175
7	[−22,−71, 39]	[24,−61, 37]	395	352
8	[−26,−68, 27]	[27,−63, 23]	214	161
9	[−29,−62, 31]	[25,−63, 29]	283	388
10	[−27,−71, 23]	[25,−55, 45]	390	510
		sIPS		
1	[−34,−55, 36]	[34,−54, 37]	65	204
2	[−34,−40, 42]	[31,−47, 50]	155	283
3	[−31,−50, 43]	[26,−60, 45]	203	145
4	[−24,−62, 43]	[29,−55, 44]	190	559
5	[−32,−53, 47]	[34,−50, 45]	472	708
6	[−35,−48, 43]	[27,−51, 48]	527	434
7	[−35,−54, 41]	[33,−52, 44]	335	919
8	[−28,−60, 45]	[28,−59, 41]	540	654
9	[−34,−53, 42]	[32,−54, 40]	486	805
10	[−28,−55, 41]	[31,−51, 44]	1057	665
		sPCS		
1				
2	[−22, 4, 60]	[30, 2, 54]	4	28
3	[−25, −4, 54]	[29, 0, 53]	35	19
4	[−31, −4, 50]	[36, −3, 49]	5	52
5	[−29, −4, 57]	[29, −4, 52]	56	142
6	[−33, 2, 54]	[36, 2, 55]	152	172
7	[−28, −5, 54]	[29, −1, 55]	19	348
8	[−37, −1, 56]	[29, 1, 53]	19	349
9	[−27, −3, 52]	[27, −3, 50]	189	209
10	[−25, −3, 56]	[26, 2, 55]	95	109

Table 1. *Continued*

Subject	Center		No. of Voxels	
	LH	RH	LH	RH
		iPCS		
1	[−47, 10,34]	[47, 6, 35]	169	222
2	[−45, 7,35]	[46, 8, 37]	25	44
3	[−48, 8,23]	[47, 8, 29]	103	121
4	[−34,−2,32]	[42, 6, 35]	30	217
5	[−46, 2,32]	[44, 4, 28]	66	135
6	[−47, 6,29]	[46, 7, 28]	216	150
7	[−42,−1,33]	[41, 6, 32]	150	422
8	[−46, 8,30]	[50, 5, 32]	124	154
9	[−43, 5,33]	[42, 5, 34]	472	483
10	[−45, 1,30]	[45, 6, 33]	130	326
		AI/FO		
1	[−35, 19, 4]	[37,17, 5]	244	532
2	[−37, 17, 3]	[38,18, 1]	267	340
3	[−35, 15, 2]	[32,18, 5]	120	115
4	[−37, 18, 4]	[37,18, 3]	354	510
5	[−34, 20, 1]	[33,19,−1]	47	34
6	[−38, 18, 2]	[37,20, 2]	289	238
7	[−35, 17, 2]	[35,18, 2]	145	388
8	[−36, 16, 4]	[35,17, 3]	218	485
9	[−37, 19, 4]	[37,17, 5]	486	551
10	[−37, 19, 3]	[33,20, 5]	53	150
		IFS		
1	[−37, 35,28]	[40,30, 28]	9	46
2	[−29, 47,17]	[34,37, 20]	47	235
3	[−44, 33,22]	[42,31, 22]	44	55
4	[−42, 29,23]	[38,30, 23]	5	96
5	[−28, 49,21]	[34,41, 18]	53	105
6	[−35, 42,26]	[41,39, 26]	105	85
7	[−40, 36,19]	[40,40, 21]	152	460
8	[−40, 33,18]	[41,33, 22]	71	308
9	[−41, 34,26]	[44,30, 27]	242	189
10	[−39, 37,21]	[38,43, 24]	33	102

Data are centers and sizes of final regions of interest (ROIs) in early visual cortex (V1, V2, V3, V3AB, V4), lateral occipital complex (LO, pFus), intraparietal sulcus [IPS0–1, IPS2–3, superior IPS (sIPS)], and the multiple-demand network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)] for each subject (see METHODS for details). Coordinates of each ROI center are described in Talairach space, where X = left-right axis (negative is left), Y = anterior-posterior axis (negative is posterior), and Z = inferior-superior axis (negative is inferior).

task relevance of each match dimension, as well as how it differed on correct and incorrect trials.

Several ROIs did show a significant difference in mean signal between the identity and viewpoint tasks (data not shown). Therefore, before performing MVPA, we first mean-centered the voxel activation pattern on each trial by calculating the mean across voxels on each trial and subtracting this value from the voxel activation pattern. This ensured that classification was based on information encoded in the relative pattern of activity across voxels in each ROI, rather than information about mean signal changes across conditions.

We performed all decoding analyses using a binary classifier based on the normalized Euclidean distance. To avoid overfitting, we used a leave-one-run-out cross-validation scheme so that each run served as the test set once. Before starting this analysis, we removed all trials that were the first in a block, because they could not be labeled as a match or nonmatch. Next, we divided data in the training set into two groups based on status as a match in the dimension of interest. For each of these two groups, we then calculated a mean voxel activation pattern (e.g., averaging the response of each voxel over all trials in the group). We also calculated the pooled variance of each voxel's response across the two groups. Next, for each trial in the test set, we calculated the normalized Euclidean distance to each of the mean patterns of the training set groups, weighting each voxel's contribution on the basis of its pooled variance. We then assigned each test set trial to the group with the minimum normalized Euclidean distance. Specifically, for a training set including n total trials, with n_a trials in *condition A* and n_b trials in *condition B*, and v voxels in each activation pattern, we can define $\bar{a}$, σ_a^2, $\bar{b}$, and σ_b^2 as vectors of size [1xv] describing the mean and variance of each voxel's response within *conditions A* and *B*, respectively. If x is a [1xv] vector describing a voxel activation pattern from a single trial in the test set, the normalized Euclidean distance from x to each of the two training set conditions is

$$d_{x\to a} = \sqrt{\sum_{i=1}^{v}\left(\frac{x_i - \bar{a}_i}{\sigma_{p_i}^2}\right)^2}$$

$$d_{x\to b} = \sqrt{\sum_{i=1}^{v}\left(\frac{x_i - \bar{b}_i}{\sigma_{p_i}^2}\right)^2}$$

where σ_p^2 is a [1xv] vector describing the pooled variance of each voxel over *conditions A* and *B*:

$$\sigma_p^2 = \frac{n_a\sigma_a^2 + n_b\sigma_b^2}{n_a + n_b}$$

The final label assigned to each test set trial by the classifier was obtained by finding the minimum value between $d_{x\to a}$ and $d_{x\to b}$. Finally, we computed a single value for classifier performance across the entire data set by calculating d' with the formula

$$d' = Z(hit\ \ rate) - Z(false\ \ positive\ \ rate),$$

where the hit rate is defined as the proportion of test samples in *condition A* accurately classified as belonging to *condition A*, and the false positive rate is the proportion of test samples in *condition B* inaccurately classified as belonging to *condition A*. The function $Z(p)$, $p \in [0,1]$ is the inverse of the cumulative distribution of the Gaussian distribution.

Because the frequency of matches in our task was less than 50%, the training set for the classifier was initially unbalanced. To correct for this, we performed downsampling on the larger training set group (nonmatch trials) by randomly sampling N trials without replacement from the larger set, where N is the number of samples in the smaller set. We performed 1,000 iterations of this random downsampling and averaged the results for d' over all iterations.

We assessed the significance of classifier decoding performance in each ROI using a permutation test in which we shuffled the labels of all trials in the training set and computed decoding performance on this shuffled data set. We repeated this procedure over 1,000 iterations to compute a null distribution of d' for each subject and each ROI. For each shuffling iteration, we performed downsampling to balance the training set as described above, but to reduce the computational time, we used only 100 iterations. To compute significance at the group level, we averaged the null distributions over all subjects to obtain a single distribution of 1,000 d' values, and averaged the d' values for the real data set over all subjects to obtain a subject-average d' value. We obtained a P value by calculating the proportion of shuffling iterations on which the shuffled d' value exceeded the real d' value, and the proportion on which the real d' value exceeded the shuffled d' value, and taking the minimum value multiplied by 2. We then performed FDR correction across ROIs within each condition and match type, at the 0.01 and 0.05 significance levels (Benjamini and Yekutieli 2001).

The above analysis was carried out separately within each ROI, task, and match dimension separately to estimate information about viewpoint and identity match status when each dimension was rele-

vant and irrelevant. Next, we entered all d' values into a three-way repeated-measures ANOVA with factors of task, ROI, and relevance. Following this, to more closely investigate interactions between ROI and relevance, we used nonparametric paired *t*-tests to compare the d' distributions for the relevant match dimension vs. the irrelevant match dimension, within each task and ROI separately. This test consisted of performing 10,000 iterations in which we randomly permuted the relevance labels corresponding to the d' values, maintaining the subject labels. After randomly permuting the labels, we calculated the difference in d' between the two conditions for each subject and used these 10 difference values to calculate a *t*-statistic. We then compared the distribution of these null *t*-statistics with the value of the *t*-statistic found with the real relevance labels and used this to generate a two-tailed *P* value. These *P* values were FDR corrected across ROIs at both the 0.05 and 0.01 levels.

In the first set of analyses (see Figs. 5 and 6), which focused on the overall performance of the classifier, we performed the above steps after removing all trials where the subject was incorrect or did not respond (on average, 7% of trials were no-response trials). In the next set of analyses (see Figs. 8 and 9), we were interested in whether information in each ROI about the task-relevant match dimension was associated with task performance. To evaluate this, we included incorrect and no-response trials in both the training and testing sets. We considered all incorrect and no-response trials as a single group, which we refer to as "incorrect." For each trial in the test set, we then used the normalized Euclidean distance (calculations described above) as a metric of classifier evidence in favor of the actual trial label, where

$$d_{x\rightarrow incorrect} - d_{x\rightarrow correct} = evidence.$$

We then compared the distributions of evidence between correct and incorrect test set trials. Because there were many more correct than incorrect test set trials, we performed an additional step of downsampling to balance the training set. We divided training set trials into four groups, based on their status as a match and the correctness of the subject's response, and downsampled the number of trials in all four groups to match the number of trials in the smallest set. This was performed over 1,000 iterations, and the resulting values for classifier evidence were averaged. We assessed significance of the difference between correct and incorrect trials using a permutation test as described above.

In the main analyses described above, we performed MVPA using all voxels from each localized ROI. Additionally, to control for differences in the number of voxels between ROIs, we repeated all analyses after restricting the number of voxels to 50 in each area. Overall, reducing the number of voxels did not lead to a dramatic change in the patterns of decoding performance across ROIs. These analyses are all reported in the figures.

Experimental design and statistical analyses. The sample size for this experiment was 10, with subjects run in 1 or 2 scanning sessions to collect at least 16–22 runs of experimental data, as well as between 2 and 6 runs of each functional localizer described above. This sample size was determined before data collection was started, based on sample sizes used by past experiments with similar methodology in our laboratory. For details of MVPA analyses and related statistics, see *MVPA decoding*. Briefly, all statistical tests, including repeated-measures ANOVAs, were performed using MATLAB R2017a (The MathWorks, Natick, MA) and were based on within-subject factors. The significance of MVPA results was assessed using permutation testing, with the final test for significance performed across all subjects. Pairwise comparisons of classifier output (e.g., classifier evidence on correct vs. incorrect trials, decoding d' for task-relevant vs. task-irrelevant match status) were performed using a nonparametric permutation-based *t*-test. Multiple comparisons correction was performed using FDR correction as described in Benjamini and Yekutieli (2001). We chose two thresholds because the 0.05 value provides slightly more power to detect weaker effects, whereas the 0.01 value provides a more conservative threshold.

Image similarity analysis. The viewpoint and identity tasks were designed so that the low-level shape similarity of the images would not be explicitly informative about the status of each image as a match. Thus, when we performed classification on the status of each image as a target in each dimension, we intended to capture information that was related to perception of the abstract dimensions of each object, rather than low-level properties such as its shape in a 2D projection. However, because of factors such as the small number of objects in our stimulus set, it is possible that there was some coincidental, systematic structure in the similarity between pairs of objects such that low-level image similarity was partially informative about the status of each image as a match in either identity or viewpoint.

To evaluate this possibility, for each trial in the sequence of images shown to each subject, we determined the image similarity between the current and previous images by unwrapping each image (1,000 × 1,000 pixels) into a single vector and calculating the Pearson correlation coefficient between each pair of vectors. For this analysis, we removed trials that were a match in both category and viewpoint (identical images by design) and sorted the remaining trials according to whether the current and previous images were actually a match in the dimension of interest. Mean image similarity between match and nonmatch trials was compared using a one-tailed *t*-test (see Fig. 6D). This resulted in a *P* value for each subject in each condition, which was used to evaluate the extent to which low-level image similarity may have been informative about match status.

In addition to using the Pearson correlation to measure similarity, we also assessed similarity by passing each image through a Gabor wavelet model meant to simulate the responses of V1 neurons to the spatial frequency and orientation content of the image (Pinto et al. 2008). We then compared the effectiveness of this V1 model and the simpler pixel model at capturing the responses in V1 from our fMRI data, by calculating a similarity matrix for each pairwise comparison of images (24 × 24), based on *1*) the V1 model, *2*) the pixel model, and *3*) the voxel activation patterns recorded in V1 for each subject. We found that across all subjects, the pixel model was more correlated with the V1 voxel responses than was the V1 model (data not shown). Therefore, in the interest of parsimony, we used the simpler pixel model for all image similarity analyses.

RESULTS

Behavioral performance. Subjects ($n = 10$) performed alternating runs of the identity task and the viewpoint task while undergoing fMRI. Runs were always presented in matched pairs so that the object sequence and visual stimulation were identical between runs of the identity task and viewpoint task. We found no significant difference in performance (d' for identity: 1.47 ± 0.24, d' for viewpoint: 1.81 ± 0.34; paired 2-tailed *t*-test, $P = 0.2054$; Fig. 3*A*;) and no significant difference in response times (RT for identity: 1.15 ± 0.06 s, RT for viewpoint: 1.17 ± 0.05 s; paired 2-tailed *t*-test, $P = 0.6124$; Fig. 3*B*) on the two tasks across subjects. Performance and response time for each task also did not differ as a function of the object set subjects had been assigned to (d' for identity, *set A*: 1.50 ± 0.22, d' for identity, *set B*: 1.43 ± 0.27, $P = 0.8930$; d' for viewpoint, *set A*: 1.76 ± 0.30, d' for viewpoint, *set B*: 1.86 ± 0.42, $P = 0.9024$; RT for identity, *set A*: 1.11 ± 0.04 s, RT for identity, *set B*: 1.19 ± 0.07 s, $P = 0.5343$; RT for viewpoint, *set A*: 1.18 ± 0.02, RT for viewpoint, *set B*: 1.16 ± 0.08, $P = 0.8843$; all are 2-tailed *t*-tests).

Univariate BOLD signal does not show match suppression. First, we examined whether the status of each image as a match in the task-relevant dimension (e.g., identity match status

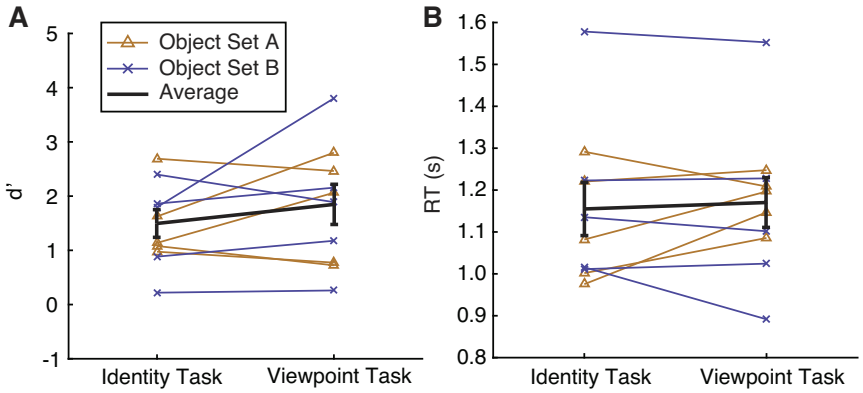


Fig. 3. Behavioral performance (d'; *A*) and response time (RT; *B*) were similar across tasks and stimulus sets. Each line represents performance of a single subject, averaged over runs of each task. Error bars are ±SE.

during the identity task; viewpoint match status during the viewpoint task) was reflected in a change in the mean amplitude of BOLD response in any visual area. On the basis of previous fMRI studies (Grill-Spector et al. 1999; Henson 2003) and electrophysiology studies (Meyer and Rust 2018; Miller et al. 1991), we predicted that the repetition of object identity or viewpoint might result in response suppression (often referred to as repetition suppression). However, in all but 1 of the 14 ROIs we examined (Fig. 4), we found no significant difference in the mean signal amplitude between matches and nonmatches. In the one ROI showing a significant effect (AI/FO), mean signal was actually higher on match trials than on nonmatch trials. We suspect that the absence of repetition suppression is due to differences in task demands between our experiment and previous work, particularly the fact that identity and viewpoint matches were task relevant in our paradigm and thus may have evoked larger attention-modulated responses that counteracted any repetition suppression effects (see DISCUSSION for details). Note that in all subsequent multivariate analyses, we first de-meaned responses across all voxels in each ROI so that single-trial voxel activation patterns were centered at zero and any small univariate effects could not contribute to decoding performance (see METHODS).

Multivariate activation patterns reflect task-relevant match status. Next, we examined how voxel activation patterns in each ROI reflected the status of a stimulus as a match in viewpoint and identity, and how representations of viewpoint and identity match status were influenced by the task relevance of each dimension. During each task, a correct behavioral response depended on the object being a match to the previous object in the relevant dimension (identity or viewpoint), whereas match status in the other dimension was irrelevant. Therefore, we expected that information about the match status of an object in each dimension would be more strongly represented when that dimension was task relevant.

Indeed, status as a match along the task-relevant dimension, measured by classifier performance (d'), was represented widely within the ROIs we examined, whereas the irrelevant match was not represented at an above-chance level in any ROI (Fig. 5). Information about the task-relevant match increased along a posterior-to-anterior axis such that match status was represented most strongly in MD and IPS ROIs but was comparatively weaker in early visual cortex and the lateral occipital complex (LOC). Relevant match decoding performance was above chance for all MD ROIs for both the identity and viewpoint tasks, and was also above chance for LO, V3AB, and V2 for both tasks. Decoding performance was above chance in V1, V4, and pFus for the identity task only.

The general pattern of decoding performance was similar across tasks, although there was a trend toward higher relevant match decoding performance in IPS for the viewpoint task than for the identity task. There was also an opposite trend in V4 and pFus for higher relevant match decoding during the identity task than during the viewpoint task. A three-way repeated-measures ANOVA with factors of task, ROI, and relevance revealed a main effect of relevance [$F(1,9) = 46.219$, $P = 10^{-4}$], a main effect of ROI [$F(13,117) = 9.930$, $P < 10^{-12}$], and a relevance × ROI interaction [$F(13,117) = 13.981$, $P < 10^{-17}$], but no main effect of task [$F(1,9) = 1.819$, $P = 0.2104$]. The interactions task × ROI [$F(13,117) = 1.838$, $P = 0.0450$] and task × relevance [$F(1,9) = 1.019$, $P = 0.3392$] were not significant at $\alpha = 0.01$. We further investigated the ROI × relevance interaction, using paired *t*-tests to compare decoding of the relevant and irrelevant dimensions, and found that the effect of relevance was significant for all MD regions in both tasks, LO in both tasks, as well as V3AB, V4, and pFus for the identity task only (Fig. 5). These results were similar when we used all voxels in each area (Fig. 5, *A* and *B*) and when we used 50 voxels in each area (Fig. 5, *C* and *D*).

Because our MD localizer failed to yield any voxels in the sPCS ROI for one subject, we used an interpolation method to fill in this value before performing the repeated-measures ANOVA (see METHODS for details). As an alternative method, we also performed the same test after removing all data from the subject that was missing sPCS, and we found similar results (Table 2). We note, however, that the task × ROI interaction term, which was not significant when the interpolation method was used, was significant when the missing subject was removed entirely [$F(13,104) = 2.530$, $P = 0.0047$].

