Neural representation of action symbols in primate frontal cortex

Lucas Y. Tian[1,7]*, Kedar U. Garzón[1], Adam G. Rouse[2], Mark A. G. Eldridge[3], Marc H. Schieber[4], Xiao-Jing Wang[5], Joshua B. Tenenbaum[6,7], Winrich A. Freiwald[1,7]*

Abstract

At the core of intelligence is proficiency in solving new problems, including those that differ dramatically from problems seen before. Problem-solving, in turn, depends on goal-directed generation of novel thoughts and behaviors[1], which has been proposed to rely on internal representations of discrete units, or symbols, and processes that can recombine them into a large set of possible composite representations[1–11]. Although this view has been influential in formulating cognitive-level explanations of behavior, definitive evidence for a neuronal substrate of symbols has remained elusive. Here, we identify a neural population encoding action symbols—internal, recombinable representations of discrete units of motor behavior—localized to a specific area of frontal cortex. In macaque monkeys performing a drawing-like task designed to assess recombination of learned action symbols into novel sequences, we found behavioral evidence for three critical features that indicate actions have an underlying symbolic representation: (i) invariance over low-level motor parameters; (ii) categorical structure, reflecting discrete classes of action; and (iii) recombination into novel sequences. In simultaneous neural recordings across motor, premotor, and prefrontal cortex, we found that

1 Laboratory of Neural Systems, The Rockefeller University, New York, NY, USA

2 Department of Neurosurgery, Department of Cell Biology & Physiology, University of Kansas Medical Center, Kansas City, KS, USA

3 Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK

4 Department of Neurology, University of Rochester, Rochester, NY, USA

5 Center for Neural Science, New York University, New York, NY, USA

6 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

7 Center for Brains, Minds and Machines, MIT & Rockefeller University

* Co-corresponding authors: ltian@rockefeller.edu (L.Y.T.) and wfreiwald@rockefeller.edu (W.A.F.)

planning-related population activity in ventral premotor cortex encodes actions in a manner that, like behavior, reflects motor invariance, categorical structure, and recombination, three properties indicating a symbolic representation. Activity in no other recorded area exhibited this combination of properties. These findings reveal a neural representation of action symbols localized to PMv, and therefore identify a putative neural substrate for symbolic cognitive operations.

Acknowledgements

We thank D. Hanuska for animal care, help with spike sorting, and project feedback; Y. Liu, V. Goudar, X. Ma, T. Wu, D. Dolnik, S. Coolsaet, S. Sharma, A. Urquieta, V. Calligy, and other members of the Freiwald, Wang, and Tenenbaum labs for project feedback; F. Buck, D. Hildebrand, E. Hosseini, P. Jaffe, K. Kay, T. Nigam, Y. Prut, J. Rhee, P. Schade, T. Suzuki, and W. Zarco for manuscript feedback; V. Sherman and A. Gonzalez for technical assistance; and L. Ying for administrative assistance. This work was supported by the National Institutes of Health, through the National Eye Institute (R01EY021594 to W.A.F.), the National Institute Of Mental Health (F32MH125573 to L.Y.T.), and the National Institute Of Neurological Disorders And Stroke (K99NS131585 to L.Y.T.), as well as the Simons Foundation's Collaboration on the Global Brain (876120SPI and AN-NC-GB-Pilot Extension-00002596-01 to X.-J.W., J.B.T., and W.A.F., and NC-GB-CULM-00003138 to X.-J.W.), the Center for Brains, Minds & Machines of the National Science Foundation (STC award CCF-1231216 to W.A.F. and J.B.T.), the Office of Naval Research (N00014-23-1-2040 to X.-J.W., and MURI N00014-21-1-2801 to J.B.T.), and the Air Force Office of Scientific Research (FA9550-22-1-0387 to J.B.T.).

Author contributions

L.Y.T., X.-J.W., J.B.T., and W.A.F conceived the study and designed the drawing task. L.Y.T., A.G.R., M.A.G.E., M.H.S., and W.A.F performed the surgeries, with input from K.U.G. L.Y.T. performed experiments and data analysis, with input from K.U.G. L.Y.T. wrote the initial draft of the manuscript.