Control analyses for visually driven match representations. To perform the identity and viewpoint tasks, subjects were required to use representations of object identity and viewpoint that were largely invariant to shape. However, a subset of trials in each task could, in principle, be solved based only on shape similarity between the previous and current objects. In the viewpoint task, because subtle differences between exemplars were not task relevant, the group of trials that could be solved based only on shape similarity included all trials that were matches in both category and viewpoint. In the identity task,

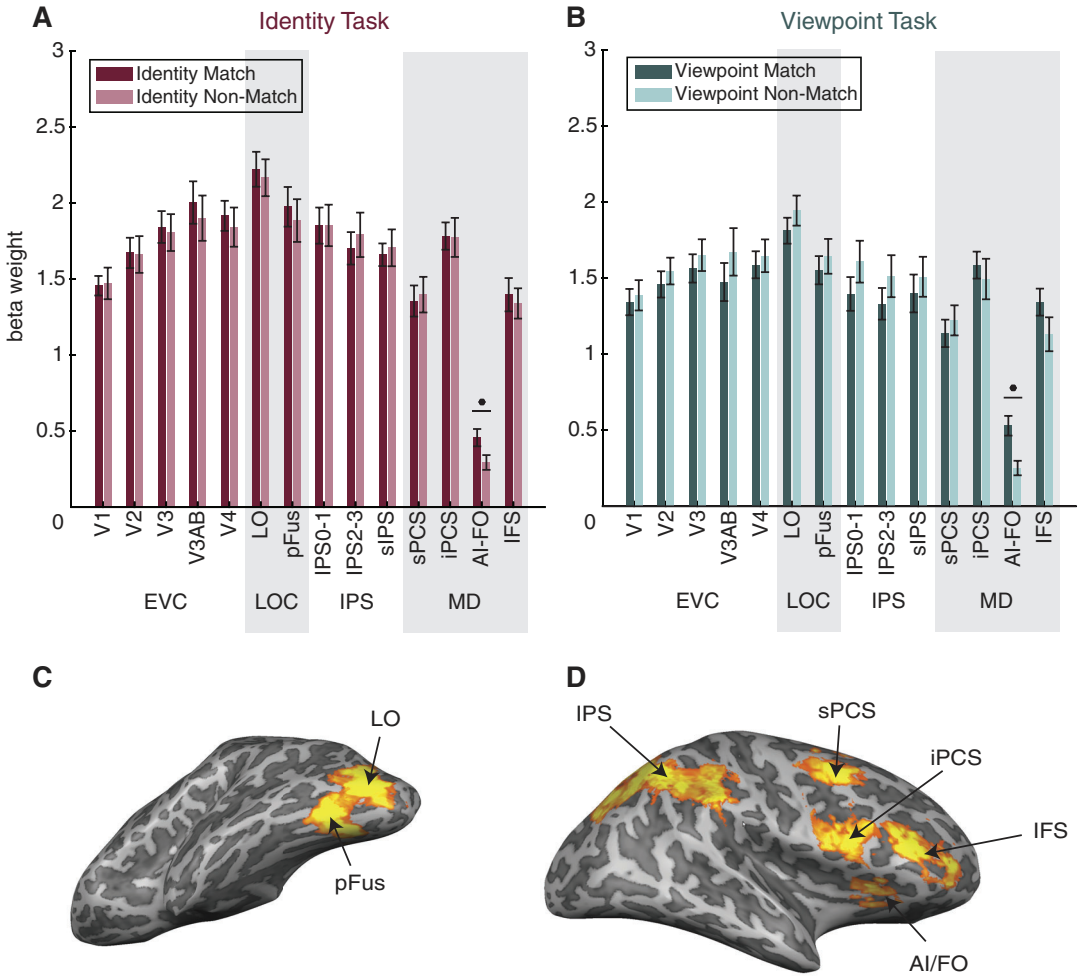


Fig. 4. Match status in the current task cannot be determined solely from mean signal change. *A* and *B*: mean beta weights are plotted on the *y*-axis, and individual regions of interest (ROIs) are plotted on the *x*-axis. In all functional MRI analyses, we used 14 ROIs: early visual cortex (EVC; comprising V1, V2, V3, V3AB, and V4), lateral occipital complex (LOC; comprising LO and pFus), intraparietal sulcus [IPS; comprising IPS0–1, IPS2–3, and superior IPS (sIPS)], and the multiple-demand (MD) network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)], all of which were defined using functional localizers (see METHODS). ROIs are organized into 4 groups for convenience. Note that although we have visually separated the groups, we consider IPS to be a part of the MD network. In the identity task (*A*), in all but one ROI, univariate activation (mean beta weight) did not differ between trials according to their status as an identity match. Similarly, in the viewpoint task (*B*), univariate activation in most ROIs did not reflect viewpoint match status. Match and nonmatch trials were compared using paired *t*-tests. *P* values were FDR-corrected over all conditions. Circles indicate significance at $q = 0.01$. Error bars are ±SE. *C*: example definitions of the LOC ROIs shown on an inflated mesh (left hemisphere of *subject 1*). *D*: example definitions of the MD ROIs shown on an inflated mesh (right hemisphere of *subject 1*). Subdivisions of IPS are not shown. For a summary of the definitions of all ROIs in all subjects, see Table 1.

the group of trials that could be solved on the basis of shape similarity included all trials that were an exact match to the previous image in category, exemplar, and viewpoint. Therefore, it is possible that some of the match status classification we observed in Fig. 5 was driven by the detection of shape similarity. To test for this possibility, we removed all trials that were a match in both category and viewpoint, leaving a set of trials in which the shape similarity was entirely uninformative about match status. We then performed classification on this reduced data set as before.

For the viewpoint task, we now found an important difference between visual and MD regions: whereas decoding of viewpoint matches remained above chance in all MD and IPS regions, it dropped to chance in LOC and early visual cortex (Fig. 6*B*). Thus, whereas early visual and LOC representations of viewpoint match status appeared to rely largely on low-level shape similarity, MD regions encoded viewpoint match status even when shape similarity could not be used to define a match.

During the identity task, however, even after removal of all trials that were a match in both category and viewpoint, decoding of identity match status remained above chance in all ROIs examined (Fig. 6*A*). The observation of above-chance identity match decoding in early visual areas was surprising given that these areas are not expected to encode abstract, viewpoint-independent representations of target status. There-

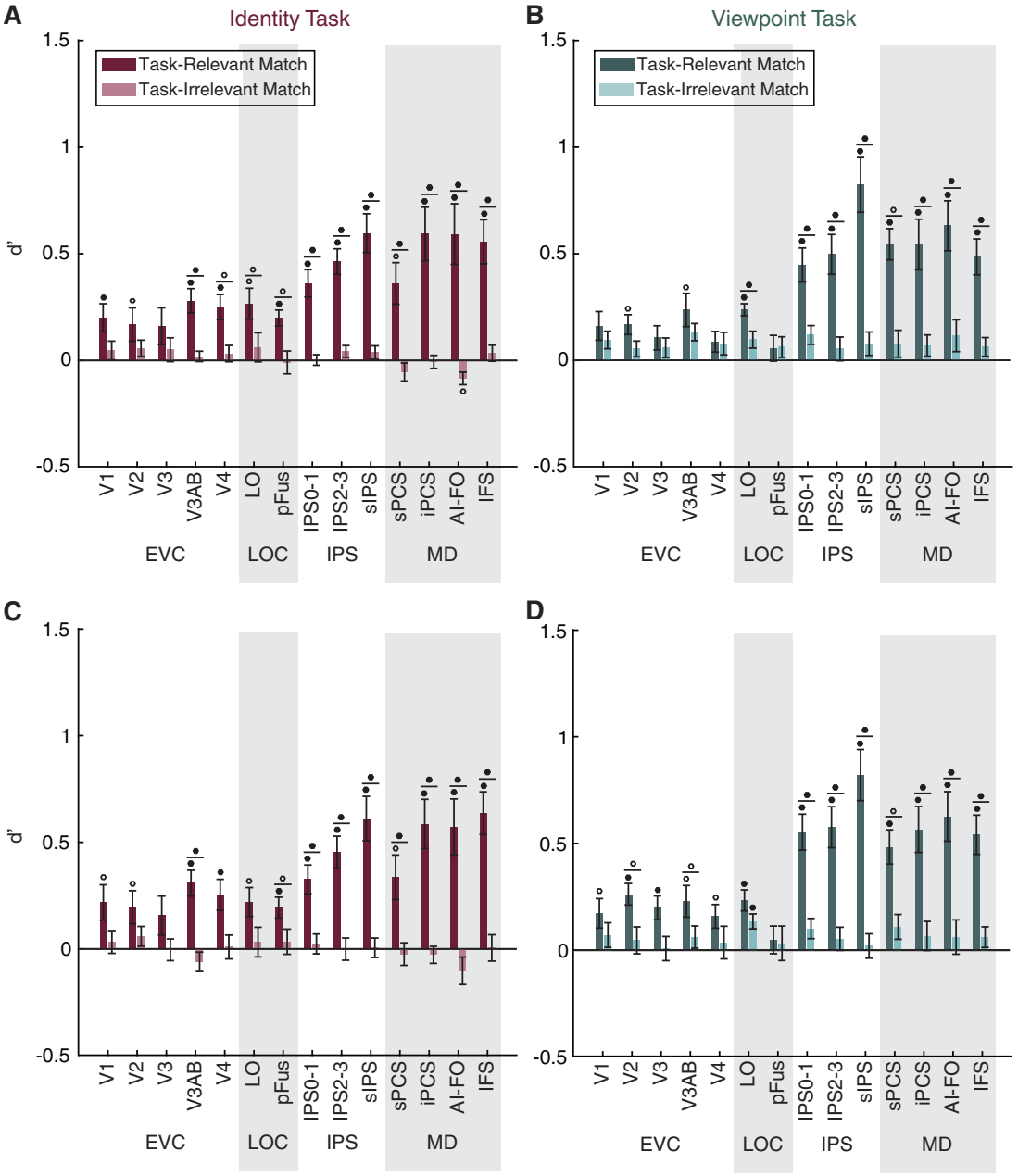


Fig. 5. Task-relevant matches are represented more strongly than task-irrelevant matches. *A–D*: a linear classifier was trained to discriminate between voxel activation patterns measured during runs of the identity task (*A* and *C*) or the viewpoint task (*B* and *D*), according to whether the viewed image was a match in the task-relevant or the task-irrelevant dimension. Classifier performance (d') is plotted on the *y*-axis for 14 regions of interest (ROIs): early visual cortex (EVC; comprising V1, V2, V3, V3AB, and V4), lateral occipital complex (LOC; comprising LO and pFus), intraparietal sulcus [IPS; comprising IPS0–1, IPS2–3, and superior IPS (sIPS)], and the multiple-demand (MD) network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)]. Circles above individual bars indicate above-chance classification performance (test against 0); circles above pairs of bars (denoted by horizontal lines) indicate significant differences between bars (paired *t*-test). *P* values were FDR-corrected over all conditions. Open circles indicate significance at $q = 0.05$; closed circles indicate significance at $q = 0.01$. Error bars are ±SE. *A* and *B* show decoding performance using all voxels in each ROI; *C* and *D* show decoding performance using only 50 voxels in each ROI.

fore, we wondered whether the remaining set of images may have had some shared features that supported match classification. Indeed, when we assessed the pixelwise similarity between images (see METHODS) belonging to each stimulus set, we found that in four of the subjects assigned to object *set A*, pixelwise similarity between images was significantly predictive of identity match status (Fig. 6*D*). We thus hypothesized that the above-chance decoding of identity match status ob-

Table 2. *Results of three-way repeated-measures ANOVA on decoding results*

	SumSq	df	MeanSq	*F* Statistic	*P* Value
		RM ANOVA, using interpolation			
(Intercept)	23.3373	1	23.3373	55.1700	**$<10^{-4}$**
Error	3.8071	9	0.4230		
(Intercept):ROI	4.7641	13	0.3665	9.9297	**$<10^{-12}$**
Error(ROI)	4.3180	117	0.0369		
(Intercept):Task	0.1433	1	0.1433	1.8193	0.2104
Error(Task)	0.7090	9	0.0788		
(Intercept):Relevance	13.4705	1	13.4705	46.2189	**0.0001**
Error(Relevance)	2.6230	9	0.2914		
(Intercept):ROI:Task	0.6030	13	0.0464	1.8375	0.0450
Error(ROI:Task)	2.9535	117	0.0252		
(Intercept):ROI:Relevance	5.5957	13	0.4304	13.9806	**$<10^{-17}$**
Error(ROI:Relevance)	3.6022	117	0.0308		
(Intercept):Task:Relevance	0.1618	1	0.1618	1.0185	0.3392
Error(Task:Relevance)	1.4296	9	0.1588		
(Intercept):ROI:Task:Relevance	0.3903	13	0.0300	1.0157	0.4411
Error(ROI:Task:Relevance)	3.4585	117	0.0296		
		RM ANOVA, with subject 1 removed			
(Intercept)	22.4745	1	22.4745	50.5309	**0.0001**
Error	3.5581	8	0.4448		
(Intercept):ROI	4.7895	13	0.3684	9.9486	**$<10^{-12}$**
Error(ROI)	3.8514	104	0.0370		
(Intercept):Task	0.1458	1	0.1458	1.6572	0.2340
Error(Task)	0.7039	8	0.0880		
(Intercept):Relevance	13.2417	1	13.2417	44.5775	**0.0002**
Error(Relevance)	2.3764	8	0.2970		
(Intercept):ROI:Task	0.7977	13	0.0614	2.5299	**0.0047**
Error(ROI:Task)	2.5227	104	0.0243		
(Intercept):ROI:Relevance	5.0628	13	0.3894	13.4838	**$<10^{-15}$**
Error(ROI:Relevance)	3.0038	104	0.0289		
(Intercept):Task:Relevance	0.0959	1	0.0959	0.5569	0.4769
Error(Task:Relevance)	1.3779	8	0.1722		
(Intercept):ROI:Task:Relevance	0.4434	13	0.0341	1.0929	0.3734
Error(ROI:Task:Relevance)	3.2455	104	0.0312		

Data are the results of a 3-way repeated-measures (RM) ANOVA with the factors region of interest (ROI), task, and relevance, performed on the decoding results shown in Fig. 5, *A* and *B*. We were unable to define the superior precentral sulcus (sPCS) ROI in 1 of 10 subjects and used two different methods to address this missing value before running the RM ANOVA. Results are shown for both an interpolation method (see METHODS for details) and with all data removed corresponding to the subject who was missing sPCS (using 9/10 subjects). Bold values indicate *P* values that were significant at $\alpha = 0.01$. For both tests, Mauchly's test revealed that the data did not violate the assumption of sphericity, so we report uncorrected *P* values. df, degrees of freedom; MeanSq, mean square; SumSq, sum of squares.

served in early visual areas might be driven by this group of subjects.

In line with this prediction, when we reanalyzed decoding performance using only subjects from *set B*, identity match decoding was no longer significant in V1, V2, V3, V4, and pFus, even though it was still well above chance in other parietal and MD areas. This pattern suggests that the above-chance decoding accuracy for identity match status in these earlier visual ROIs may have been related to low-level image features (Fig. 6*C*). In contrast, even after this additional source of low-level image similarity was removed, MD and IPS regions still encoded robust representations of identity match status. Furthermore, match decoding in the MD regions was individually strong in the majority of subjects (Fig. 7).

Behavioral performance is closely linked to activation patterns in the MD network. Having established that several ROIs, including all MD network ROIs, represent the status of an object as a match in the task-relevant dimension, we next sought to determine whether these representations were related to behavioral performance. To answer this question, we used the normalized Euclidean distance to calculate a continuous measure of the classifier's evidence at predicting the correct label for each trial in the test set (see METHODS). We then calculated the mean classifier evidence for all correct and incorrect trials. Because the previous analysis (Fig. 5) indicated no significant effect of task on relevant match decoding, we combined all trials across the identity and viewpoint tasks for this analysis (Fig. 8).

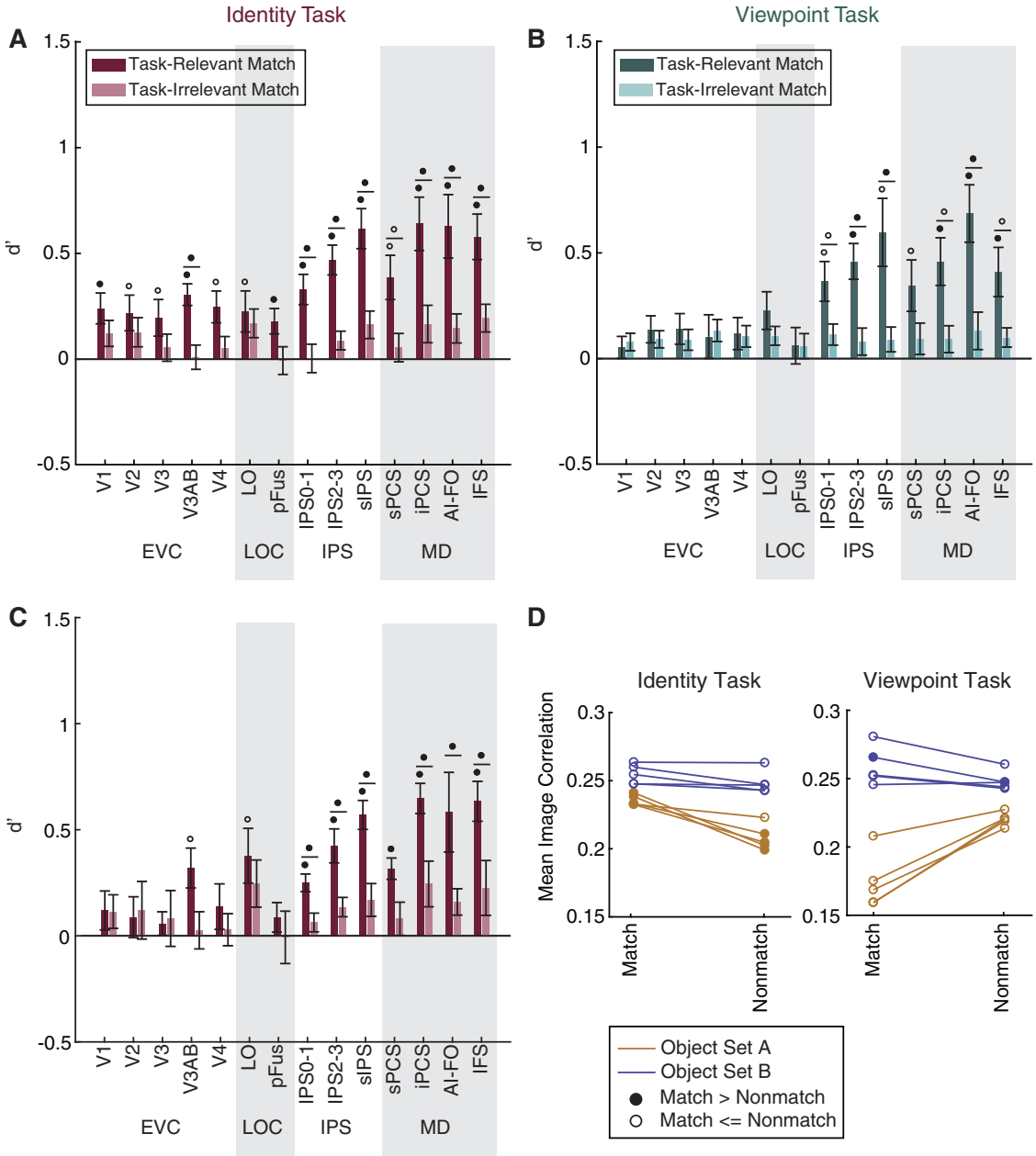
In all regions of the MD network and IPS, we found that classifier evidence was significantly higher on correct than incorrect trials. In contrast, we found no significant effect of behavior on classifier evidence in early visual cortex or in the LOC. Thus representations in IPS and MD ROIs, but not any of the other regions that we evaluated, were significantly associated with task performance. We also verified that this effect was similar using data from each task separately (Fig. 9) and when voxel number was matched across ROIs (Fig. 8*B* and Fig. 9, *C* and *D*).

DISCUSSION

In this study we used fMRI and pattern classification methods to investigate the role of different brain areas in signaling task-relevant matches across identity-preserving transforma-

tions. Specifically, subjects performed a task that required them to identify matches in either the identity or the viewpoint of novel objects. Consistent with previous work, we found that areas of ventral visual cortex represent information about the status of an object as a relevant match (Fig. 5; Lueschow et al. 1994; Miller and Desimone 1994; Pagan et al. 2013; Woloszyn and Sheinberg 2009). However, our results also suggest a key role for the MD network in this match identification task, with MD match representations showing specificity to task-relevant dimensions (Fig. 5), as well as invariance to low-level visual features (Fig. 6). Importantly, the present data also establish a significant link between task performance and the strength of MD match representations (Fig. 8). In contrast, in early visual and ventral object-selective cortex, we found comparatively weaker evidence for representations of match status and no significant associations with task performance. These results suggest that MD regions play a key role in computing flexible and abstract target representations that are likely to be important for task performance.

In contrast to our MVPA results, univariate signal amplitude in almost all ROIs was not significantly modulated by match status. This finding differs from many past fMRI



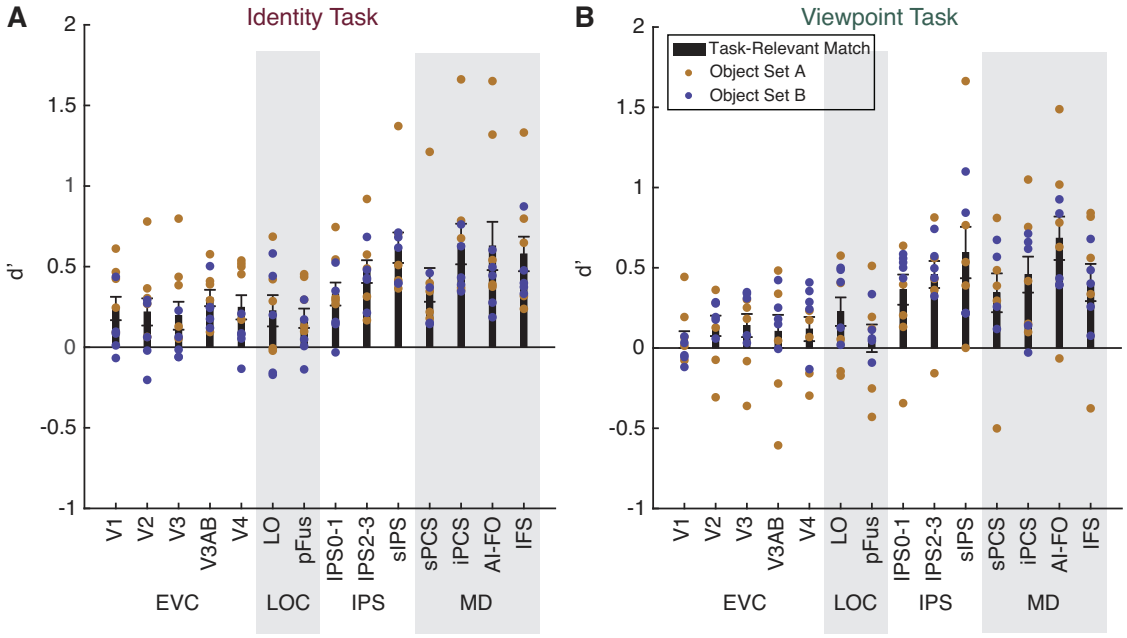


Fig. 7. After stimulus similarity confounds were addressed, most individual subjects in both stimulus sets still show above-chance match decoding in multiple-demand (MD) network regions of interest (ROIs): superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS). Relevant match decoding performance was calculated after removal of all trials in which the current and previous objects had a high degree of shape similarity (same analysis as Fig. 6, *A* and *B*). Black bars show the subject average ± SE and are identical to the dark colored bars in Fig. 6, *A* and *B*. Colored dots indicate individual subject decoding performance, with different stimulus sets shown in different colors. EVC, early visual cortex (comprising V1, V2, V3, V3AB, and V4); LOC, lateral occipital complex (comprising LO and pFus); IPS, intraparietal sulcus [comprising IPS0–1, IPS2–3, and superior IPS (sIPS)].

studies (Grill-Spector et al. 1999; Henson 2003; Turk-Browne et al. 2007) that have observed response suppression as a result of object repetition (or "repetition suppression"). One explanation for this divergence in findings is that in our task, the repetition of identity and viewpoint is task relevant, meaning that repetition-related signals are mixed with signals related to task performance. This account is consistent with prior electrophysiology studies reporting that when an object's match status is task relevant, neural response modulations are heterogeneous, including both enhancement and suppression (Engel and Wang 2011; Lui and Pasternak 2011; Miller and Desimone 1994; Pagan et al. 2013; Roth and Rust 2018). This type of signal would be detectable using multivariate decoding methods but in univariate analyses may be obscured by averaging across all voxels in an ROI (Kamitani and Tong 2005; Norman et al. 2006; Serences and Saproo 2012). Therefore, our data are consistent with the interpretation that match representations are not an automatic by-product of stimulus repetition, but are linked to the task relevance of each stimulus.

Representations of identity and viewpoint match status were weaker in early visual and ventral ROIs compared with ROIs in the MD network. Moreover, several control analyses suggest that the representations in some of these occipital and ventral regions were driven primarily by low-level image statistics, as opposed to an object's status as a match in the task-relevant dimension. First, after removing all trials in which the current and previous objects were matches in both category and viewpoint, we found that viewpoint match decoding during the viewpoint task dropped to chance in all ROIs except for those

Fig. 6. Control analyses related to Fig. 5: viewpoint and identity match information in MD regions is not driven by low-level image statistics. *A* and *B*: to address the possibility that match status could have been inferred from low-level visual properties, we removed all trials in which an object had a high degree of shape similarity to the previous object and repeated the analyses of Fig. 5. *A*: identity match information remained above chance in all regions. *B*: viewpoint match information dropped to chance in the early visual cortex (EVC; comprising V1, V2, V3, V3AB, and V4) and lateral occipital complex (LOC; comprising LO and pFus) but remained above chance in multiple-demand (MD) network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)] regions of interest (ROIs). For individual subject data, see Fig. 7. C: after identifying that the pairwise similarity between images in object *set A* was informative about identity match status, we reanalyzed the identity task data using only the subjects shown object *set B*. Identity match classification in early visual and ventral visual cortex drops below significance when subjects shown object *set A* are removed but remains above chance in MD regions. *P* values were computed at the subject level over these 5 subjects and FDR-corrected across ROIs. Open circles indicate significance at $q = 0.05$; closed circles indicate significance at $q = 0.01$. Circles above individual bars indicate above-chance classification performance (test against 0); circles above pairs of bars (denoted by horizontal lines) indicate significant differences between bars (paired *t*-test). Error bars are ±SE. *D*: image correlation is predictive of identity match status for several subjects in object *set A*. To assess the possibility that identity match classification in EVC may have been driven by low-level similarity between pairs of images, we used the Pearson correlation coefficient to calculate the similarity between all pairs of object images (excluding image pairs that were matched in both category and viewpoint and thus highly similar). For each subject, we calculated the mean correlation coefficient between pairs of images that were a match in each feature vs. images that were not a match. In the plot, closed circles indicate that this mean value was higher for matching pairs than for nonmatching pairs ($\alpha = 0.01$). This finding suggests that the above-chance classification of identity match status observed in *set A* subjects may have been driven by low-level image features.

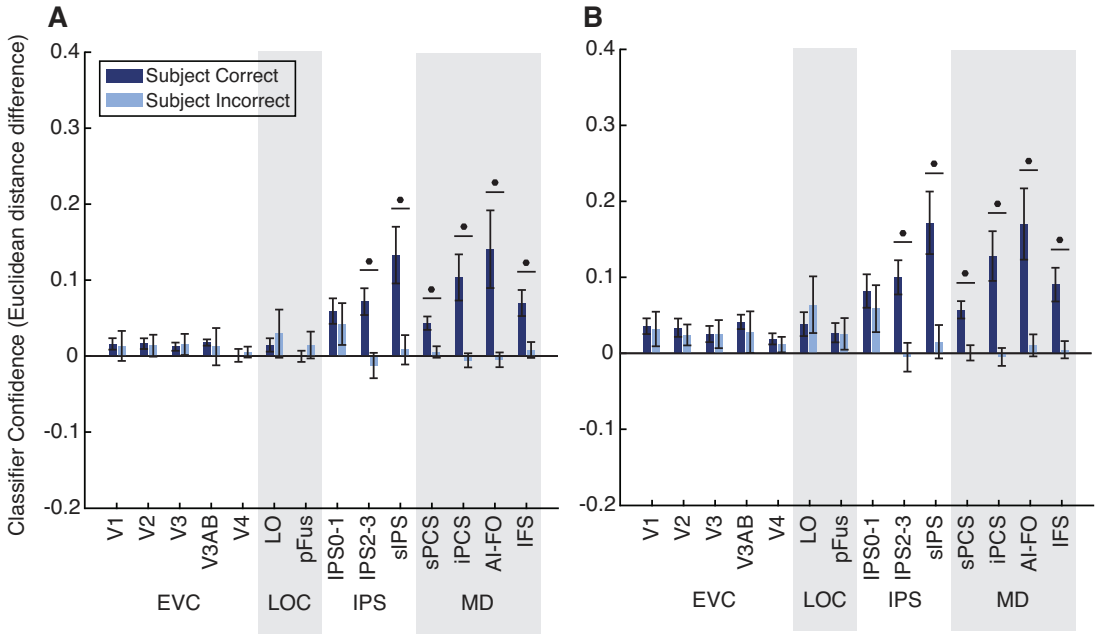


Fig. 8. Classifier evidence is associated with task performance. *A* and *B*: trials were classified according to their status as a match in the task-relevant feature, and the normalized Euclidean distance was used as a metric of evidence in favor of the correct classification (distance to incorrect label minus distance to correct label, plotted on the *y*-axis). Circles above pairs of bars indicate a significant difference between correct and incorrect trials (paired *t*-test). Closed circles indicate significance at $q = 0.01$ (FDR corrected). Error bars are ±SE. *A* shows results using all voxels; *B* shows results using 50 voxels per region of interest (ROI): early visual cortex (EVC; comprising V1, V2, V3, V3AB, and V4), lateral occipital complex (LOC; comprising LO and pFus), intraparietal sulcus [IPS; comprising IPS0–1, IPS2–3, and superior IPS (sIPS)], and the multiple-demand (MD) network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)].

in the MD network. Additionally, a post hoc image-based analysis revealed that one of our stimulus sets had a subtle bias in pixel similarity such that the degree of image similarity between each pair of objects was informative about the status of the pair as an identity match. When we repeated our decoding analysis using only subjects who had seen the nonbiased stimulus set, we found that identity match decoding dropped to chance in V1, V2, V3, V4, and pFus, although it remained above chance in V3AB and LO (and in regions of IPS and the MD network). Together, these findings suggest that low-level visual features may be partially responsible for the viewpoint and identity match representations we initially observed in early visual and ventral visual areas. We note, however, that this is not a powerful test; a more targeted experiment would be needed to rigorously assess the role of image similarity in driving match representations in early and ventral areas.

Our findings suggest that representations of object target status in MD regions show more invariance to identity preserving transformations than representations in ventral regions, which are more strongly influenced by low-level visual features. This view is consistent with several previous studies in nonhuman primates. For instance, in tasks that require identification of targets based on their membership in abstract categories whose boundaries are not predicted from visual similarity, neurons in both prefrontal cortex (PFC; Cromer et al. 2010; Freedman et al. 2001, 2003; Roy et al. 2010) and premotor cortex (Cromer et al. 2011) encode objects' target status. Studies that directly compare PFC and inferotemporal cortex (ITC) responses during these tasks have found that PFC neurons show a higher degree of abstract category selectivity (Freedman et al. 2003), as well as a stronger influence of task-related effects on object representation (McKee et al. 2014), compared with ITC neurons. Past work in humans is also consistent with a high degree of abstraction in MD representations. A recent fMRI study found that abstract face-identity information is represented more strongly in IPS than in LO and the fusiform face area (Jeong and Xu 2016). Another study found that during the delay period of a working memory task, object information could be decoded from PFC only when the task was nonvisual and from the posterior fusiform area only when the task was visual, supporting a dissociation between these regions in representing abstract and visual object information, respectively (Lee et al. 2013). Overall, our findings provide additional support for the conclusion that target representations in frontoparietal cortex reflect abstract signals that flexibly update to guide behavior, while ventral representations are more linked to the details of currently-viewed images.

That said, our results do not rule out the possibility of abstract object information or target-related information in the ventral stream. Many past studies have demonstrated viewpoint-invariant object information in human and primate ventral visual cortex (Anzellotti et al. 2014; Erez et al. 2016; Freiwald and Tsao 2010; Tanaka 1996). Furthermore, several studies have found that match representations in ITC are capable of generalizing across changes in size and position of objects (Lueschow et al. 1994; Roth and Rust 2018). Despite this, our multivariate decoding analysis failed to detect an association with behavior in any ventral ROI. One explanation

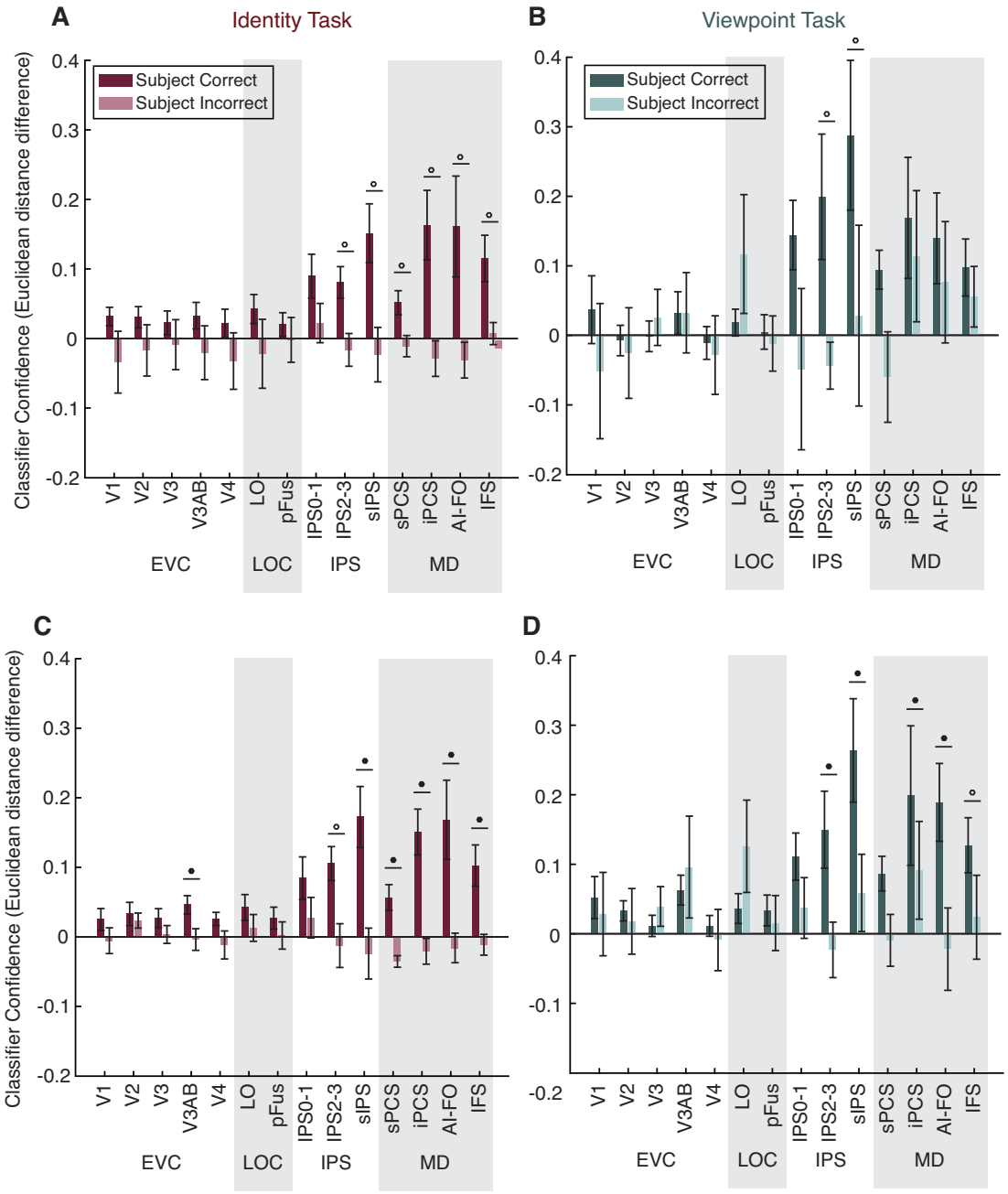


Fig. 9. Classifier evidence is associated with performance on both tasks individually. *A–D*: analysis is identical to that in Fig. 8 but was carried out using trials from the identity task only (*A* and *C*) or the viewpoint task only (*B* and *D*). *A* and *B* show results using all voxels; *C* and *D* show results using 50 voxels per region of interest (ROI): early visual cortex (EVC; comprising V1, V2, V3, V3AB, and V4), lateral occipital complex (LOC; comprising LO and pFus), intraparietal sulcus [IPS; comprising IPS0–1, IPS2–3, and superior IPS (sIPS)], and the multiple-demand (MD) network [superior precentral sulcus (sPCS), inferior precentral sulcus (iPCS), anterior insula/frontal operculum (AI/FO), and inferior frontal sulcus (IFS)].

for this finding is that target information was present at a level of granularity that is too fine to be detectable with our methods, or that the signal-to-noise regime in ventral object-selective cortex prevented us from detecting abstract object representations (Dubois et al. 2015). Furthermore, it is possible that linearly decodable information about abstract object properties may be more readily detectable at later stages of the ventral visual stream, such as perirhinal cortex, compared with LO and pFus (Erez et al. 2016; Pagan et al. 2013). Future studies will be needed to determine the extent to which abstract matches in identity and viewpoint may be represented within ventral visual cortex.

Additionally, although match representations in early visual cortex and the LOC were weaker than those measured in MD regions, we did observe a consistent modulation of these representations by task such that the relevant match was generally represented more strongly than the irrelevant match. This modulation was significant in LO during both tasks and in V4, V3AB, and pFus during the identity task (Fig. 5). One interpretation for the enhanced representation of relevant match status is the presence of top-down feedback that influences the content of visual representations in early visual and ventral object-selective cortex. This feedback could act to enhance representations of visual properties that are informative for computing match status, which would result in an enhancement of signals related to repetition of these properties. Alternatively, these feedback signals could contain nonspecific information about the presence of a relevant target. In either case, our data do not provide strong evidence that the match representations in early visual and ventral ROIs are involved in task performance.

Finally, one novel finding of the present study is that MD regions encode representations of matches in viewpoint across objects, in addition to matches in identity. This is especially interesting given that subjects had to be trained to recognize an arbitrary viewpoint of each object as its front. Therefore, the viewpoint dimension of an object may be regarded as a learned, semantic dimension rather than an intuitive physical dimension of the object. This may be another reason why we do not see invariant representations of viewpoint match signals in the ventral visual areas examined, because these regions may not encode such arbitrary dimensions that define potentially relevant objects. Further work will be needed to understand the extent to which neural representations of object viewpoint can be separated from visual features such as the axis of elongation.

In conclusion, our findings support a role of MD network regions in computing high-level and abstract target representations based on features such as viewpoint and identity, even across changes in retinal input patterns. The match representations in these regions are modulated by task relevance and are associated with behavioral performance on a trial-by-trial basis. In contrast, early visual and ventral object selective cortex contained weaker evidence for match representations, with some modulation by task, consistent with a role for top-down feedback in enhancing representations of relevant information at the earliest levels of processing.

ACKNOWLEDGMENTS

We thank Alexandra Woolgar and Kalanit Grill-Spector for code and advice regarding localizing regions of the multiple demand network and the lateral occipital complex, respectively; Kelvin Lam for collecting behavioral pilot data; and Timothy Brady, Edward Vul, Vy Vo, and Rosanne Rademaker for useful discussions.

GRANTS

This work was supported by funding provided by National Eye Institute Grant R01-EY025872 (to J. T. Serences) and a James S. McDonnell Foundation Scholar Award (to J. T. Serences).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

M.M.H. and J.T.S. conceived and designed research; M.M.H. performed experiments; M.M.H. analyzed data; M.M.H. and J.T.S. interpreted results of experiments; M.M.H. prepared figures; M.M.H. drafted manuscript; M.M.H. and J.T.S. edited and revised manuscript; M.M.H. and J.T.S. approved final version of manuscript.

REFERENCES

Andersson JL, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20: 870–888, 2003. doi:10.1016/S1053-8119(03)00336-7.

Andresen DR, Vinberg J, Grill-Spector K. The representation of object viewpoint in human visual cortex. *Neuroimage* 45: 522–536, 2009. doi:10.1016/j.neuroimage.2008.11.009.

Anzellotti S, Fairhall SL, Caramazza A. Decoding representations of face identity that are tolerant to rotation. *Cereb Cortex* 24: 1988–1995, 2014. doi:10.1093/cercor/bht046.

Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29: 1165–1188, 2001. doi:10.1214/aos/1013699998.

Biederman I. Recognizing depth-rotated objects: a review of recent research and theory. *Spat Vis* 13: 241–253, 2000. doi:10.1163/156856800741063.

Bracci S, Daniels N, Op de Beeck H. Task context overrules object- and category-related representational content in the human parietal cortex. *Cereb Cortex* 27: 310–321, 2017. doi:10.1093/cercor/bhw419.

Conway BR. The organization and operation of inferior temporal cortex. *Annu Rev Vis Sci* 4: 381–402, 2018. doi:10.1146/annurev-vision-091517-034202.