L.Y.T., X.-J.W., J.B.T., and W.A.F. revised the manuscript with input from the other authors. L.Y.T., X.-J.W., J.B.T., and W.A.F. acquired funding. X.-J.W., J.B.T., and W.A.F. supervised the project.

Competing interest

The authors declare no competing interest.

Introduction

Understanding the mechanisms of intelligence requires explaining the most powerful forms of generalization, in particular, generalization to new situations or problems differing considerably from those previously encountered. For example, if children are asked to “draw an animal that does not exist”, they can generalize from prior experience to produce an imaginary animal, such as a dog-like creature with six legs, three camel humps, and three pig tails[12]. An influential hypothesis is that this seemingly unbounded generalization depends on an internal representation of discrete units (symbols) that can be recombined into composite representations, in a process called compositional generalization[1–11]. Symbols support generalization because they enable the systematic derivation of a large set of possible representations from a smaller set of components, such as new animals imagined as rule-based combinations of discrete parts (e.g., *new animal = 1 torso + 8 arms + 4 legs*). Importantly, this hypothesis is not restricted to concepts that people explicitly represent as symbol systems in language[13], computer programs[14], or mathematics[9,10], but applies broadly even to abilities, such as drawing, which are not superficially symbolic, but which seem to reflect internal symbolic representations[5,7,11]. These abilities span various domains, including, in humans, geometry[7,15], perception[11], handwriting[16], drawing[17,18], dance[19], music[7], and speech[20,21], and, in non-human animals, logical reasoning[22,23], social cognition[24], navigation[25,26], artificial grammars[27–30], communication[26], numerical cognition[31,32], and physical reasoning[11].

Despite behavioral evidence for internal, symbolic representations, we lack definitive evidence for whether and how symbols are implemented in neuronal substrates. This is especially problematic in light of the uncertainty over how symbols reconcile with hypothesized mechanisms of cognition that do not presuppose symbols, including those based on distributed processing in neural networks[13,33–35], dynamical systems[36–39], and map-like representations[40,41]. Given that symbols are discrete units that are systematically recombined, a neural population representing symbols should exhibit three basic properties in its activity patterns: (i) invariance, (ii) categorical structure, and (iii) recombination. Invariance means that activity exhibits a form of abstraction in which it is largely independent of

variables that are irrelevant to the task goal. Categorical structure means the neural population should express a set of distinct activity patterns, one for each symbol, and be biased towards these discrete activity states even in the setting of continuous variation in task parameters. Recombination implies that a symbol's activity pattern should occur in all contexts in which it is used in composition with other symbols.

Recordings during cognitive tasks have revealed a diversity of invariant representations, including of rules[42,43], actions[44,45], sequences[28,46,47], numerical concepts[48–50], perceptual categories[51,52], visual objects[53], cognitive maps[41,54,55], and abstract concepts[56,57]. Collectively, these findings reveal a striking capacity for invariance in neural activity, and implicate critical roles for specific regions in this capacity, including prefrontal cortex[28,42,43,47,48,51,52] and medial temporal lobe[41,43,49,54,56,57]. However, it is unclear whether these representations exhibit the other properties expected for symbols. First, with a few exceptions[51,52], these prior studies did not systematically test for categorical structure by assessing whether activity varies discretely with continuous variation in task parameters. Second, evidence for recombination in neural activity is also rare, with a notable exception being hippocampal activity encoding spatial paths appearing to reuse parts of previous paths in novel sequences[58,59]. However, whether these continuous spatial paths reflect recombination of discrete, categorical components has not been tested. Third, an important point that applies to the navigation tasks in these hippocampal studies, and generally to tasks in prior studies finding invariant representations, is that they did not, to our knowledge, implement behavioral tests of compositional generalization, making it challenging to determine whether and how any identified neural representation contributes to the cognitive processes that support the ability to solve novel problems. For these reasons, we still lack evidence for a neural representation of symbols—one that jointly exhibits invariance, categorical structure, and recombination in the behavioral setting of compositional generalization.