Cromer JA, Roy JE, Buschman TJ, Miller EK. Comparison of primate prefrontal and premotor cortex neuronal activity during visual categorization. *J Cogn Neurosci* 23: 3355–3365, 2011. doi:10.1162/jocn_a_00032.

Cromer JA, Roy JE, Miller EK. Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron* 66: 796–807, 2010. doi:10.1016/j.neuron.2010.05.005.

DiCarlo JJ, Cox DD. Untangling invariant object recognition. *Trends Cogn Sci* 11: 333–341, 2007. doi:10.1016/j.tics.2007.06.010.

Dubois J, de Berker AO, Tsao DY. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. *J Neurosci* 35: 2791–2802, 2015. doi:10.1523/JNEUROSCI.4037-14.2015.

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

Engel SA, Glover GH, Wandell BA. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb Cortex* 7: 181–192, 1997. doi:10.1093/cercor/7.2.181.

Engel TA, Wang X-J. Same or different? A neural circuit mechanism of similarity-based pattern match decision making. *J Neurosci* 31: 6982–6996, 2011. doi:10.1523/JNEUROSCI.6150-10.2011.

Erez J, Cusack R, Kendall W, Barense MD. Conjunctive coding of complex object features. *Cereb Cortex* 26: 2271–2282, 2016. doi:10.1093/cercor/bhv081.

Erez Y, Duncan J. Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. *J Neurosci* 35: 12383–12393, 2015. doi:10.1523/JNEUROSCI.1134-15.2015.

Fedorenko E, Duncan J, Kanwisher N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci USA* 110: 16616–16621, 2013. doi:10.1073/pnas.1315235110.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291: 312–316, 2001. doi:10.1126/science.291.5502.312.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci* 23: 5235–5246, 2003. doi:10.1523/JNEUROSCI.23-12-05235.2003.

Freiwald WA, Tsao DY. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330: 845–851, 2010. doi:10.1126/science.1194908.

Grill-Spector K. The neural basis of object perception. *Curr Opin Neurobiol* 13: 159–166, 2003. doi:10.1016/S0959-4388(03)00040-0.

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24: 187–203, 1999. doi:10.1016/S0896-6273(00)80832-6.

Guo F, Preston TJ, Das K, Giesbrecht B, Eckstein MP. Feature-independent neural coding of target detection during search of natural scenes. *J Neurosci* 32: 9499–9510, 2012. doi:10.1523/JNEUROSCI.5876-11.2012.

Henson RN. Neuroimaging studies of priming. *Prog Neurobiol* 70: 53–81, 2003. doi:10.1016/S0301-0082(03)00086-8.

Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat Neurosci* 19: 613–622, 2016. doi:10.1038/nn.4247.

Ito M, Tamura H, Fujita I, Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol* 73: 218–226, 1995. doi:10.1152/jn.1995.73.1.218.

Jackson J, Rich AN, Williams MA, Woolgar A. Feature-selective attention in frontoparietal cortex: multivoxel codes adjust to prioritize task-relevant information. *J Cogn Neurosci* 29: 310–321, 2017. doi:10.1162/jocn_a_01039.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage* 62: 782–790, 2012. doi:10.1016/j.neuroimage.2011.09.015.

Jeong SK, Xu Y. Behaviorally relevant abstract object identity representation in the human parietal cortex. *J Neurosci* 36: 1607–1619, 2016. doi:10.1523/JNEUROSCI.1016-15.2016.

Jerde TA, Curtis CE. Maps of space in human frontoparietal cortex. *J Physiol Paris* 107: 510–516, 2013. doi:10.1016/j.jphysparis.2013.04.002.

Kadohisa M, Petrov P, Stokes M, Sigala N, Buckley M, Gaffan D, Kusunoki M, Duncan J. Dynamic construction of a coherent attentional state in a prefrontal cell population. *Neuron* 80: 235–246, 2013. doi:10.1016/j.neuron.2013.07.041.

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

Lee SH, Kravitz DJ, Baker CI. Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nat Neurosci* 16: 997–999, 2013. doi:10.1038/nn.3452.

Lueschow A, Miller EK, Desimone R. Inferior temporal mechanisms for invariant object recognition. *Cereb Cortex* 4: 523–531, 1994. doi:10.1093/cercor/4.5.523.

Lui LL, Pasternak T. Representation of comparison signals in cortical area MT during a delayed direction discrimination task. *J Neurophysiol* 106: 1260–1273, 2011. doi:10.1152/jn.00016.2011.

Marr D, Nishihara HK. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc R Soc Lond B Biol Sci* 200: 269–294, 1978. doi:10.1098/rspb.1978.0020.

McKee JL, Riesenhuber M, Miller EK, Freedman DJ. Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *J Neurosci* 34: 16065–16075, 2014. doi:10.1523/JNEUROSCI.1660-14.2014.

Meyer T, Rust NC. Single-exposure visual memory judgments are reflected in inferotemporal cortex. *eLife* 7: e32259, 2018. doi:10.7554/eLife.32259.

Miller E, Erickson C, Desimone R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J Neurosci* 16: 5154–5167, 1996. doi:10.1523/JNEUROSCI.16-16-05154.1996.

Miller EK, Desimone R. Parallel neuronal mechanisms for short-term memory. *Science* 263: 520–522, 1994. doi:10.1126/science.8290960.

Miller EK, Li L, Desimone R. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* 254: 1377–1379, 1991. doi:10.1126/science.1962197.

Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10: 424–430, 2006. doi:10.1016/j.tics.2006.07.005.

Pagan M, Urban LS, Wohl MP, Rust NC. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat Neurosci* 16: 1132–1139, 2013. doi:10.1038/nn.3433.

Pinto N, Cox DD, DiCarlo JJ. Why is real-world visual object recognition hard? *PLoS Comput Biol* 4: e27, 2008. doi:10.1371/journal.pcbi.0040027.

Riesenhuber M, Poggio T. Models of object recognition. *Nat Neurosci* 3, *Suppl*: 1199–1204, 2000. doi:10.1038/81479.

Roth N, Rust NC. Inferotemporal cortex reflects behaviorally-relevant target match information during invariant object search (Preprint). *bioRxiv* 152181, 2018. doi:10.1101/152181.

Roth P. Missing data: a conceptual review for applied psychologists. *Pers Psychol* 47: 537–560, 1994. doi:10.1111/j.1744-6570.1994.tb01736.x.

Roy JE, Riesenhuber M, Poggio T, Miller EK. Prefrontal cortex activity during flexible categorization. *J Neurosci* 30: 8519–8528, 2010. doi:10.1523/JNEUROSCI.4837-09.2010.

Serences JT, Saproo S. Computational advances towards linking BOLD and behavior. *Neuropsychologia* 50: 435–446, 2012. doi:10.1016/j.neuropsychologia.2011.07.013.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268: 889–893, 1995. doi:10.1126/science.7754376.

Sprague TC, Serences JT. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat Neurosci* 16: 1879–1887, 2013. doi:10.1038/nn.3574.

Stigliani A, Weiner KS, Grill-Spector K. Temporal processing capacity in high-level visual cortex is domain specific. *J Neurosci* 35: 12412–12424, 2015. doi:10.1523/JNEUROSCI.4822-14.2015.

Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC. Visual topography of human intraparietal sulcus. *J Neurosci* 27: 5326–5337, 2007. doi:10.1523/JNEUROSCI.0991-07.2007.

Tanaka K. Neuronal mechanisms of object recognition. *Science* 262: 685–688, 1993. doi:10.1126/science.8235589.

Tanaka K. Inferotemporal cortex and object vision. *Annu Rev Neurosci* 19: 109–139, 1996. doi:10.1146/annurev.ne.19.030196.000545.

Tarr MJ, Williams P, Hayward WG, Gauthier I. Three-dimensional object recognition is viewpoint dependent. *Nat Neurosci* 1: 275–277, 1998. doi:10.1038/1089.

Turk-Browne NB, Yi DJ, Leber AB, Chun MM. Visual quality determines the direction of neural repetition effects. *Cereb Cortex* 17: 425–433, 2007. doi:10.1093/cercor/bhj159.

Valyear KF, Culham JC, Sharif N, Westwood D, Goodale MA. A double dissociation between sensitivity to changes in object identity and object orientation in the ventral and dorsal visual streams: a human fMRI study. *Neuropsychologia* 44: 218–228, 2006. doi:10.1016/j.neuropsychologia.2005.05.004.

Vaziri-Pashkam M, Xu Y. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *J Neurosci* 37: 8767–8782, 2017. doi:10.1523/JNEUROSCI.3392-16.2017.

Vinberg J, Grill-Spector K. Representation of shapes, edges, and surfaces across multiple cues in the human visual cortex. *J Neurophysiol* 99: 1380–1393, 2008. doi:10.1152/jn.01223.2007.

Wandell BA, Dumoulin SO, Brewer AA. Visual field maps in human cortex. *Neuron* 56: 366–383, 2007. doi:10.1016/j.neuron.2007.10.012.

Ward EJ, Isik L, Chun MM. General transformations of object representations in human visual cortex. *J Neurosci* 38: 8526–8537, 2018. doi:10.1523/JNEUROSCI.2800-17.2018.

Winawer J, Witthoft N. Human V4 and ventral occipital retinotopic maps. *Vis Neurosci* 32: E020, 2015. doi:10.1017/S0952523815000176.

Woloszyn L, Sheinberg DL. Neural dynamics in inferior temporal cortex during a visual working memory task. *J Neurosci* 29: 5494–5507, 2009. doi:10.1523/JNEUROSCI.5785-08.2009.

Chapter 2 is a reprint of the material as it appears in: Henderson, M., & Serences, J. T. (2019). Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features. *Journal of Neurophysiology*, 121(4): 1410-1427. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3: Prospective response planning degrades spatial memory representations in human visual cortex

Abstract

Human neuroimaging studies suggest that sensory cortex is recruited to support the storage of information in working memory (WM). In contrast, many non-human primate (NHP) studies suggest a limited role for sensory areas and instead implicate both sustained and transient codes in prefrontal cortex. This divergence is often attributed to differences in recording method – single-unit physiology measures spiking directly, while fMRI may pick up on non-spike related modulations. However, an alternative explanation is that tasks used with human subjects generally require retrospective sensory-like representations, while NHP tasks often encourage prospective motor-based strategies. Here, we tested this hypothesis by training human subjects to perform a spatial working memory task where the required manual motor response was either predictable or unpredictable. Using fMRI, we found that the amount of information about spatial position in early visual and parietal regions of interest was lower for predictable than unpredictable trials. Further, we show that a representation of the planned motor response emerged in primary motor, premotor, and primary somatosensory cortex on the same trials where sensory decodability declined. These results suggest that there is not one locus of WM representations. Instead, the neural networks supporting WM can be strategically reconfigured depending on task demands.

Introduction

Working memory (WM) allows organisms to hold information in mind about past experiences and use it to guide future behavior. Depending on the task, this can be accomplished by maintaining a representation of recent sensory inputs (a "retrospective" code), or by forming a plan for an upcoming action (a "prospective" code). Maintaining sensory-like retrospective information is likely to be useful in tasks requiring memory for fine sensory details, while re-

mapping information into an action-oriented format may reduce representational complexity when the stimulus-response mapping is known in advance.

Many nonhuman primate (NHP) electrophysiology studies have employed tasks that encourage the use of a prospective motor-based strategy, such as a delayed saccade task where the position of a cue briefly presented at the beginning of a trial is predictive of the saccade that must be made at the end of the trial. Single unit recordings made during these tasks suggest an important role for the prefrontal cortex (PFC) in maintaining remembered information across brief delay periods (Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971; Goldman-Rakic, 1995). Other NHP studies have dissociated memorized visual features from upcoming actions, but generally utilize a small set of discrete memory stimuli, which may encourage animals to form a discrete, categorical representation of a memory item rather than maintaining the precise details of its original appearance (Funahashi, Chafee, & Goldman-Rakic, 1993; Mendoza-Halliday, Torres, & Martinez-Trujillo, 2014; Miller, Erickson, & Desimone, 1996). These studies also generally suggest a key role for frontal cortex and a more limited role for sensory cortex in WM storage (Mendoza-Halliday et al., 2014).

In contrast, human neuroimaging studies typically use tasks that require subjects to remember precise values of continuously varying visual features, and to report remembered features using responses that cannot be planned during the delay period (Albers, Kok, Toni, Dijkerman, & De Lange, 2013; Christophel, Hebart, & Haynes, 2012; Ester, Serences, & Awh, 2009; Harrison & Tong, 2009; Lorenc, Sreenivasan, Nee, Vandenbroucke, & D'Esposito, 2018; Rademaker, Chunharas, & Serences, 2019; Serences, Ester, Vogel, & Awh, 2009; Xing, Ledgeway, McGraw, & Schluppeck, 2013). These studies typically find that patterns of voxel activation measured in visual cortex encode information about specific feature values held in

memory, supporting the role of visual cortex in maintaining detailed visual representations during WM. The difference in results between human and primate studies may be accounted for in part by differences in measurement modality. For example, the BOLD response is driven in part by modulations of local field potentials (Boynton, 2011; Goense & Logothetis, 2008; Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001; Logothetis & Wandell, 2004), and local field potentials recorded in monkey visual cortex represent motion direction during WM, even though single unit spike rates may not (Mendoza-Halliday et al., 2014).

While different methods of assessing neural responses may play a role in the discrepancies between the human and NHP literatures, an important alterative is that task demands may underlie the observed differences. Specifically, the tasks that have typically been used in human neuroimaging research, which most often require retrospective or sensory-like codes, may encourage top-down recruitment of the same early sensory areas that support high-precision perceptual representations (Awh & Jonides, 2001; Gazzaley & Nobre, 2012; Pasternak & Greenlee, 2005; Serences, 2016; Sreenivasan, Curtis, & D'Esposito, 2014). In contrast, the tasks commonly used in NHP research may invoke prospective or categorical codes that rely more heavily on areas involved in planning and motor production, relying less on sensory cortex (Cisek & Kalaska, 2010; Myers, Stokes, & Nobre, 2017). This hypothesis predicts that the involvement of sensory cortex in representing features in WM depends in part on whether a task allows for response planning. Here, we directly tested this hypothesis using a spatial WM task in which the manual response required on each trial was either predictable or unpredictable (Figure 3.1A). We hypothesized that the discriminability of spatial position information from retinotopic visual cortex would decrease when responses were predictable, in accordance with a decreased reliance on visual cortex for information storage. We further predicted this would be

accompanied by a response-related representation emerging in motor, premotor and somatosensory cortex.

Results

While in the MRI scanner, subjects performed a spatial working memory task in which they were required to remember the spatial position of a briefly presented target across a delay period (a small white dot appearing anywhere along an imaginary circle with radius 7°). Participants used their left or right index finger to report which side of a response probe disk the target had been presented on (Figure 3.1A). We manipulated subjects' ability to pre-plan their motor response by presenting a "preview" of the response disk stimulus at the beginning of the delay period. The preview disk was preceded by a cue indicating whether the trial was part of the "predictable" or "unpredictable" condition. On predictable trials, the preview disk was rotated randomly with respect to orientation of the response disk, but on unpredictable trials, the preview disk was at a random orientation. Predictable and unpredictable trials were randomly intermixed throughout each run.

Task performance was overall better in the predictable compared to the unpredictable condition (Figure 3.1C-D). Participants were significantly faster (predictable mean +/- SEM: 0.57 +/- 0.03 sec; unpredictable: 1.08 +/- 0.06 sec; $t(5)=-9.618$; $p<0.001$; paired t-test) and more accurate in the predictable condition (predictable mean +/- SEM: 93.92 +/- 2.12%; unpredictable: 89.83 +/- 1.12%; $t(5)=3.463$; $p=0.018$; paired t-test). These behavioral benefits, particularly the response time difference, support the idea that subjects used the cue to plan their response during the predictable condition.

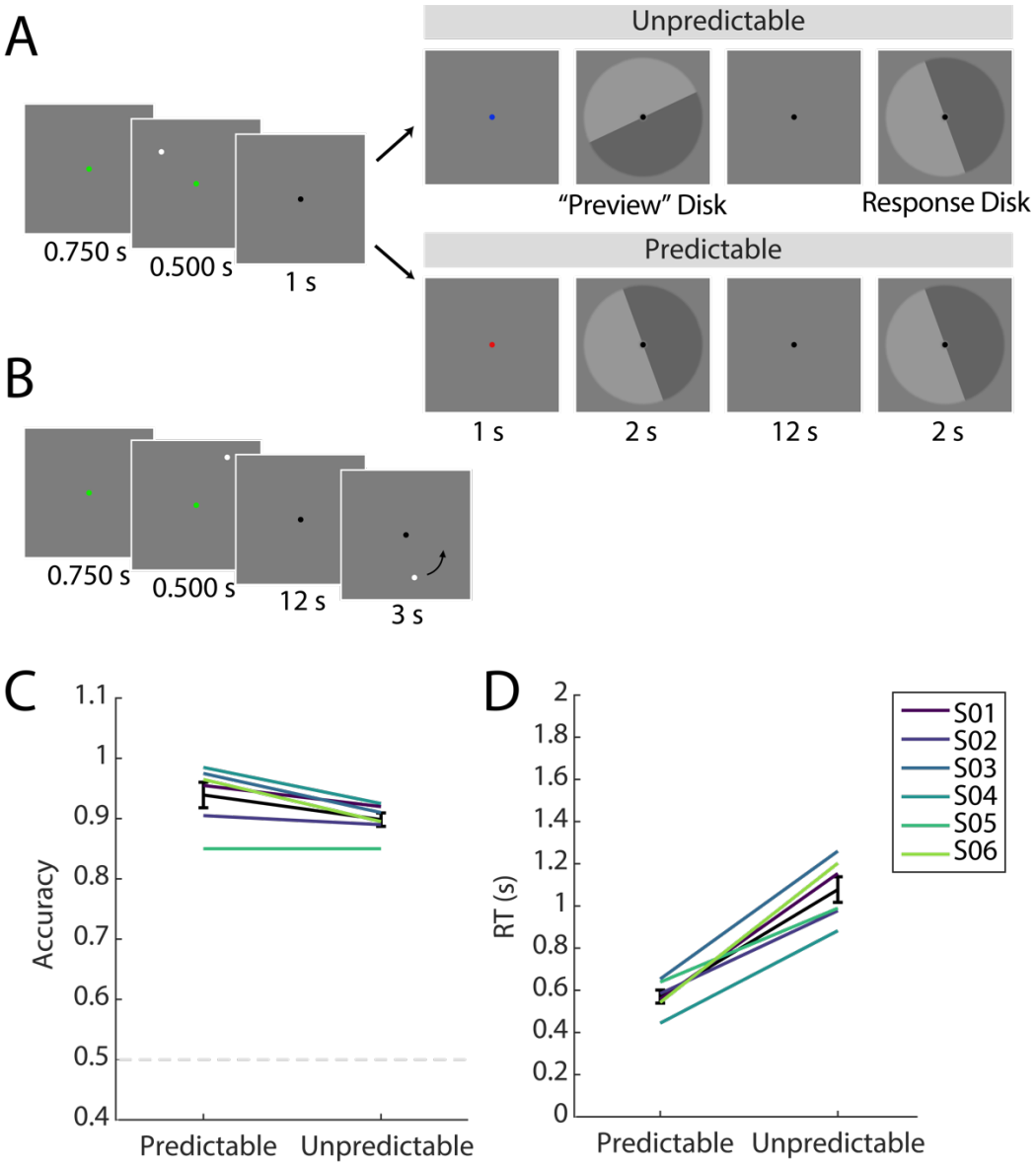


Figure 3.1. Ability to plan a response improves accuracy and response speed during a spatial working memory task. **(A)** During each trial of the main scanner task, subjects remembered the position of a spatial target dot presented at a random angular position (7° eccentricity). After a 16 sec delay, they made a binary response to indicate which side of a probe disk stimulus the target position was on. Partway through the delay, a "preview" of the response probe disk was shown that either predicted (bottom row) or was random with respect to (top row) the orientation of the final probe disk (see *Methods: Main Behavioral Task in the Scanner* for more details). **(B)** In a separate spatial working memory mapping task, subjects remembered a target location and responded by moving a probe dot continuously to the remembered position (see *Methods: Spatial Working Memory Mapping Task)*. **(C)** Accuracy for each subject, averaged over all trials of each main task condition. Error bars represent mean +/- standard error of the mean (SEM). **(D)** Reaction time for each subject, averaged over all trials of each main task condition; error bars as in B.

Next, we examined the average fMRI responses in visual and sensorimotor cortical areas. Our visual regions of interest (ROIs) were retinotopically-defined areas in occipital and parietal cortex, and our sensorimotor ROIs were independently localized response-related regions of primary motor cortex (M1), primary somatosensory cortex (S1), and premotor cortex (PMc, see *Methods: Sensorimotor Cortex Localizer Task)*. First, we used deconvolution to calculate the average hemodynamic response function (HRF) for voxels in each ROI during each task condition (see *Methods: Univariate Analyses*). This allowed us to test whether any visual areas exhibited sustained delay period activation, and whether delay period activation differed between task conditions. Based on past work, we expected to find sustained delay period activation in parietal but not occipital ROIs (Curtis & D'Esposito, 2003; D'Esposito, 2007; Ester, Sprague, & Serences, 2015; Riggall & Postle, 2012). Additionally, we predicted that univariate BOLD delay activation in parietal cortex would be higher for the predictable relative to the unpredictable condition, reflecting a greater reliance on spatial information when motor coding was not available (Curtis, Rao, & D'Esposito, 2004). In sensorimotor cortex, we predicted the opposite pattern, indicating the emergence of motor plan representations during predictable trials only (Calderon, Van Opstal, Peigneux, Verguts, & Gevers, 2018; Donner, Siegel, Fries, & Engel, 2009).

As expected, we found that the BOLD signal in all retinotopic ROIs increased following visual stimulation, both after the presentation of the spatial target item and again following the preview disk presentation (Figure 3.2). Likewise, we replicated the typical finding that early retinotopic areas did not show sustained delay period activation. In addition, there was no effect of task condition (predictable vs,. unpredictable) on delay period activity in V1-V4, V3AB, or LO2. In contrast, mean BOLD responses in parietal areas IPS0-3, as well as LO1, showed

elevated late delay period activity in the unpredictable condition relative to the predictable condition (significance tested using a Wilcoxon signed-rank test with permutation, all p-values <0.05). Sensorimotor areas S1, M1, and PMc showed a condition difference in the opposite direction, with higher mean BOLD responses in the predictable compared to the unpredictable condition at several timepoints early in the delay period.

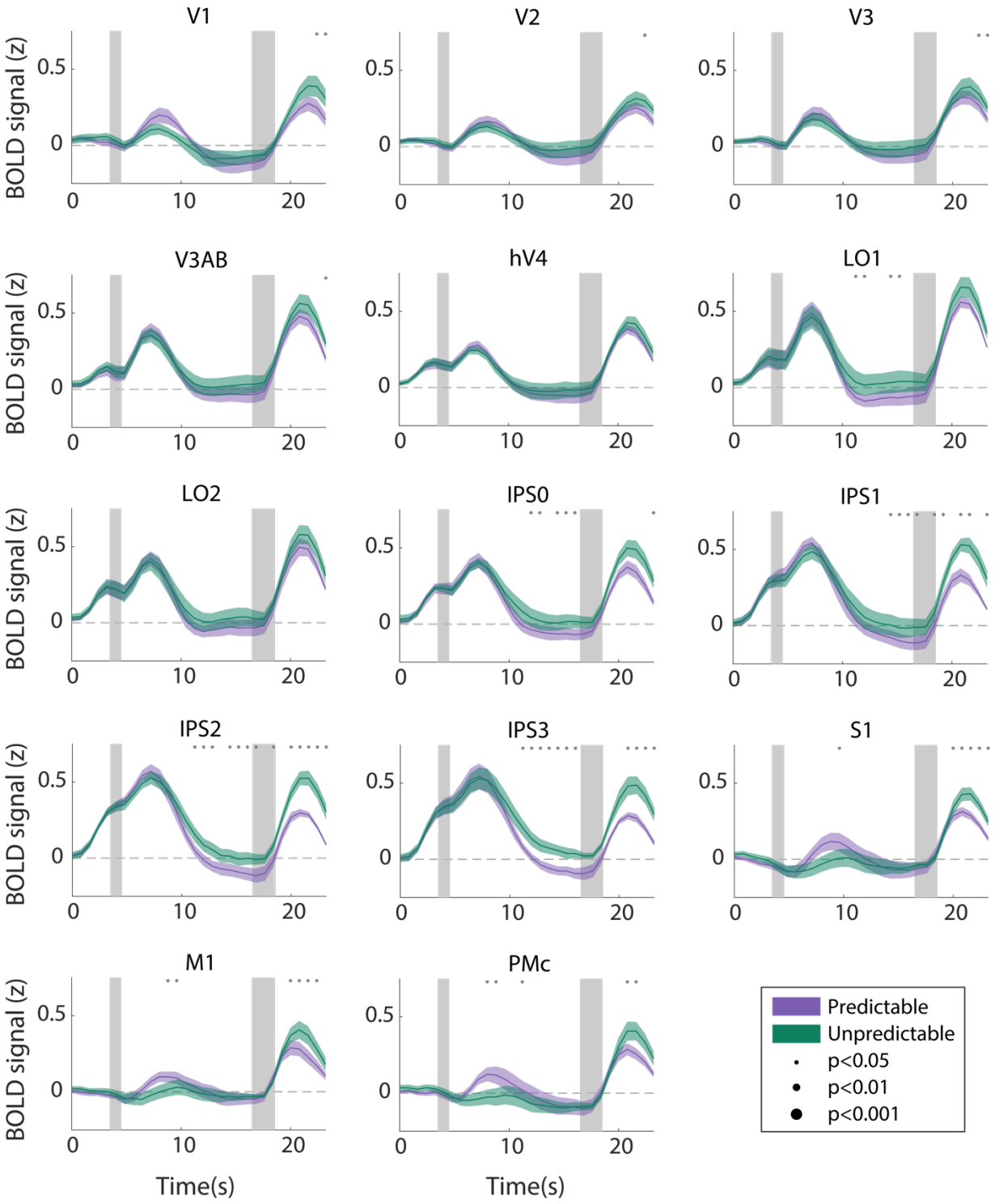


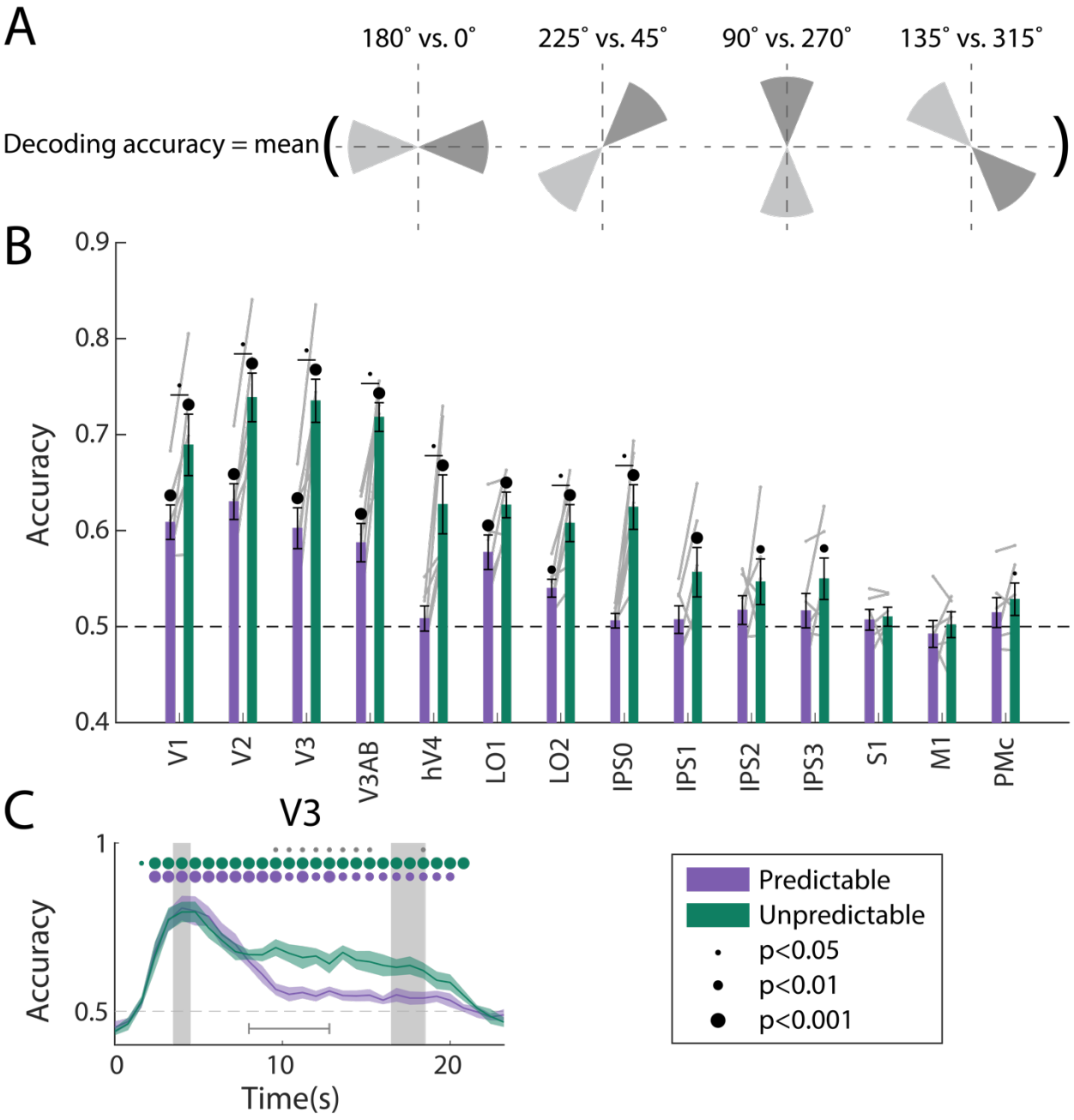
Figure 3.2. Hemodynamic response function in each ROI during the predictable (purple) and unpredictable (green) conditions. Shaded gray rectangles indicate the time periods when the "preview" disk was onscreen (3.5-4.5 sec) and when the response disk was onscreen (16.5-18.5 sec). Shaded error bars represent ±1 SEM across subjects. Gray dots indicate timepoints showing a significant condition difference, evaluated using non-parametric statistics (see *Methods: Univariate Analyses* for details).

The univariate results described above suggest the possibility of a handoff between parietal and sensorimotor ROIs during response planning, but leave open the question of how representations in early visual cortex differ between conditions. Previous work suggests that region-level patterns of voxel activation in early visual cortex contain information about remembered items even in the absence of sustained univariate activation (Kamitani & Tong, 2005; Norman, Polyn, Detre, & Haxby, 2006; Sprague, Saproo, & Serences, 2015). Therefore, we next examined the information content of population-level patterns, by using a multivariate linear classifier to decode the angular spatial position of the remembered location based on multi-voxel activation patterns measured during the delay period (Figure 3.3). Before decoding, we subtracted the mean signal across voxels from each single-trial activation pattern to ensure that any small differences in the mean response did not contribute to classification accuracy. We then sorted the continuous angular position labels into 8 bins, and used a decoding scheme with four binary classifiers, each trained to discriminate between two bins separated by 180° (see Figure 3.3A). The final decoding values reflect the average of these binary classifiers' performance, where chance performance is 50% (see *Methods: Spatial Position Decoding*). To facilitate an independent comparison between the two task conditions , we used a separate spatial working memory mapping task as a training set for this decoding analysis (Sprague, Boynton, & Serences, 2019; Figure 3.1B; see *Methods: Spatial Working Memory Mapping Task*).

The results of this multivariate analysis further support the idea of a decreased reliance on sensory codes when motor planning was available. Across both conditions, spatial decoding accuracy was strongest in early visual areas V1, V2, V3, and V3AB, and became progressively weaker at more anterior regions of the visual hierarchy (Figure 3.3B; decoding using data averaged between 8-12.8 sec after target onset). Spatial decoding was at chance in the three

sensorimotor areas S1, M1, and PMc. We also observed a pronounced effect of task condition, where decoding was significantly higher for the unpredictable condition than the predictable condition. A two-way repeated measures ANOVA revealed a main effect of ROI, a main effect of condition, and an ROI x condition interaction (ROI: $F(13,65)=24.548$, $p<0.001$; Condition: $F(1,5)=35.537$, $p=0.001$; ROI x Condition $F(13,65)$: 5.757, $p<0.001$; p-values obtained using permutation test; see *Methods: Spatial Position Decoding*). Follow-up pairwise comparisons showed that the difference between unpredictable and predictable decoding was significant in V1, V2, V3, V3AB, hV4, LO2, and IPS0 (p-values for V1=0.034, V2=0.048, V3=0.034, V3AB=0.034, hV4=0.038, LO2=0.038, IPS0=0.030; two-tailed p-values obtained fusing a Wilcoxon signed-rank test with permutation, uncorrected). In all the IPS subregions (IPS0-3), as well as hV4, spatial decoding was above chance for the unpredictable condition, but at chance for the predictable condition. In all other visual areas, decoding was above chance in both conditions. Time-resolved analyses revealed that this difference between the conditions was not present early in the trial, when the trial condition was not yet known, but emerged around 6 seconds after the presentation of the preview disk stimulus, and persisted until the response probe appeared (Figure 3.3C, Supplementary Figure 3.1).

Figure 3.3. Response planning is associated with weaker spatial memory representations. **(A)** Schematic of decoding procedure. Continuous values of angular spatial position were divided into 8 discrete bins, and 4 binary decoders were trained to discriminate between patterns corresponding to bins 180° apart in angular position. The final decoding accuracy was obtained by averaging accuracy over these 4 decoders. **(B)** Average spatial decoding accuracy for each ROI for trials of each task condition. The spatial decoder was always trained on data from an independent spatial working memory mapping task, and tested on data measured during the delay period of the main task (averaged within a window 8-12.8 sec from start of trial; see *Methods: Spatial Position Decoding* for more details). Error bars reflect ±1 SEM across subjects, and light gray lines indicate individual subject data. Dots above bars and pairs of bars indicate the statistical significance of decoding within each condition, and of condition differences, respectively, both evaluated using non-parametric statistics. Dot sizes reflect significance level. **(C)** Spatial decoding accuracy over time for an example ROI. Shaded gray rectangles indicate the periods of time when the "preview" disk was onscreen (3.5-4.5 sec) and when the response disk was onscreen (16.5-18.5 sec). Shaded error bars represent ±1 SEM across subjects, colored dots indicate significance of decoding within each condition, and gray dots indicate significant condition differences, with dot sizes reflecting significance levels as in B. Gray bracket in **C** indicates the time range in which data were averaged to produce **B** (8-12.8 sec)**.** See Supplementary Figure 3.1 for time-resolved decoding in all ROIs.


A
180˚ vs. 0˚
225˚ vs. 45˚
90˚ vs. 270˚
135˚ vs. 315˚
Decoding accuracy = mean
B
Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
V1
V2
V3
V3AB
hV4
LO1
LO2
IPS0
IPS1
IPS2
IPS3
S1
M1
PMc
C
V3
Accuracy
1
0.5
0
10
20
Time(s)
Predictable
Unpredictable
p<0.05
p<0.01
p<0.001

The observed difference in spatial decoding accuracy between conditions may reflect a weakening of spatial memory representations when response planning was available. Alternatively, because we used an independent mapping task as a training set for the decoder, the decrease in spatial decoding performance for the predictable condition may indicate a difference in representational format between the two task conditions (e.g. Lorenc, Vandenbroucke, Nee, de Lange, & D'Esposito, 2020; Vaziri-Pashkam & Xu, 2017). For example, there may be as much information encoded in early visual in the predictable and unpredictable conditions, but responses patterns in the predictable condition might deviate more from the patterns evoked by the spatial working memory mapping task. To evaluate this possibility, we ran the same spatial decoding analysis using data from one condition at a time to both train and test the decoder (see *Methods: Spatial Position Decoding* for cross-validation procedure). Within-condition spatial decoding showed a similar pattern of results to the analysis using the independent training set, though the condition difference was slightly smaller (Supplementary Figure 3.2). A few areas (hV4, IPS0, IPS2, IPS3) that had chance level spatial decoding performance for the predictable condition in the previous analysis, showed above-chance accuracy for both conditions in this analysis. Thus, overall decoding accuracy in the predictable condition benefitted slightly from within-condition training. However, as in the previous analysis, a two-way repeated measures ANOVA still revealed a main effect of condition, as well as a main effect of ROI and ROI x condition interaction (ROI: $F(13,65)=12.873$, $p<0.001$; Condition: $F(1,5)=11.461$, $p=0.017$; ROI x Condition: $F(13,65)=2.581$, $p=0.011$; p-values obtained using permutation test). This supports the interpretation that spatial memory representations in visual cortex were degraded in quality in the predictable condition relative to the unpredictable condition, rather than changing in format without any degradation.

Finally, we investigated how information related to the preparation of a motor response was reflected in BOLD activation patterns. Since motor responses were always made with the left or the right index finger, we trained a binary linear classifier to predict which finger was associated with the required response on each trial, based on delay period signal in each ROI (Figure 3.4A, see *Methods: Response Decoding* for detailed classification procedure). During the predictable condition, responses could be decoded with above chance accuracy in each of the three sensorimotor ROIs, S1, M1, and PMC (accuracy +/- SEM for S1: 60.4 +/- 3.8%, $p<0.001$; M1: 61.8 +/- 4.1%, $p<0.001$; PMc: 61.4 +/- 3.6%, decoding using data averaged between 8-12.8 sec after target onset; $p<0.001$; one-tailed p-values against 50% obtained using permutation test, uncorrected). In contrast, in the unpredictable condition, where the required response was not yet known during the delay period, response decoding was at chance in all ROIs. All retinotopic visual ROIs showed chance level response decoding for both conditions. Supporting this dissociation between visual and motor ROIs, a two-way repeated measures ANOVA revealed a main effect of ROI and an ROI x condition interaction (ROI: $F(13,65) = 4.00$, $p<0.001$; Condition $F(1,5) = 3.80$, $p=0.106$; ROI x Condition: $F(13,65) = 2.94$, $p=0.001$; p-values obtained using permutation test). A time-resolved decoding analysis revealed that response information began to emerge approximately 4 seconds after the onset of the preview disk stimulus, decreased slightly toward the end of the delay period, then rose steeply after the response probe onset when the participants actually made a response (Figure 3.4B, Supplementary Figure 3.3). The probe-related increase in response decoding appeared sooner for the predictable condition than the unpredictable condition, in agreement with the speeding of response times in the predictable condition observed behaviorally (Figure 3.1D).

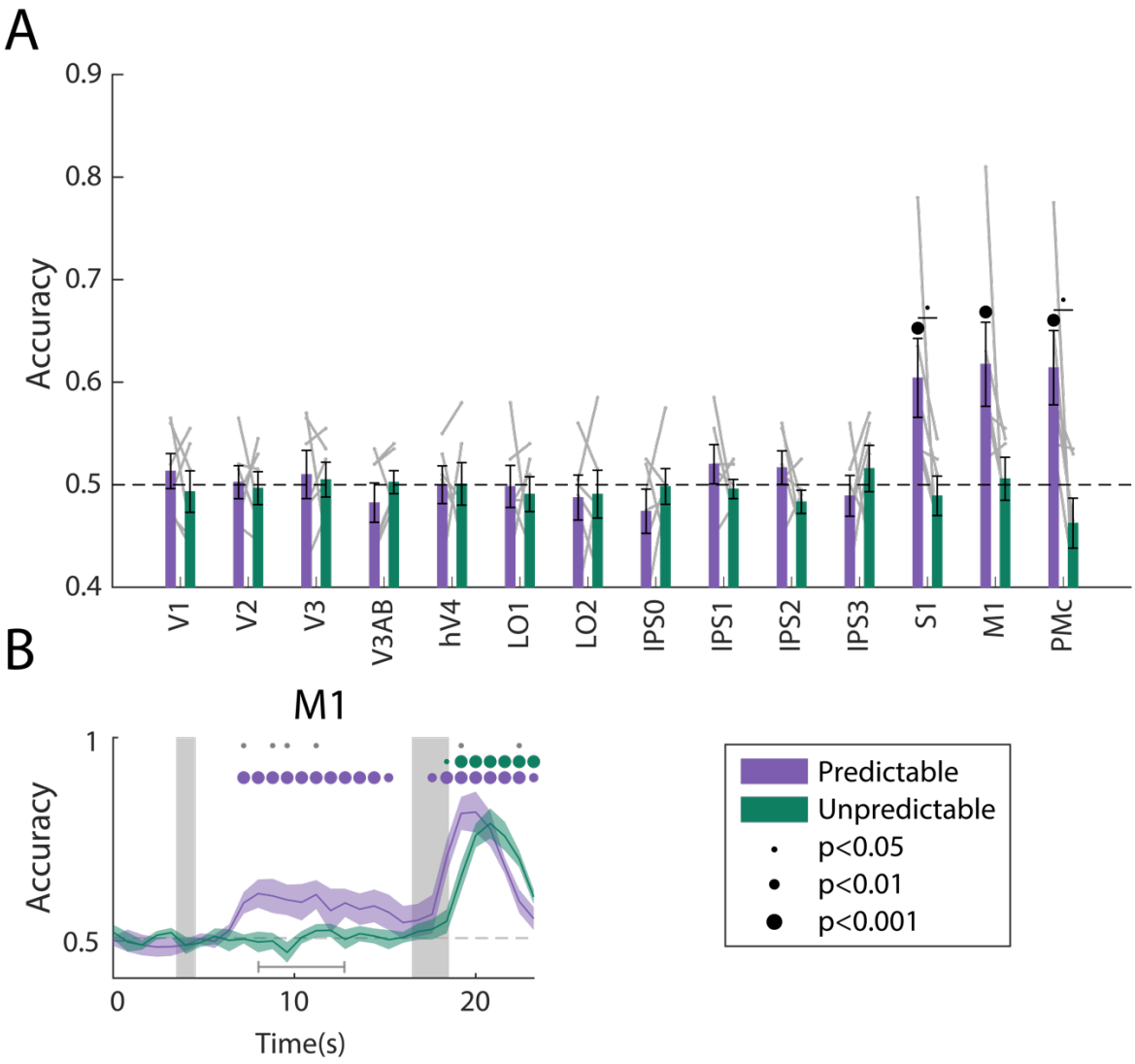


Figure 3.4. Response plans can be decoded from sensorimotor ROIs during the delay period. **(A)** A linear decoder was trained to classify the finger (left or right index) associated with the correct response on each trial, using data measured during the delay period of trials in each task condition separately (averaged within a window 8-12.8 sec from start of trial; see *Methods: Response Decoding* for more details). Error bars reflect ±1 SEM across subjects, and light gray lines indicate individual subjects. Dots above bars and pairs of bars indicate the statistical significance of decoding within each condition, and of condition differences, respectively, both evaluated using non-parametric statistics. Dot sizes reflect significance level. **(B)** Response decoding accuracy over time one example ROI. Shaded gray rectangles indicate the periods of time when the "preview" disk was onscreen (3.5-4.5 sec) and when the response probe disk was onscreen (16.5-18.5 sec). Shaded error bars represent ±1 SEM across subjects. Colored dots indicate significance of decoding within each condition, and gray dots indicate significant condition differences, with dot sizes reflecting significance levels as in A. Gray bracket in **B** indicates the time range in which data were averaged to produce **A** (8-12.8 sec)**.** For time-resolved decoding in all ROIs, see Supplementary Figure 3.3.