To identify neural populations that may express these properties of symbols, we established a task paradigm involving symbol-based compositional generalization, implemented in macaque monkeys, a

species with extensive abilities for abstract cognition[28,42,43,47,48,52,54,60] (**Fig. 1a, b**). This paradigm was designed to capitalize on the brain's remarkable ability to generate novel, goal-directed actions, an ability hypothesized to involve symbolic representations in the form of discrete units of action (action symbols) that are recombined into sequences[16,17,19,21,61–68]. Action symbols may be especially important when complex actions must be generated to solve problems, such as in the problem of imitating a new dance by observation, which may depend on internal representations of symbols corresponding to dance poses[19,69]. Action symbols also form the basis of models that capture behavioral findings in action sequencing, including in handwriting[16,61], drawing[17,62], object manipulation[63], and tool use[64]. These studies suggest that a task involving compositional generalization in action sequencing would enable the systematic investigation of action symbols. Here, we establish such a task paradigm. We then show that behavior in this task exhibits invariance, categorical structure, and recombination, indicating an internal symbolic representation of action. In multi-area neuronal recordings, we found a neural population, localized to ventral premotor cortex (PMv), which encodes actions in a manner that reflects these three properties of symbols. This finding therefore identifies a neural substrate of action symbols in PMv.

Results

Learned stroke primitives in a drawing-like task paradigm.

We developed a drawing-like task paradigm for macaque monkeys modeled after studies of symbolic structure in human drawing[17] and handwriting[16,61]. We trained two subjects to draw geometric figures by tracing them on a touchscreen (**Fig. 1a,c,d**; videos of behavior in **Supplementary Videos 1-10**; experimental setup in **Extended Data Fig. 1**). On each trial, subjects were presented with an image of a figure, and were rewarded for making an accurate copy of the image, quantified as the spatial similarity between the drawing and image (primarily using the Hausdorff distance; see Methods).

Once subjects understood the core requirements of the task paradigm (i.e., to make accurate traces of images), they practiced drawing a diverse set of simple shapes, each using one stroke (**Fig. 1e**). Each subject converged on a set of consistently reused stroke spatio-temporal trajectories, one for each shape, which we call the subject's "primitives" (**Fig. 1e**). Analysis of stroke trajectories revealed that primitives were idiosyncratic to each subject and shape. We devised a "trajectory distance" metric, which measures the mean Euclidean distance between the velocity time series of two strokes (after normalizing scale and duration; see Methods). For each subject, each shape was drawn in a consistent manner across trials (example drawings in **Fig. 1f**; low trajectory distance in **Fig. 1g**). However, stroke trajectories differed for the same shape across subjects (compare drawings between subjects in **Fig. 1e,f**; high trajectory distance in **Fig. 1g**) and for the same subject across shapes (**Fig. 1e, g**).

These findings—that stroke primitives were idiosyncratic to each shape and subject—raised the possibility that these primitives reflect learned action symbols. To test this possibility, we tested how subjects generalize to draw new figures designed to assess three behavioral properties that, in combination, indicate an underlying symbolic representation: motor invariance, categorical structure, and recombination (**Fig. 1b**).