Discussion

Past studies of visual WM have produced apparently conflicting results, leading to competing theories about the brain regions that support WM. In particular, human neuroimaging studies often find robust representations of remembered visual features in early visual cortex, whereas NHP electrophysiology studies suggest a more limited role for sensory areas and instead emphasize the importance of PFC. In this experiment, we evaluated the hypothesis that these differences might be related to differences in task demands. Our results show that decoding of the remembered location from activation patterns in visual cortex was lower when subjects had the opportunity to prospectively plan their motor response (predictable condition) than when they were unable to plan (unpredictable condition). This effect was consistent across a range of early visual and parietal ROIs (Figure 3.3B), supporting the idea that the recruitment of visual cortex for WM storage is task-dependent. Conversely, during the same trials where spatial representations became weaker in visual cortex, several sensorimotor ROIs showed above-chance decoding of the subject's planned response during the delay period prior to response execution (Figure 3.4A). Together, these results suggest that rather than having one fixed mechanism, visual WM can be supported by different coding schemes that are adapted to current task demands.

Our demonstration of a hand-off between visual and sensorimotor codes builds on some past work suggesting that the neural systems underlying WM may be flexibly reconfigured based on task demands. For instance, IPS shows higher delay activation during a task where oculomotor responses are decoupled from spatial memory items (i.e. retrospective coding), whereas oculomotor areas show higher activation when responses and memoranda are yoked (i.e. prospective coding; Curtis, Rao, & D'Esposito, 2004). This finding parallels our observation of

higher mean delay period activation in IPS for the predictable condition relative to the unpredictable condition (Figure 3.2). Together, these results suggest specialized networks for retrospective versus prospective coding formats. Other prior work has examined the effect of attentional priority on WM representational format, where high priority items are represented in active codes associated with spiking activity, and lower priority items may be maintained by spike-silent mechanisms such as synaptic weight remodeling (Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012; Lorenc et al., 2020; Mongillo, Barak, & Tsodyks, 2008; Rose et al., 2016; Sprague, Ester, & Serences, 2016; Stokes, 2015; Wolff, Jochim, Akyürek, & Stokes, 2017). Prioritized WM representations can be further reconfigured so that they are optimized for future behavior, which may include the activation of circuits related to motor output (Myers et al., 2017; Nobre & Stokes, 2019; Schneider, Barth, & Wascher, 2017; Souza & Oberauer, 2016; van Ede, Chekroud, Stokes, & Nobre, 2019). Adding to these findings, our study is the first to simultaneously show a drop in spatial information in sensory cortex and an increase in response information in sensorimotor cortex during response preparation. Together, much evidence supports the idea that information may shift between visual and sensorimotor codes – and brain areas – during maintenance in WM.

Together, these results add to an existing body of work supporting the dissociation between sustained delay activation and multivariate decoding performance as indices of WM storage (Emrich, Riggall, La Rocque, & Postle, 2013; Ester et al., 2015; Harrison & Tong, 2009; Riggall & Postle, 2012; Serences et al., 2009). Specifically, while we found widespread task differences in multivariate spatial decoding accuracy, early visual areas V1, V2, V3, V4, V3AB, and LO2 showed no sustained BOLD responses or differences in the mean BOLD signal between conditions during the delay period. The lack of sustained responses and mean signal

differences suggests that the observed differences in decoding performance were not due to global changes in signal due to arousal or task difficulty, but instead due to differences in the amount of information represented within population-level patterns of visual cortex activation.

In addition to the large body of neuroimaging work supporting the involvement of visual cortex in WM storage, some NHP electrophysiology studies have shown spiking activity in V1 that reflects the contents of WM (Supèr, Spekreijse, & Lamme, 2001; Van Kerkoerle, Self, & Roelfsema, 2017). These findings are apparently inconsistent with other NHP studies that fail to detect information about remembered features in spiking activity in other early visual regions like the middle temporal (MT) area (Mendoza-Halliday et al., 2014). One explanation for these differences may be that WM representations in sensory cortex are thought to be mediated by top-down projections terminating in superficial and deep layers of cortex, resulting in spiking activity that is largely localized to these layers (Lawrence et al., 2018; Van Kerkoerle et al., 2017). Therefore, if recordings are made from layers whose inputs are dominated by feedforward information from the lateral geniculate nucleus, such as layer 4, spiking activity related to WM may be missed. In addition, task differences may contribute to the different results. For instance, the curve-tracing task used by Van Kerkoerle and colleagues is likely to require precise spatial information that cannot easily be re-mapped into a non-sensory format. In contrast, as mentioned previously, many tasks used in NHP studies require memory for one of just a few discrete features or items, which may reduce the need for recruitment of precisely tuned neural populations. Along with the results of our study, this highlights the important role of task demands in determining the mechanisms for information storage in WM. Comparing decoding performance in human visual cortex for tasks requiring fine versus coarse representations is one possible direction for future studies.

We observed above-chance delay period decoding of subjects' planned responses in M1, S1, and PMc. Each of these areas has previously been shown to exhibit motor preparatory activity in the context of tasks involving delayed reaching or finger pressing (Ariani, Pruszynski, & Diedrichsen, 2020; Calderon et al., 2018; Cisek & Kalaska, 2005; Donner et al., 2009). Response information became detectable around 5 seconds after the onset of the preview disk, which was around the same time that spatial decoding accuracy in visual cortex began to show a difference between the predictable and unpredictable conditions (Figure 3.3C, 3.4B). The time resolution of this experiment does not allow us to draw firm conclusions about the relative timing of each process, but these results are broadly consistent with a theory in which response selection and decision making unfold in parallel within sensorimotor regions (Cisek & Kalaska, 2010; Donner et al., 2009; Klein-Flügge & Bestmann, 2012; van Ede et al., 2019). Additionally, information about planned responses declined slightly toward the end of the delay period, dropping below chance just before the response probe appeared (Figure 3.4B). This time course is consistent with a transient signal related to the initial formation of a response plan, which aligns with recent EEG findings in human motor cortex following a retro-cue in a visual WM task (Gresch, Boettcher, Nobre, & van Ede, 2020). More precisely measuring the time course of spatial information and motor planning in this task remains an avenue for future work.

Overall, our findings suggest that the neural mechanisms that underly WM are dynamic and can adjust to the demands of the current cognitive task. When subjects had the option to divert to a motor-based strategy, spatial position discriminability dropped in early visual and parietal cortex. These findings may provide a partial reconciliation of apparently disparate findings from human and nonhuman primate WM research, which have placed different levels of emphasis on sensory-like codes for WM in visual cortex. Our experiment also highlights the

importance of considering task constraints when making comparisons between the results of different paradigms and across different species.

Materials and Methods

Participants. 6 subjects (2 male) between the ages of 20 and 34 were recruited from the UCSD community (mean age 27.2 $\pm$ 2.7 years), having normal or corrected-to-normal vision. An additional subject (male) participated in an early pilot version of the study, but did not complete the final version of the experiment and is not included here. The study protocol was submitted to and approved by the Institutional Review Board at UCSD, and all participants provided written informed consent in accordance with this IRB. Each subject performed a behavioral training session lasting approximately 30 minutes, followed by 3 or 4 scan sessions, each lasting approximately 2 hours. Participants were compensated at a rate of $10/hour for behavioral training and $20/hour for the scanning sessions. Subjects were also given "bonus" money for correct performance on certain trials in the main behavioral task (see *Main Behavioral Task in the Scanner)*, up to a maximum of $40 total.

Magnetic Resonance Imaging (MRI). All MRI scanning was performed on a General Electric (GE) Discovery MR750 3.0T research-dedicated scanner at the UC San Diego Keck Center for Functional Magnetic Resonance Imaging (CFMRI). Functional echo-planar imaging (EPI) data were acquired using a Nova Medical 32-channel head coil (NMSC075-32-3GE-MR750) and the Stanford Simultaneous Multi-Slice (SMS) EPI sequence (MUX EPI), with a multiband factor of 8 and 9 axial slices per band (total slices = 72; 2 mm^3 isotropic; 0 mm gap; matrix = 104 x 104; FOV = 20.8 cm; TR/TE = 800/35 ms; flip angle = 52°; inplane acceleration

= 1). Image reconstruction procedures and un-aliasing procedures were performed on servers hosted by Amazon Web Services, using reconstruction code from CNI (Center for Neural Imaging at Stanford). The initial 16 TRs collected at sequence onset served as reference images required for the transformation from k-space to the image space. Two short (17 s) "topup" datasets were collected during each session, using forward and reverse phase-encoding directions. These images were used to estimate susceptibility-induced off-resonance fields (Andersson, Skare, & Ashburner, 2003) and to correct signal distortion in EPI sequences using FSL topup functionality (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012).

Each subject participated in at least three sessions of functional scanning for this study, as well as a separate retinotopic mapping session. During the retinotopic mapping session, we also acquired a high-resolution anatomical scan, which produced higher quality contrast between grey and white matter and was used for segmentation, flattening, and visualizing retinotopic mapping data. For 4 out of the 6 subjects, the anatomical scan was obtained using an Invivo 8-channel head coil with accelerated parallel imaging (GE ASSET on a FSPGR T1-weighted sequence; 1x1x1 mm^3 voxel size; 8136 ms TR; 3172 ms TE; 8° flip angle; 172 slices; 1 mm slice gap; 256x192 cm matrix size), and for the remaining 2 subjects this scan was collected using the same 32-channel head coil used for functional scanning. Anatomical scans collected with the 32-channel head coil were corrected for inhomogeneities in signal intensity using GE's "Phased array uniformity enhancement" (PURE) method.

Pre-Processing of MRI Data. All pre-processing of MRI data was performed using software tools developed and distributed by FreeSurfer and FSL (available at https://surfer.nmr.mgh.harvard.edu and http://www.fmrib.ox.ac.uk/fsl). First, the recon-all utility in the FreeSurfer analysis suite (Dale, Fischl, & Sereno, 1999) was used to perform cortical

surface gray-white matter volumetric segmentation of anatomical T1 scans. The segmented T1 data were used to define cortical meshes on which to define retinotopic ROIs used for subsequent analyses (see *Retinotopic Mapping Procedures)*, and were also used to align cross-session data into a common space, as follows. For each scan session, the first volume of the first run of the session was used as a template to align the functional data from that scan session to the anatomical data. Co-registration was was performed using FreeSurfer's manual and automated boundary-based registration tools (Greve & Fischl, 2009). The resulting transformation matrices were then used to transform every four-dimensional functional volume into a common space, using FSL FLIRT (Jenkinson, Bannister, Brady, & Smith, 2002; Jenkinson & Smith, 2001). Next, motion correction was performed using FSL MCFLIRT (Jenkinson et al., 2002), without spatial smoothing, with a final sinc interpolation stage, and 12 degrees of greedom. Finally, slow drifts in the data were removed using a high-pass filter (1/40 Hz cutoff). No additional spatial smoothing was performed.

Following this initial pre-processing, we normalized the time series data from each scan run by z-scoring each voxel's signal across the entire run. This and all subsequent analyses were performed in Matlab 9.5 (Natick, MA). We then epoched the data based on the start time of each trial. Since trial events were jittered slightly with respect to TR onsets, we rounded trial start times to the nearest TR. This epoched data was used for all time-resolved decoding analyses (Figure 3.3, Figure 3.4). For data from the main task (see *Main Behavioral Task in the Scanner)*, we also obtained a single estimate for the delay period signal in each voxel on each trial, by taking an average over the timepoints from 10-16 TRs after trial onset. For data from the spatial working memory mapping task (see *Spatial Working Memory Mapping Task)*, we obtained a

single estimate for the delay period signal in each voxel by averaging over the timepoints from 6-12 TRs after trial onset.

Univariate Analyses. In order to estimate a hemodynamic response function (HRF) for each voxel during trials in each condition (Figure 3.2), we used linear deconvolution. This was done by constructing a finite impulse response model (Dale, 1999), including a series of regressors for each task condition: one regressor marking the onset of spatial memory target items for that task condition, followed by a series of temporally shifted versions of that regressor to model the BOLD response at each subsequent time point in the trial. The model also included a constant regressor for each of the 20 total runs. The data used as input to this GLM was z-scored on a voxel-by-voxel basis within runs as described in the previous section. Estimated HRFs were then averaged across all voxels within each ROI. To evaluate whether the mean BOLD signal in each ROI differed significantly between conditions, we used a permutation test. First, for each ROI and timepoint, we computed a Wilcoxon signed rank statistic comparing the activation values for each subject from condition 1 to the activation values for each subject from condition 2. Then, we performed 1000 iterations of shuffling the condition labels within each subject (swapping the condition labels for each subject with 50% probability). We then computed a signed rank statistic from the shuffled values. Finally, we computed a two-tailed p value for each ROI by computing the number of iterations on which the shuffled signed rank statistic was >= the real statistic, and the number of iterations on which the shuffled statistic was <= the real statistic, and taking the smaller of these two values. We obtained the final p-value by dividing this value by the total number of iterations and multiplying by 2.

General Stimulus Presentation Procedures. For all tasks described here, stimuli were projected onto a screen 21.3 cm wide x 16 cm high, fixed to the inside of the scanner bore just

above the subject's chest. The screen was viewed through a tilted mirror attached to the headcoil, with a viewing distance of 49 cm. This resulted in a maximum vertical eccentricity of 9.3°. The background was always a mid-gray color (RGB=[128,128,128]), and the fixation point was always a black circle with radius 0.2°. All stimuli were generated using Ubuntu 14.04, Matlab 9.3, and the Psychophysics toolbox (Brainard, 1997; Kleiner et al., 2007).

Retinotopic Mapping Procedures. We followed previously published retinotopic mapping protocols to define the visual areas V1, V2, V3, V3AB, hV4, IPS0, IPS1, IPS2, and IPS3 (Engel, Glover, & Wandell, 1997; Jerde & Curtis, 2013; Sereno et al., 1995; Swisher, Halko, Merabet, McMains, & Somers, 2007; Wandell, Dumoulin, & Brewer, 2007; Winawer & Witthoft, 2015). Subjects performed mapping runs in which they viewed either a contrast-reversing black and white checkerboard stimulus (4Hz) configured as a rotating wedge (10 cycles, 36 s/cycle), an expanding ring (10 cycles, 32 s/cycle), or a bowtie (8 cycles, 40 s/cycle). To increase the quality of data from parietal regions, subjects performed a covert attention task on the rotating wedge stimulus, which required them to detect contrast dimming events that occurred occasionally (on average, 1 event occurred every 7.5 seconds) in a row of the checkerboard (mean accuracy = 74.4 $\pm$ 3.6%). The maximum eccentricity of the stimulus was limited to 9.3° (degrees visual angle).

Main Behavioral Task in the Scanner. For all main task runs, subjects performed a spatial working memory task in the scanner. Each trial of the main task began with the fixation point turning green for 750 ms, to alert the subject that the spatial memory target was about to appear. Immediately following this, a white dot (radius=0.15°) appeared for 500 ms in the periphery of the screen. Subjects were required to remember the precise position of this target dot. The target dot could appear anywhere on an imaginary ring 7° away from fixation, with a

random polar angle position (see below paragraph for details on polar angle position grid). Next, the fixation point turned back to black for 1 second, then turned either red (RGB=[100]) or blue (RGB=[001]) for 2 seconds. This color cue indicated to the subject whether the current trial was a "predictable" or "unpredictable" trial (the mapping between color and task was counter-balanced across subjects, but fixed for all runs within a given subject). Next, a disk stimulus appeared for 1 second. This stimulus consisted of a circle 10° in radius, divided into two equal halves, with each side colored in a different shade of gray (dark gray RGB=[102.5, 102.5, 102.5], light gray RGB=[153.5, 153.5, 153.5]). The disk could be rotated about its center by an amount between 0° and 360°. To avoid the disk overlapping with the fixation point, an aperture of radius 0.4° was cut out of the middle of the disk, creating a donut-like shape. The inner and outer edges of the donut were smoothed with a Gaussian kernel implemented using fspecial.m (size=0.5°, sigma=0.5°), without disrupting the sharpness of the boundary between the two halves of the disk. The first appearance of this disk stimulus was the "preview" disk presentation, and was followed by a 12 second delay period. Following the delay period, another disk stimulus appeared for 2 seconds, serving as the response probe. At this point, subjects responded with a button press to indicate which side of the disk stimulus the memory position dot had been presented on (e.g. light gray or dark gray). Subjects were instructed not to make any physical finger movements until the response probe appeared. Responses were always made with the left or right index finger, with the mapping between luminance and finger counter-balanced across sessions within each subject. If the trial was in the predictable condition, the orientation of the probe stimulus was identical to the preview disk stimulus shown at the beginning of the delay period, but if the trial was in the unpredictable condition, the orientation of the preview disk was random with respect to the final disk orientation. Thus, for the predictable condition, subjects

had complete knowledge of the required motor response as soon as the preview disk appeared, but for the unpredictable condition, they had no ability to predict the required motor response. Trials of the two task conditions (e.g. predictable/unpredictable) were randomly interleaved within every scan run. At the start of each scan run, the subject was shown an instruction screen which reminded them of the color/task mapping and the luminance/finger mapping that was in effect for that session.

Each run of the main task included 20 total trials, with each trial followed by an inter-trial interval jittered randomly within the range 1-5 seconds. The total length of each run was 466 seconds. Subjects performed 10 runs of this task in each of two scan sessions for a total of 20 runs or 400 trials. All counter-balancing of stimulus positions and task conditions was performed within each session separately. Within each session, there were 100 trials of each task, and each of these 100 trials had a unique polar angle position, with positions sampled uniformly from a 100 point grid (e.g. the minimum spacing between possible positions was 3.6°). The rotation of the response probe disk stimulus on each trial, which determined the spatial boundary to which the memory position was compared, also took on 100 unique, uniformly spaced values within each task in each session. The grid of possible disk rotation increments was shifted by 1.8° relative to the grid of spatial memory positions, so that the memory position was never exactly on the boundary of the disk. To ensure that the joint distribution of memory positions and boundary positions was close to uniform, we broke the 100 point grid for each variable into 10 bins of 10 positions each. Across all 100 trials of each task, each combination of the bin for spatial memory position and the bin for boundary position was used once. For the predictable condition, the preview disk always took on the same rotation value as the response probe disk, but for the unpredictable condition, the uninformative preview disk stimulus was assigned a

random rotation amount. Across all unpredictable trials within each session, the rotation of the preview disk took on the same 100 values as the response probe disk rotation, but in a random order. The final trial sequence for each session was generated by randomly shuffling together all the trials from both task conditions, and splitting them evenly into 10 runs. As a result, the proportion of trials from each task condition and stimulus position bin was balanced across each session, but not balanced within individual runs.