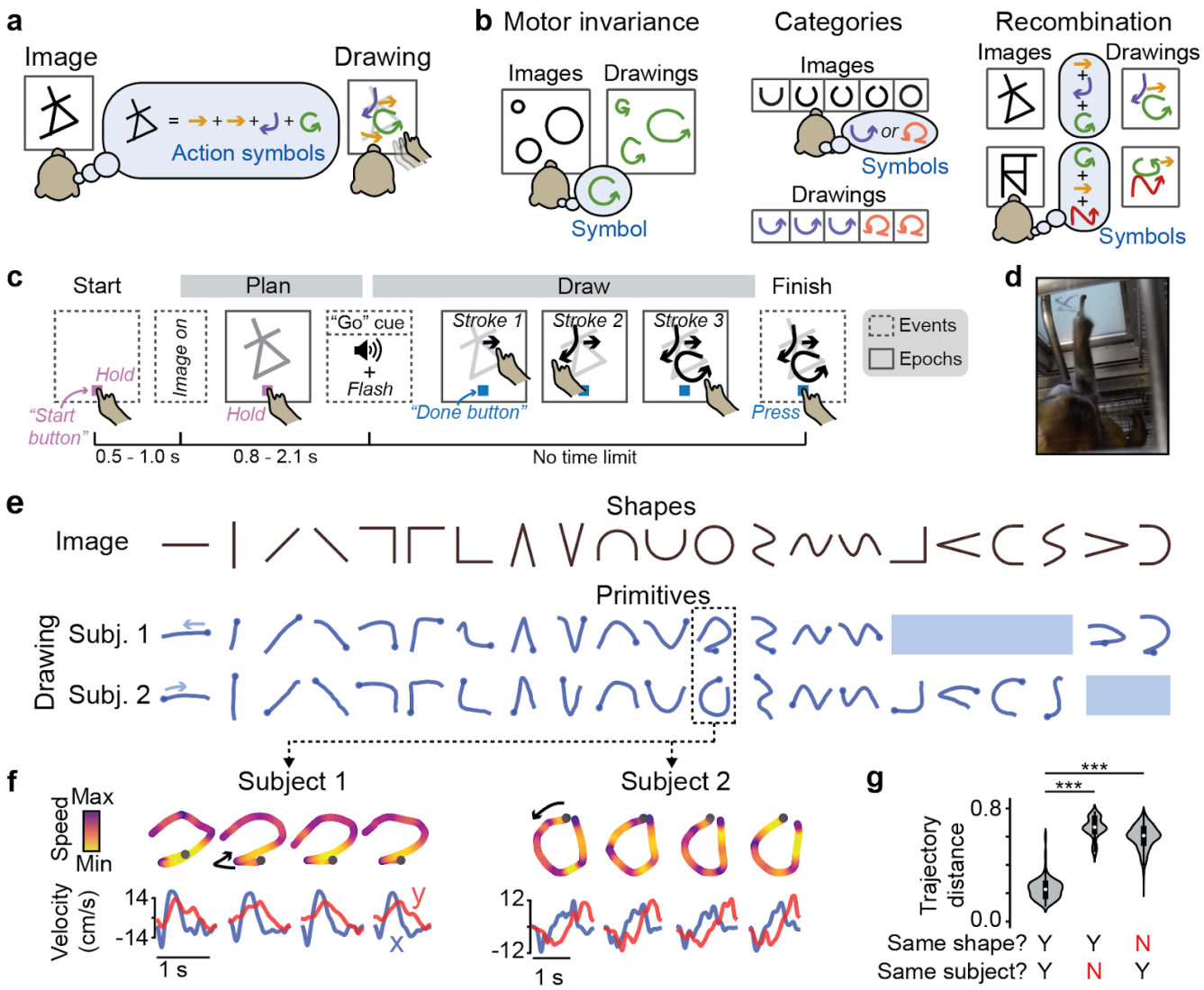


Fig. 1. Learned stroke primitives in a drawing-like task paradigm.

(a) Experimental paradigm. Given an image of a (potentially novel) figure, we hypothesize that subjects internally represent actions as a composite of learned symbols.

(b) Three essential features of a hypothesized symbolic representation, which we test in behavior and neural activity: invariance, categorical structure, and recombination. Schematics depict predicted behavior if subjects represent actions in a symbolic manner. Predictions consistent with alternative hypotheses are presented in **Figs. 2a, f, k**.

(c) Trial structure, showing a sequence of discrete events (dashed box) and sustained epochs (solid box) represented on a touchscreen. During the "plan" epoch, subjects see the image but are not allowed to remove their finger from the "start button". During the "draw" epoch, subjects produce strokes, and can report completion at any time by pushing the "done" button (which appears after the go cue). They then receive juice reward based on drawing performance (see Methods). The range of epoch durations reflects the range of mean durations across task variants and sessions (further trial-by-trial jitter was also added within each session; see Methods). Note that "buttons" refers to virtual touchscreen buttons.

(d) Photograph, showing top-down view of subject drawing (during in-cage training). The metal tube is the reward spout. See videos of behavior in **Supplementary Videos 1-10**.

(e) Learned stroke primitives for each subject, one for each practiced shape. Stroke onsets are marked with a ball. Blue shading marks images for which subjects did not readily learn stereotyped stroke primitives, either because they used two strokes or they executed a single stroke variably across trials (see Methods). Drawings are averages over 15 trials.

(f) Example single-trial stroke trajectories for the "circle" shape, depicted in two ways, instantaneous speed as a heatmap on the stroke trajectory (top) or as velocity vs. time (bottom). Shown are four trials per subject.

(g) Summary of pairwise trajectory distances across trials, where pairs of trials are grouped along two dimensions: whether they refer to the same shape and to the same subject. Violin plots represent kernel density estimates, with overlaid medians and quartiles. Starting from a pool of 300 trials (15 shapes x 10 trials x 2 subjects), each data point represents a single trial's average distance to a specific shape, such that N = 300 for both "same shape, same subject" and "same shape, different subject" (15 shapes x 10 trials x 2 subjects), N = 4200 for "same subject, different shape" (15 shapes x 10 trials x 14 other shapes x 2 subjects). ***, $p < 0.0005$, two-sided Wilcoxon signed-rank test, performed on data first averaged over trials (N = 15 shapes, each averaged across subjects and trial-pairs).

Stroke primitives exhibit motor invariance over location and size.

If subjects represent stroke primitives as symbols, then stroke primitives should exhibit motor invariance. Each primitive's idiosyncratic trajectory should generalize across low-level motor parameters (e.g., muscle activity patterns), as seen in handwriting and other skills[70]. To test motor invariance, we presented each shape at a size and location that varied across trials, including novel location-size conditions. Motor invariance predicts similar stroke trajectories across location and size ("Symbols", **Fig. 2a**). Alternatively, if responses were memorized by rote, one for each specific stimulus, the subject would have difficulty generalizing, as seen in inflexible automatized skills[71] ("Fail" in **Fig. 2a**). A third alternative strategy could prioritize efficiency by minimizing movement from the starting position of the hand. This would predict different trajectories dependent on location and size ("Efficient" in **Fig. 2a**). We found that stroke trajectories were similar across locations and sizes (**Figs. 2b-e**), indicating that stroke primitives exhibit motor invariance.