To encourage subjects to encode the spatial positions with high precision, we rewarded subjects monetarily for correct performance on "hard" trials where the spatial memory item was close to the boundary. These "hard" trials were identified as those where the spatial memory item and the boundary belonged in the same bin, according to the angular position bins described in the above paragraph. Subject received $1 for correct performance on each "hard" trial, for a maximum of $40. Across subjects, the average bonus received was $32.83 +/- 2.86.

Spatial Working Memory Mapping Task. Subjects also performed an additional working memory task while in the scanner, which served as the training set for our classification analyses (see *Spatial Position Decoding*). As in the main working memory task, each trial began with the fixation point briefly turning green, followed by a spatial memory target item. The timing and visual appearance of these events was identical to those in the main task. The disappearance of the target memory item was followed by a 12 second delay period, then a white probe dot (radius=0.15°) appeared in the periphery of the screen, at the same eccentricity at which the memory dot had been presented (7°). Subjects were required to move this probe dot to the position at which the original memory item had been presented, using the four fingers of their right hand to press different buttons. The four buttons corresponded to movements of the dot quickly (120°/s) or slowly (40°/s) in a counter-clockwise or clockwise direction. The response

period lasted 3 seconds, and the position of the dot at the end of this period was taken as the subject's final response. The start position of the probe dot was random with respect to the position of the target memory dot.

Each run of the spatial working memory mapping task consisted of 20 trials, with trials separated by an inter-trial interval jittered randomly within the range 1-5 seconds. The total run length was 406 seconds. Subjects performed 10 runs of this task, all taking place during a single scan session. Across all 200 trials of the task, the spatial memory positions took on 200 distinct values uniformly spaced between 0-360° (e.g. spacing between adjacent points was 1.8°). To create a more even sampling of this space within individual runs, we binned these 200 positions into 20 bins of 10 positions each, and generated a sequence where each run sampled from each bin once. The random starting position of the probe dot on each trial was generated using an identical procedure, but independently of the memory positions, so that there was no association between the spatial memory item position and the start position of the probe. Average angular error on this task was $7.0 \pm 0.9°$.

Spatial Localizer Task. To identify voxels having spatial selectivity within the range of positions spanned by the memory positions, we used an independent mapping task for thresholding. In this task, participants viewed black and white checkerboard stimuli flickering at a rate of 4 Hz. Stimuli were wedges with a width of 15° (polar angle), spanning an eccentricity range of 4.4°-9.3° (visual angle), and were positioned at 24 different locations around an imaginary circle. The grid of possible wedge center positions was offset from the cardinal axes by 7.5° (e.g. a wedge was never centered on the horizontal or vertical meridian). Each run included 4 presentations of each position, for a total of 96 trials. The sequence of positions was random with the constraint that the stimulus never appeared in the same quadrant on back-to-

back trials. Trials were each 3 seconds long and were not separated by an inter-trial interval. The total run length was 313 seconds. During each run, subjects performed a contrast change-detection task at fixation, where they responded with a button press anytime the fixation point brightness increased or decreased. A total of 20 brightness changes occurred in each run, at times that were random with respect to trial onsets. The magnitude of brightness changes was adjusted manually at the start of each run to control the difficulty. Average detection performance (hit rate) was 76.7 $\pm$ 4.2%. Subjects performed between 8 and 16 total runs of this task. For some subjects, some of these runs were collected as part of a separate experiment. In all cases, data from different sessions were combined using the cross-session alignment procedure outlined above (*Pre-Processing of MRI Data)*.

Data from this task were analyzed using a General Linear Model (GLM) implemented in FSL's FEAT (FMRI Expert Analysis Tool, version 6.00). Brain extraction (Smith, 2002) and pre-whitening (Woolrich, Ripley, Brady, & Smith, 2001) were performed on individual runs before analysis. Predicted BOLD responses for each wedge position were generated by convolving the stimulus sequence with a canonical gamma hemodynamic response (phase=0s, s.d.=3s, lag=6s). Individual runs were combined using a standard weighted fixed effects analysis. For each of the 24 wedge positions, we identified voxels that were significantly more activated by that position than all other positions ($p<0.05$, false discovery rate corrected). We then merged the sets of voxels that were identified by any of these 24 tests, and used this merged map to select voxels from each retinotopic ROI for further analysis. For the sensorimotor ROIs (see *Sensorimotor Cortex Localizer Task*), the spatial localizer was not used for thresholding.

Sensorimotor Cortex Localizer Task. To identify ROIs in motor and somatosensory cortex that were selective for contralateral index finger button presses, we used a button pressing

localizer task. Subjects attended to brief color changes at the fixation point and responded by pressing buttons with their left or right index finger. Color changes were either magenta (RGB=[200, 0, 226.6]) or cyan (RGB=[1.6, 79.1, 155.0]), and lasted 1 second each, separated by an inter-trial interval randomly jittered in the range of 2-6 seconds. Each run was 319 seconds long, and included 60 total trials, with the 30 trials of each color randomly interleaved. The color/finger mapping was switched on alternating runs. Subjects were instructed to respond as quickly as possible to each color change. Average performance on this task was 92.8 $\pm$ 3.0 % and average reaction time was 530 $\pm$ 23 ms.

Data from this task were analyzed using a general linear model in FSL's FEAT, as described previously (see *Spatial Localizer Task*). Predicted BOLD responses for each button press were generated by convolving the stimulus sequence with a gamma hemodynamic response function (phase=0s, s.d.=3s, lag=5s). We identified voxels that showed significantly more activation for the contralateral finger than the ipsilateral finger ($p<0.05$, false discovery rate corrected). This procedure was done separately within each hemisphere. We then defined each ROI by intersecting the map of above-threshold voxels with the anatomical definitions of Brodmann's areas identified by FreeSurfer's recon-all segmentation procedure (Dale et al., 1999; Fischl et al., 2008). The intersection of the functionally-defined mask with Brodmann's area (BA) 6 was used to define premotor cortex (PMC), BA 4 was used to define primary motor cortex (M1), and BA 1,2 and 3 were combined to define primary somatosensory cortex (S1) (Brodmann, 1909; Fulton, 1935; Penfield & Boldrey, 1937).

Spatial Position Decoding. Linear classification was used to measure the representation of spatial position information in each ROI. Before performing classification, we first mean-centered the voxel activation pattern from each trial by subtracting the mean across all voxels

from each trial. To perform decoding, we first binned all trials of the main task (see *Main Behavioral Task in the Scanner)* and the spatial working memory mapping task (see *Spatial Working Memory Mapping Task*) into 8 angular position bins that were each 45° in size, with the first bin centered at 0°. We then performed binary classification between pairs of bins that were 180° apart, using a linear classifier based on the normalized Euclidean distance (more details given in Henderson & Serences, 2019). This meant that we constructed four separate classifiers, each operating on approximately ¼ of the data, with a chance decoding value of 50%. For the main analyses (Fig 2, Fig S1), the training set for these classifiers consisted of data from the spatial working memory mapping task (for trials in the two position bins of interest), and the test set consisted of data from the main task. For the within-condition analyses (Fig 3), the training and testing sets both consisted of data from a single task condition. The classifier was cross-validated by leaving out two trials at a time (leaving out one trial with each of the two labels ensured that the training set was always perfectly balanced), and looping so that every trial served as a test trial once. Decoding accuracy was evaluated for each binary decoder within test set trials from each task condition separately, and performance was averaged over the four decoders to give a single value of accuracy for each ROI and task condition.

To test whether decoding performance was significantly above chance in each ROI and task condition, permutation testing was used. On each of 1000 iterations, we shuffled the labels for the training set data and trained a classifier on this shuffled data, then computed its accuracy at predicting the labels for each task condition. For each iteration, we then computed a Wilcoxon signed rank statistic comparing the 6 real decoding values to the 6 shuffled decoding values, implemented with custom code. A signed rank statistic greater than 0 indicated the median of the real decoding values was greater than the median of the shuffled decoding values, and a statistic

less than zero indicated the median of the null decoding values was greater than the median of the real values. We obtained a one-tailed p-value for each ROI and task condition across all subjects by counting the number of iterations on which the signed rank statistic was less than or equal to zero, and dividing by the total number of iterations.

To test whether decoding performance differed significantly between the two task conditions within each ROI, we used a permutation test. First, for each ROI, we computed a Wilcoxon signed rank statistic comparing the 6 decoding values from condition 1 to the 6 decoding values from condition 2. Then, we performed 1000 iterations of shuffling the condition labels within each subject (swapping the condition labels for each subject with 50% probability). We then computed a signed rank statistic from the shuffled values. Finally, we computed a two-tailed p value for each ROI by computing the number of iterations on which the shuffled signed rank statistic was >= the real statistic, and the number of iterations on which the shuffled statistic was <= the real statistic, and taking the smaller of these two values. We obtained the final p-value by dividing this value by the total number of iterations and multiplying by 2.

The above procedures were used for both time-averaged (see *Pre-Processing of MRI Data)* and time-resolved decoding analyses. For the main time-resolved decoding analyses (Fig 3), the training set data always consisted of spatial working memory localizer data averaged over a fixed window in the delay period (see *Pre-Processing of MRI Data)*, and the testing set consisted of main task data at the TR of interest. For the within-condition time-resolved decoding analyses (Fig S2), the training set consisted of data at the TR of interest, and the testing set consisted of data at the same TR (cross-validated using the procedure described above). The signed rank tests described above were used to measure significance of decoding within each condition and timepoint and to compare decoding between conditions within each timepoint.

For the time-averaged decoding accuracies, we also performed a two-way repeated measures ANOVA with factors of ROI, condition, and ROI x condition interaction, implemented using ranova.m. We obtained p-values for each effect by performing a permutation test where we shuffled the decoding scores within each subject 1000 times, and computed the F statistic for each effect on the shuffled data. The p-value corresponding to each effect was the number of times the shuffled F statistic for that effect was greater than or equal to the real F statistic, divided by the total number of iterations (similar method used in Rademaker et al., 2019). F statistics reported in the text reflect the F statistic obtained using the real (unshuffled) data.

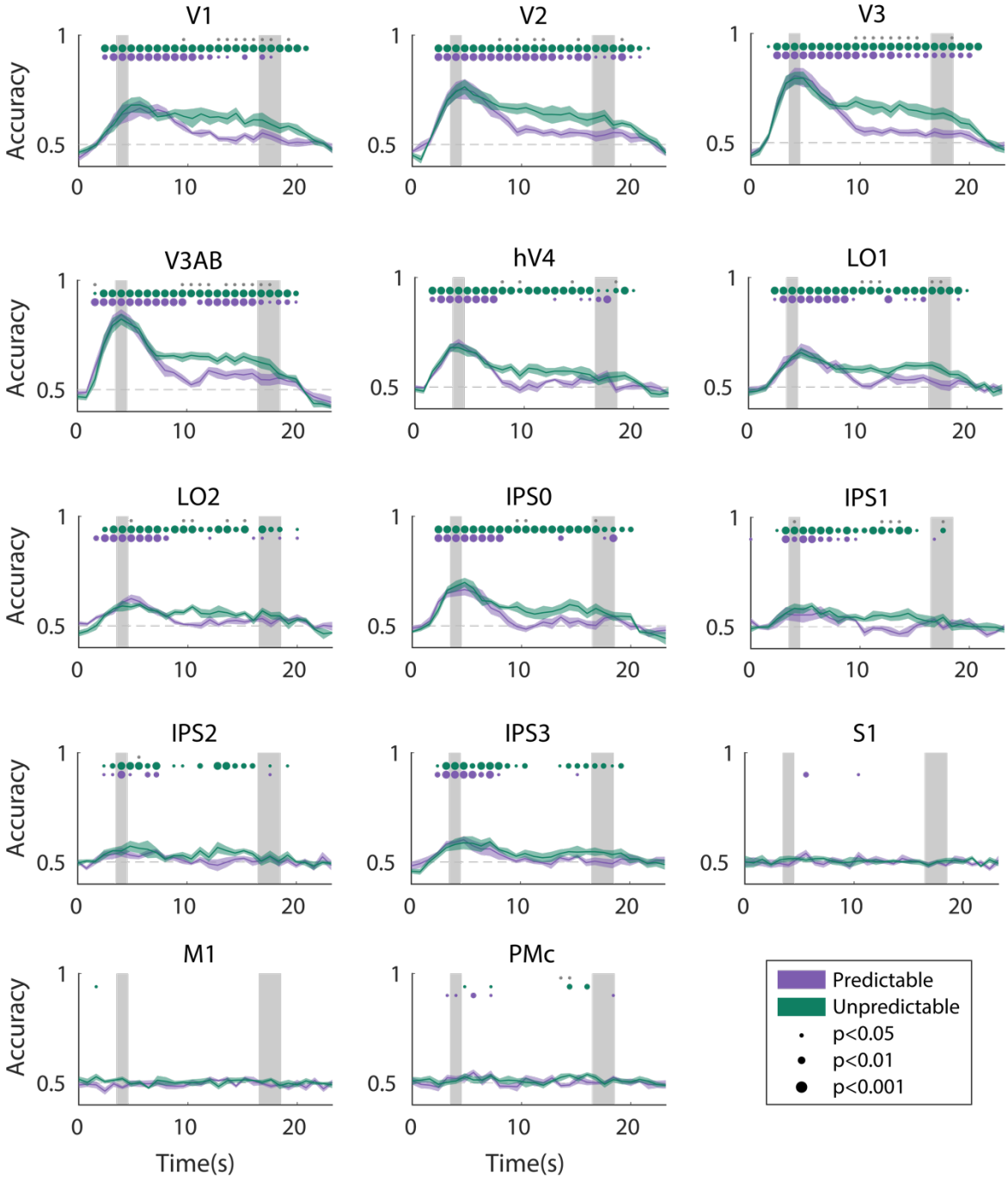
Response Decoding. We performed linear classification to measure the representation of information related to response planning in each ROI. Response classification was always done using data from one task condition at a time. To perform classification, all trials were labeled according to the finger (left or right index) that corresponded to the correct response on that trial. For the main analyses shown in this paper, this was done using all trials (including trials where the incorrect button or no button was pressed), as this ensured that the training set for the classifier was perfectly balanced, but similar results were obtained when using correct trials only, or when using the subject's actual response as the label for the decoder. We performed decoding using a linear classifier based on the normalized Euclidean distance (more details given in Henderson & Serences, 2019). The decoder was always trained on data from one session and tested on the other session. Because the mapping of disk side luminance to finger was always switched between the two sessions, this ensured that the information detected by the classifier was not related to the luminance of the disk side corresponding to the response.

The above procedure was used for both time-averaged and time-resolved response decoding. For time-resolved decoding, the training and testing set both consisted of data from the

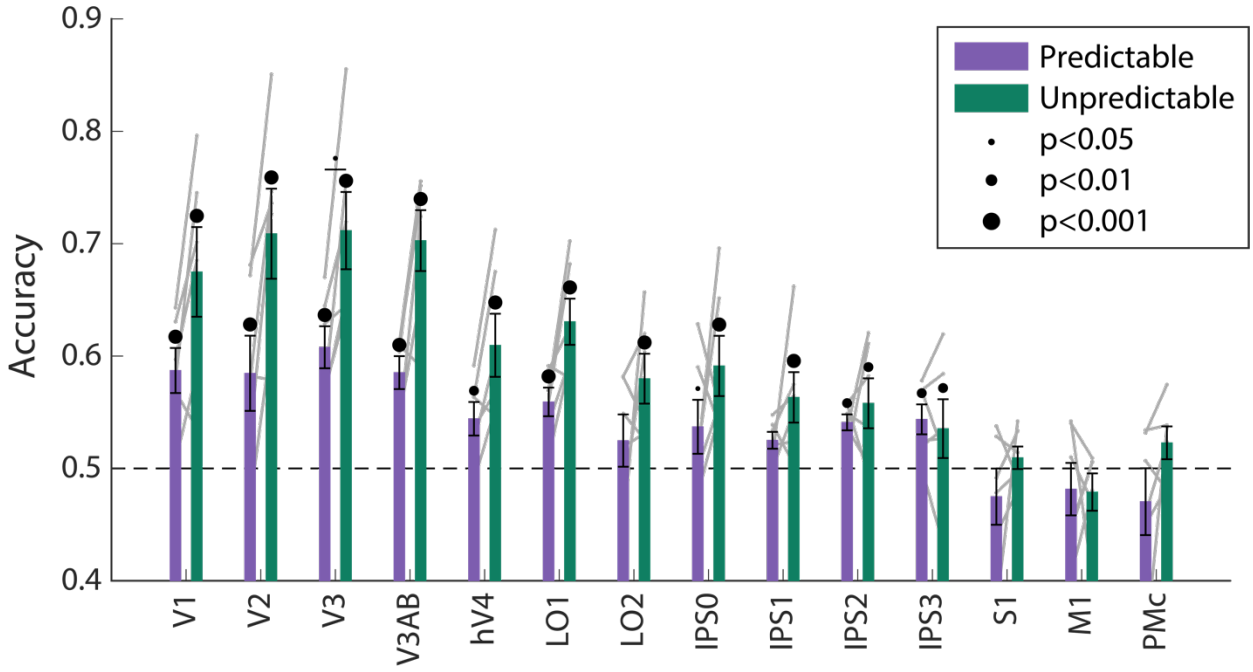
TR of interest (but different sessions, as described above). All statistical tests on the results of response decoding were performed in an identical manner to the statistics on the results of spatial decoding (see *Spatial Position Decoding*).

Chapter 3 is currently in preparation for publication. The author list and working title is Henderson, M. M., Rademaker, R. L., & Serences, J.T. (2021). Prospective response planning degrades spatial memory representations in human visual cortex. *In prep*. The dissertation author was the primary investigator and author of this paper.

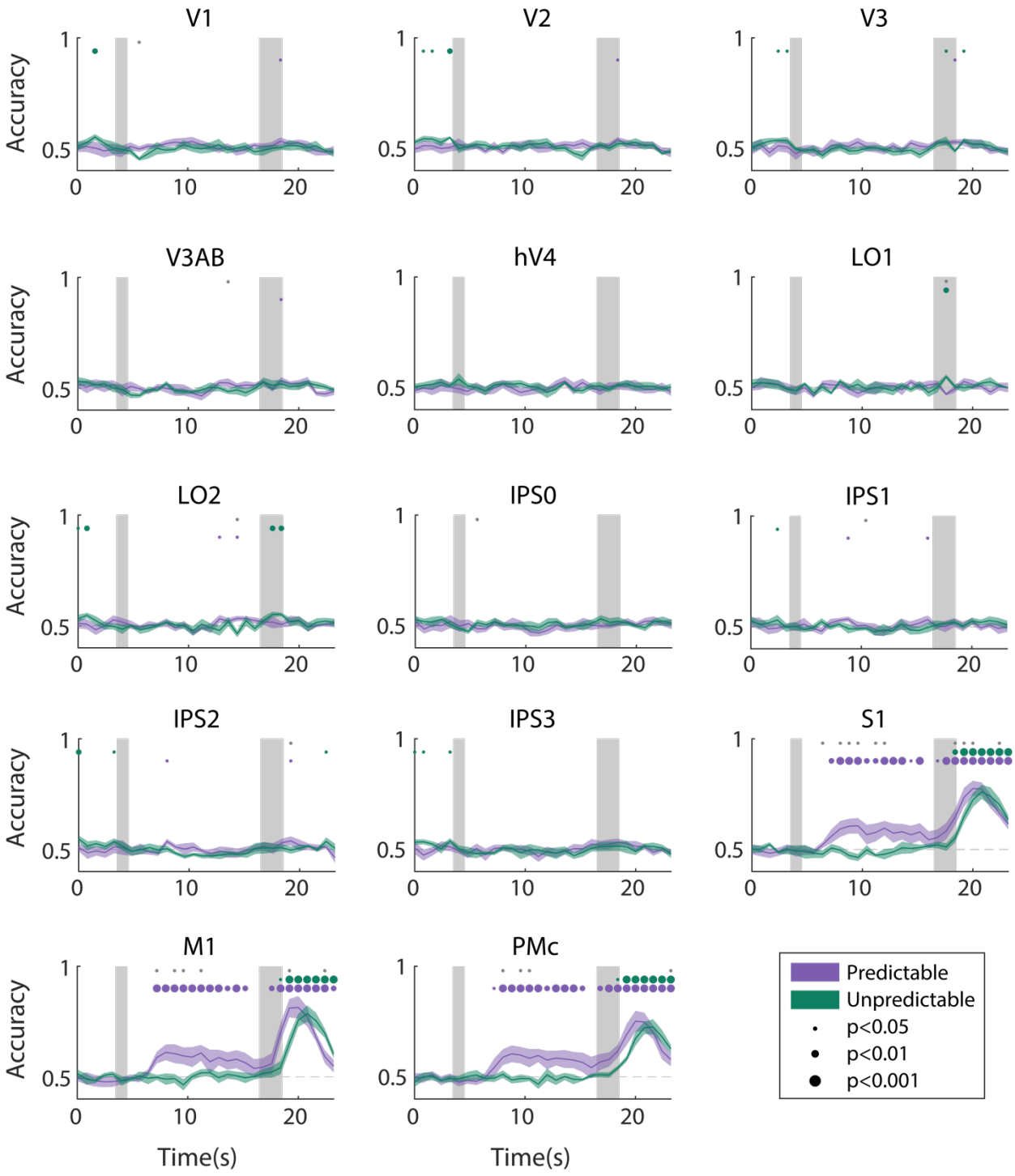
Supplementary Figures



Supplementary Figure 3.1. Time-resolved spatial decoding accuracy in every ROI. All decoding was done using the spatial working memory mapping task as a training set (see *Methods: Spatial Position Decoding* for details). Shaded gray rectangles indicate the periods of time when the "preview" disk was onscreen (3.5-4.5 sec) and when the response probe disk was onscreen (16.5-18.5 sec). Shaded error bars represent ±1 SEM across subjects, colored dots indicate significance of decoding within each condition, and gray dots indicate significant condition differences, with dot sizes reflecting significance levels.