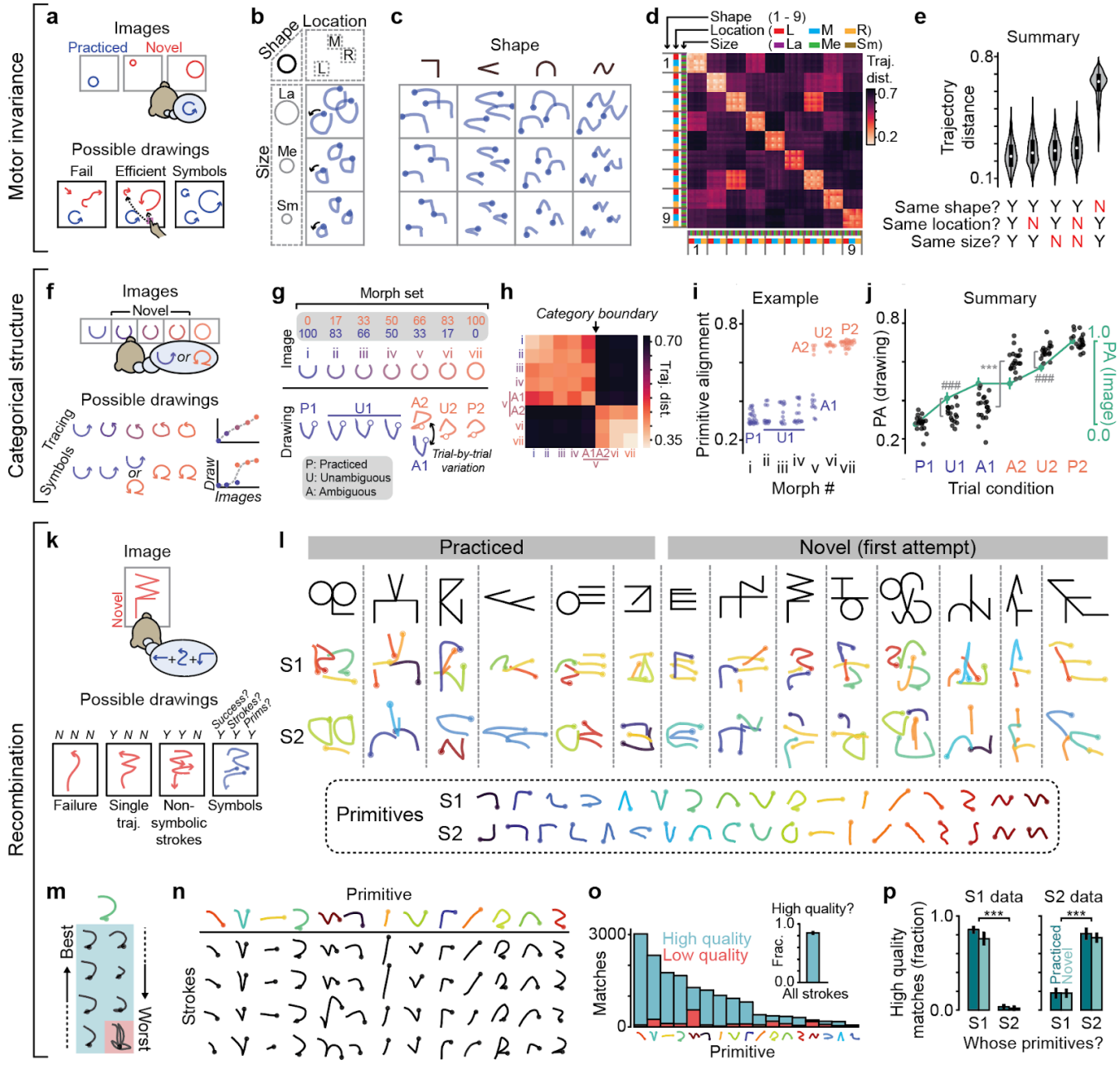


Fig. 2. Behavioral evidence for action symbols: motor invariance, categorical structure, and recombination into sequences.

(a) Experiment designed to test motor invariance over location and size. Each trial presents one shape, which varies in size and location across trials. Shown are three possibilities for how subjects will draw; only a symbolic representation predicts reuse of stroke primitives.

(b) Example drawings for a "circle" shape, across trials varying in location (left, middle, right) and size (small, medium, large). Three trials are overlaid on each panel (subject 2).

(c) Example drawings for more shapes (columns), varying in location and size.

(d) Heatmap of pairwise trajectory distances between each combination of shape, location, and size, plotting the mean distance across trial-pairs (N = 9 trials per shape-location-size condition).

(e) Summary of pairwise trajectory distances, where pairs of trials are grouped along three dimensions: same shape, same size, and same location. Violin plots represent kernel density estimates, with overlaid medians

and quartiles. Each underlying datapoint is one trial's average distance to one shape/size/location condition. N = 648 (same shape/size/location = YYY), 1296 (YYN), 1296 (YNY), 2592 (YNN), 5184 (NYY), taken from a pool of 729 trials (81 shape-location-size conditions x 9 trials each).

(f) Experiment designed to test for categorical structure. Given novel images generated by morphing between two well-practiced shapes, we consider two hypotheses, "Tracing" and "Symbols". A tracing strategy predicts that drawings should closely match each interpolated figure, and thus their properties will vary linearly across the morphed images. A symbolic representation predicts drawings with two hallmarks of categorical structure: sigmoidal non-linearity and trial-by-trial discrete variation.

(g) Example experiment for one morph set. Across trials, images varied linearly (randomly sampled) between two well-practiced images: morph *i* (mixture ratio of 100% "U", 0% circle) and morph *vii* (0% "U", 100% circle). Example drawings are shown for each morph. Two drawings are shown for morph *v*, because it elicited trial-by-trial variation between two primitives (A1, A2). We define "trial conditions" that group morphs based on whether they are well-practiced images (P), are unambiguously drawn as one primitive (U), or are ambiguously drawn as one primitive or the other across trials (A).

(h) Heatmap of pairwise trajectory distances between morphs for the example morph set in panel g. The category boundary is defined as the morph that was found to elicit trial-by-trial variation in drawing (i.e., "ambiguous" morph). Trials for morph *v* are split into two trial conditions based on the drawing (A1, A2). N = 5 - 30 trials (range across morphs, counting A1 and A2 separately).