Supplementary Figure 3.2. Spatial decoding performance differs across conditions, even when training and testing a decoder within each task condition separately. See *Methods: Spatial Position Decoding* for details on classification procedure. Error bars reflect ±1 SEM across subjects, and light gray lines indicate individual subjects. Dots above bars and pairs of bars indicate the statistical significance of decoding within each condition, and of condition differences, respectively, both evaluated using non-parametric statistics. Dot sizes reflect significance level.



Supplementary Figure 3.3. Time-resolved response decoding accuracy in every ROI. All decoding was done using data from the same task condition for training and testing (see *Methods: Response Decoding* for details). Shaded gray rectangles indicate the periods of time when the "preview" disk was onscreen (3.5-4.5 sec) and when the response probe disk was onscreen (16.5-18.5 sec). Shaded error bars represent ±1 SEM across subjects, colored dots indicate significance of decoding within each condition, and gray dots indicate significant condition differences, with dot sizes reflecting significance levels.

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, *20*(2), 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7

Ariani, G., Pruszynski, J. A., & Diedrichsen, J. (2020). Motor planning brings human primary somatosensory cortex into movement-specific preparatory states. *BioRxiv*, 2020.12.17.423254. https://doi.org/10.1101/2020.12.17.423254

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11239812

Boynton, G. M. (2011). Spikes, BOLD, attention, and awareness: A comparison of electrophysiological and fMRI signals in V1. *Journal of Vision*, *11*(5), 12–12. https://doi.org/10.1167/11.5.12

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9176952

Brodmann, K. (1909). Physiology of the cortex as an organ. In *Brodmann's Localisation in the Cerebral Cortex* (pp. 239–262). https://doi.org/10.1007/0-387-26919-3_10

Calderon, C. B., Van Opstal, F., Peigneux, P., Verguts, T., & Gevers, W. (2018). Task-Relevant Information Modulates Primary Motor Cortex Activity Before Movement Onset. *Frontiers in Human Neuroscience*, *12*, 93. https://doi.org/10.3389/fnhum.2018.00093

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814. https://doi.org/10.1016/j.neuron.2005.01.027

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience*, Vol. 33, pp. 269–298. https://doi.org/10.1146/annurev.neuro.051508.135409

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

Curtis, C. E., Rao, V. Y., & D'Esposito, M. (2004). Maintenance of Spatial and Motor Codes during Oculomotor Delayed Response Tasks. *Journal of Neuroscience*, *24*(16), 3944–3952. https://doi.org/10.1523/JNEUROSCI.5640-03.2004

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761–772. https://doi.org/10.1098/rstb.2007.2086

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2–3), 109–114. https://doi.org/10.1002/(SICI)1097-0193(1999)8:2/33.0.CO;2-W

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of Choice-Predictive Activity in Human Motor Cortex during Perceptual Decision Making. *Current Biology*, *19*(18), 1581–1585. https://doi.org/10.1016/j.cub.2009.07.066

Emrich, S. M., Riggall, A. C., La Rocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience*, *33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Engel, S., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*(2), 181–192. https://doi.org/10.1093/cercor/7.2.181

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience*, *29*(48), 15258–15265. https://doi.org/10.1523/JNEUROSCI.4388-09.2009

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, *87*(4), 893–905. https://doi.org/10.1016/j.neuron.2015.07.013

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T. T., … Zilles, K. (2008). Cortical Folding Patterns and Predicting Cytoarchitecture. *Cerebral Cortex August*, *18*, 1973–1980. https://doi.org/10.1093/cercor/bhm225

Fulton, J. F. (1935). A note on the definition of the "motor" and "premotor" areas. *Brain*, *58*(2), 311–316.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1016/j.neuron.2012.12.039

Funahashi, S., Chafee, M. V., & Goldman-Rakic, P. S. (1993). Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature*, *365*(6448), 753–756. https://doi.org/10.1038/365753a0

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science (New York, N.Y.)*, *173*(3997), 652–654. https://doi.org/10.1126/SCIENCE.173.3997.652

Gazzaley, A., & Nobre, A. C. (2012, February 1). Top-down modulation: Bridging selective

attention and working memory. *Trends in Cognitive Sciences*, Vol. 16, pp. 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Goense, J. B. M., & Logothetis, N. K. (2008). Neurophysiology of the BOLD fMRI Signal in Awake Monkeys. *Current Biology*, *18*(9), 631–640. https://doi.org/10.1016/j.cub.2008.03.054

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

Gresch, D., Boettcher, S. E. P., Nobre, A. C., & van Ede, F. (2020). Prospective action imprinting into visual working memory. *Journal of Vision*, *20*(11), 1017. https://doi.org/10.1167/jov.20.11.1017

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

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Henderson, M., & Serences, J. T. (2019). Human frontoparietal cortex represents behaviorally relevant target status based on abstract object features. *Journal of Neurophysiology*, *121*(4). https://doi.org/10.1152/jn.00015.2019

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1016/S1053-8119(02)91132-8

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

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

Jerde, T. A., & Curtis, C. E. (2013). Maps of space in human frontoparietal cortex. *Journal of Physiology Paris*, *107*(6), 510–516. https://doi.org/10.1016/j.jphysparis.2013.04.002

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

Klein-Flügge, M. C., & Bestmann, S. (2012). Time-dependent changes in human corticospinal excitability reveal value-based competition for action during decision processing. *Journal of Neuroscience*, *32*(24), 8373–8382. https://doi.org/10.1523/JNEUROSCI.0270-12.2012

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? A free cross-platform toolkit for psychophysiscs with Matlab and GNU/Octave. In *Cognitive and Computational Psychophysics* (Vol. 36). https://doi.org/10.1068/v070821

Lawrence, S. J. D., Van Mourik, T., Kok, P., Koopmans, P. J., Norris, D. G., & De Lange

Correspondence, F. P. (2018). Laminar Organization of Working Memory Signals in Human Visual Cortex Highlights d Early visual cortex contains item-specific working memory signals. *Current Biology*, *28*, 3435–3440. https://doi.org/10.1016/j.cub.2018.08.043

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. https://doi.org/10.1038/35084005

Logothetis, N. K., & Wandell, B. A. (2004). INTERPRETING THE BOLD SIGNAL. *Annu. Rev. Physiol*, *66*, 735–769. https://doi.org/10.1146/annurev.physiol.66.082602.092845

Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. E., & D'Esposito, M. (2018). Flexible coding of visual working memory representations during distraction. *Journal of Neuroscience*, *38*(23), 5267–5276. https://doi.org/10.1523/JNEUROSCI.3061-17.2018

Lorenc, E. S., Vandenbroucke, A. R. E., Nee, D. E., de Lange, F. P., & D'Esposito, M. (2020). Dissociable neural mechanisms underlie currently-relevant, future-relevant, and discarded working memory representations. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-67634-x

Mendoza-Halliday, D., Torres, S., & Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nature Neuroscience*, *17*(9), 1255–1262. https://doi.org/10.1038/nn.3785

Miller, E., Erickson, C. a, & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(16), 5154–5167. https://doi.org/10.1.1.41.2959

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science (New York, N.Y.)*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017, June 1). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*, Vol. 21, pp. 449–461. https://doi.org/10.1016/j.tics.2017.03.010

Nobre, A. C., & Stokes, M. G. (2019, October 9). Premembering Experience: A Hierarchy of Time-Scales for Proactive Attention. *Neuron*, Vol. 104, pp. 132–146. https://doi.org/10.1016/j.neuron.2019.08.030

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006, September). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, Vol. 10, pp. 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. In *Brain* (Vol. 60). https://doi.org/10.1093/brain/60.4.389

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Simultaneous representation of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, *22*, 1336–1344. https://doi.org/10.1101/339200

Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Schneider, D., Barth, A., & Wascher, E. (2017). On the contribution of motor planning to the retroactive cuing benefit in working memory: Evidence by mu and beta oscillatory activity in the EEG. *NeuroImage*, *162*(April), 73–85. https://doi.org/10.1016/j.neuroimage.2017.08.057

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Sereno, M., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., … Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*(5212), 889–893. https://doi.org/10.1126/science.7754376

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, and Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Sprague, T. C., Boynton, G. M., & Serences, J. T. (2019). The importance of considering model choices when interpreting results in computational neuroimaging. *ENeuro*, *6*(6). https://doi.org/10.1523/ENEURO.0196-19.2019

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Sprague, T. C., Saproo, S., & Serences, J. T. (2015). Visual attention mitigates information loss in small- and large-scale neural codes. *Trends in Cognitive Sciences*, *19*(4), 215–226. https://doi.org/10.1016/j.tics.2015.02.005

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent

neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Supèr, H., Spekreijse, H., & Lamme, V. A. F. (2001). A neural correlate of working memory in the monkey primary visual cortex. *Science*, *293*(5527), 120–124. https://doi.org/10.1126/science.1060496

Swisher, J. D., Halko, M. A., Merabet, L. B., McMains, S. A., & Somers, D. C. (2007). Visual Topography of Human Intraparietal Sulcus. *Journal of Neuroscience*, *27*(20), 5326–5337. https://doi.org/10.1523/JNEUROSCI.0991-07.2007

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, *22*(3), 477–483. https://doi.org/10.1038/s41593-018-0335-6

Van Kerkoerle, T., Self, M. W., & Roelfsema, P. R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature Communications*, *8*(1), 1–14. https://doi.org/10.1038/ncomms13804

Vaziri-Pashkam, M., & Xu, Y. (2017). Goal-Directed Visual Processing Differentially Impacts Human Ventral and Dorsal Visual Representations. *The Journal of Neuroscience*, *37*(36), 8767–8782. https://doi.org/10.1523/JNEUROSCI.3392-16.2017

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383. https://doi.org/10.1016/j.neuron.2007.10.012

Winawer, J., & Witthoft, N. (2015). Human V4 and ventral occipital retinotopic maps. *Visual Neuroscience*, *32*, E020. https://doi.org/10.1017/S0952523815000176

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

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

Xing, Y., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2013). Decoding working memory of stimulus contrast in early visual cortex. *Journal of Neuroscience*, *33*(25), 10301–10311. https://doi.org/10.1523/JNEUROSCI.3754-12.2013

CONCLUSION

In this dissertation I have presented three experiments which demonstrate how the brain's limited processing capacity encourages selectivity in visual representations. In **Chapter 1,** I demonstrated how an artificial vision system learns to allocate many feature detectors to representing orientations that are commonly encountered during its training phase, supporting the idea that low-level biases in vision are related to the efficient coding of natural scene statistics. In **Chapter 2,** I evaluated how selectivity operates over faster time scales, using a task requiring subjects to determine the target status of currently viewed items. This experiment supported the role of frontal and parietal cortical regions in selection of target items during complex recognition tasks. In **Chapter 3,** I examined how the neural mechanisms of visual working memory (WM) can selectively adapt to the demands of a given task, representing remembered information in either a retrospective, sensory-like format, or a prospective, action-like format depending on how the information will ultimately be used. In this section, I discuss the broader relevance of this work and some of its limitations.

In **Chapter 1** I showed that non-uniformities in the visual statistics of images used to train a convolutional neural network (CNN) can lead to pronounced, systematic biases in the feature representations learned by the network. Though I demonstrated this principle for edge orientation, a low-level visual feature, a similar principle may hold for higher-level visual features such as object or face identity. For instance, if a computer vision model trained to perform face recognition is trained on a dataset comprised of mainly light-skinned male faces, it may perform poorly at identifying dark-skinned female faces (Buolamwini & Gebru, 2018). Given the current popularity of CNNs and related models for commercial applications, it is key to understand how dataset biases can influence model performance. Future work should

investigate how the effects on tuning properties and information content that I demonstrated for orientation might differ for other image properties.

In **Chapter 1** I described the characteristics of orientation biases in CNNs, but did not examine the role of these biases in supporting the network's task performance. Each CNN was trained to perform categorization of object images, a task that does not explicitly require fine orientation discrimination. Despite this, the networks developed orientation-tuned units and developed especially precise representations of orientations close to those most common in the training images. This suggests that the ability to discriminate orientations, particularly commonly encountered orientations, is beneficial for identifying objects. One way to investigate this further would be to selectively lesion units from a trained model, and evaluate how removal of different types of units impacts object recognition accuracy. If removing cardinal-tuned units selectively impairs object recognition performance, to a greater extent than removing oblique-tuned units or units with no orientation selectivity, this would suggest that orientation biases support object categorization. Alternatively, this could be tested by measuring network activations in response to test object images, and using gradient descent based methods to determine which units are most diagnostic for the network's category judgment (Bau et al., 2020). If cardinal-tuned units have a high weighting toward the network's final decision, this would support the theory that these units play an important functional role.

In **Chapter 2** I introduced an original stimulus set consisting of novel objects that have several unique properties. First, the images were designed so that the abstract features of identity and viewpoint were dissociated from the 2D projection made by each image onto the retina. As a result, they can be used to isolate information about high level object properties from lower-level shape information. Though I used these images within the context of a one-back matching task,

other tasks could be developed with this stimulus set to further examine how these high-level properties are computed in the brain. In addition, while other existing stimulus sets dissociate identity information from shape information, the novel objects presented here also introduce the concept of viewpoint as an abstract feature dissociable from shape. Specifically, we defined one viewpoint of the object arbitrarily as its "front" and taught subjects to identify viewpoints in a coordinate system relative to this view. The results of our experiment showed that subjects were capable of recognizing object viewpoint in this framework, but leave open the question of whether they were learning a true geometric representation of the coordinate system or using an alternative strategy. Future work could explore this issue further, for instance by determining whether the "front" of an object is represented distinctly from other views, or whether the similarity of images in a neural representational space maps onto the theorized distance of their viewpoint vectors.

An additional question that was not addressed in **Chapter 2** is how objects' identity and viewpoint were represented by the brain, independent of their status as matches in the current task. Multivariate decoding of fMRI data revealed robust representations of items' status as targets in the current task, but our initial analyses (not shown) showed weak decoding of both identity and viewpoint from all areas examined. This is likely due in part to the low spatial resolution of fMRI. Decoding of features such as orientation from fMRI data is thought to be driven in part by clustering of similarly-tuned neurons at a sub-millimeter scale (Boynton, 2011; Pratte, Sy, Swisher, & Tong, 2016). Though regions of the primate ventral visual system are known to encode representations of high-level features such as face identity, these population codes may lack the topographic clustering that characterizes representations of simpler features like orientation, thus making it difficult to reliably decode high-level object properties from

fMRI data (Dubois, de Berker, & Tsao, 2015). Single-unit electrophysiology in either humans or nonhuman primates (NHPs) may be required to fully reveal how the identity and viewpoint of our novel object images is computed within the visual system. If such a method were used, this would make it possible to address additional questions regarding how representations of currently viewed object properties are multiplexed with sought object properties, and how the task-relevance of object properties impacts the fidelity with which they are represented.

Chapter 3 showed that the brain can flexibly engage different coding schemes and cortical regions for maintaining information in visual WM, depending on task demands. This adds to a growing body of work suggesting that rather than having one fixed mechanism, WM can be supported by multiple neural mechanisms, including persistent spiking codes as well as more dynamic, temporally-evolving codes (Orhan & Ma, 2019; Stokes, 2015). More broadly, our work suggests that task demands can have a substantial influence on whether information will be decodable from a given brain region within the context of a particular experiment. As discussed in **Chapter 3**, this is particularly relevant when comparing experiments done in humans with experiments done in NHPs. In addition to the fact that many tasks used in classic NHP studies of WM allow for prospective motor response planning, tasks used with NHPs often require a much lower level of sensory detail than tasks used in human experiments (e.g. remembering one of four distinct pictures vs. remembering one of 180 similar orientations). As a result of this, the neural representations measured in NHP brains during task performance might be more coarse or categorical than those measured in the human brain (Serences, 2016). Additionally, NHPs are typically trained on a task for many months over the course of an experiment, and this over-training may lead to changes in the neural strategy used to solve a task. For instance, the relative importance of attentional gain and noise reduction as mechanisms for selective attention in

humans has been shown to shift over the course of repeated training sessions (Itthipuripat, Cha, Byers, & Serences, 2017). One avenue for future work would be to examine whether a similar principle holds for WM. For instance, it may be the case that highly trained human subjects rely more strongly on frontal cortex than sensory cortex for storage even when a task requires memory for detailed sensory features. Experiments such as this would provide a way to reconcile cross-species differences in findings, as well as evaluate the plasticity of the mechanisms supporting WM.

The results of **Chapter 3** suggest that when the mapping between a remembered sensory feature and the required response is fully predictable, information is re-mapped from a sensory-like to a motor-like representational format. However, this leaves some unresolved questions regarding the nature of this re-mapping. For example, the limited time resolution of fMRI (samples collected every 800 ms) and the delayed nature of the hemodynamic response function prevents us from measuring the precise timing with which spatial information degrades in sensory cortex, and with which response information appears in motor cortex. This experiment could be adapted for use with electroencephalography (EEG) in order to measure the temporal dynamics of this process more precisely. Specifically, we might predict that the drop in spatial decoding in sensory cortex would precede the appearance of response decoding in motor cortex. This would lend further support to the idea that information "flows" from sensory to motor cortex over the course of the delay period.

In sum, the research described in this dissertation demonstrates multiple ways in which the requirement for selectivity shapes neural processing in the visual system. This can be observed in the feedforward responses of sensory neurons, which demonstrate selectivity for stimuli that were frequently encountered during evolution and development of the visual system,

leading to tradeoffs in perception of other stimuli. Selection of relevant information can also be observed in the real-time behavior of animals, as when identifying relevant objects based on remembered information about sought object properties. Finally, when representing information across delays in WM, the brain appears to selectively re-map representations into a format most appropriate for the current task, which may be a way of maximizing coding efficiency. These observations illustrate how the brain's limited computational resources place a fundamental constraint on the function of the visual system. The ability to adapt to this constraint, by efficiently and selectively encoding sensory stimuli, appears to be a key function underlying adaptive human behavior.

References

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071–30078. https://doi.org/10.1073/pnas.1907375117

Boynton, G. M. (2011). Spikes, BOLD, attention, and awareness: A comparison of electrophysiological and fMRI signals in V1. *Journal of Vision*, *11*(5), 12–12. https://doi.org/10.1167/11.5.12

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *. In *Proceedings of Machine Learning Research* (Vol. 81). Retrieved from PMLR website: http://proceedings.mlr.press/v81/buolamwini18a.html

Dubois, J., de Berker, A. O., & Tsao, D. Y. (2015). Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(6), 2791–2802. https://doi.org/10.1523/JNEUROSCI.4037-14.2015

Itthipuripat, S., Cha, K., Byers, A., & Serences, J. T. (2017). Two different mechanisms support selective attention at different phases of training. *PLoS Biology*, *15*(6), e2001724. https://doi.org/10.1371/journal.pbio.2001724

Orhan, A. E., & Ma, W. J. (2019). A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature Neuroscience*, *22*(2), 275–283. https://doi.org/10.1038/s41593-018-0314-y

Pratte, M. S., Sy, J. L., Swisher, J. D., & Tong, F. (2016). Radial bias is not necessary for orientation decoding. *NeuroImage*, *127*, 23–33. https://doi.org/10.1016/j.neuroimage.2015.11.066

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004