(i) Primitive alignment vs. morph number for morph set in panel g. N = 5 - 30 trials (range across images, counting A1 and A2 separately). For primitive alignment using image data, instead of drawing data, see **Extended Fig. 2c**).

(j) Summary of primitive alignment (PA) vs. trial condition across all experiments, showing scores based on image data (green, mean and 95% CI) and drawing trajectories (black). Each drawing datapoint represents a single morph set [N = 20 morph set, combining subjects 1 (13) and 2 (7)]. The reason that primitive alignments using drawing data for P1 and P2 are not exactly 0 and 1 is due to trial-by-trial variation in behavior. ###, $p<0.0005$ testing for sigmoidal nonlinearity, i.e., that, compared to the image, primitive alignment for drawings is closer to 0 (U1) and 1 (U2). Drawing scores were first normalized so that the means of P1 and P2 trials were 0 and 1, respectively, and data were aggregated across U1 and U2; ***, $p<0.0005$, testing for trial-by-trial switching (drawing distribution in A2 higher than in A1); two-sided Wilcoxon signed-rank tests.

(k) Experiment designed to test for recombination of primitives into sequences, using character tasks. Four possible drawing responses are shown, each differing along three properties: success in copying the figure, use of multiple strokes, and reuse of a subject's own primitives. Combined evidence for all three properties is consistent with a symbolic representation.

(l) Example drawings in character tasks by subjects 1 and 2 given the same set of images. Strokes are color-coded by their assigned primitive from the subject's own primitive set (see Methods). "Primitives" plots each subject's "ground-truth" primitives taken from the single-shape task.

(m) Example character strokes that were assigned to one particular primitive ("reversed C"), uniformly ranging from best to worst match. Stroke background color indicates whether strokes are high (blue) or low (red) quality matches. Data from subject 1.

(n) Example character strokes organized by the primitive they were assigned to (columns), showing four example high-quality matches. Strokes are ordered from high (top row) to median (bottom row) quality (based on trajectory distance to primitive), most to least frequent (columns), showing only the thirteen most frequent (see frequency distribution in panel o). Data from subject 1.

(o) Frequency histogram of strokes matched to each primitive. Stacked bars separate high- and low-quality matches. Inset, fraction of all strokes that received a high-quality score and were thus considered cases of primitive reuse. Data from subject 1.

(p) Summary of primitive recombination, showing fraction of strokes labeled as high-quality match to a primitive and were thus considered cases of primitive reuse, including only characters performed by both subjects. Results show each subject's strokes (data) tested against each subject's primitives. Bars show mean and 95% CI. N (characters) = 133 (subject 1, practiced), 90 (S1, novel), 70 (S2, P), 135 (S2, N); note that a character could contribute a data point to both practiced and novel groups, thus explaining the different sample sizes between subjects. ***, $p < 0.0005$, two-sided Wilcoxon signed-rank test, performed independently for practiced and novel data.

Stroke primitives exhibit categorical structure.

If subjects represent stroke primitives as categorically structured action symbols, we would expect them to preferentially draw using their idiosyncratic set of learned primitives when challenged with new figures that interpolate, or "morph", between the shapes the primitives were associated with during learning (**Fig. 2f,** "Symbols"). This would indicate that stroke primitives are represented as a set of discrete types of action, similar to stroke categories in Chinese characters[61] or phonemes in speech[20]. In contrast, if subjects simply trace images without interpreting them as action symbols, then we would expect drawings to closely match the interpolated figures (**Fig. 2f**, "Tracing"). We presented subjects with images that were randomly sampled on each trial by linearly morphing between two practiced shapes (each "morph set" consisted of two practiced shapes and four to five morphed figures). We tested whether the resulting drawings reflected a categorical boundary in the subject's interpretation of these images, which would manifest in two behavioral hallmarks (**Fig. 2f,** "Symbols"): a steep sigmoidal relationship between image variation and drawing variation[72,73]—such that images on the same side of the category boundary are drawn similarly, while images across are drawn differently—and trial-by-trial variation between distinct stroke primitives for images close to the boundary.

We found evidence for both of these hallmarks of categorical structure. In the example shown in **Fig. 2g** (more examples in **Extended Data Fig. 2a,b**), the images varied in the extent to which the top was closed, ranging linearly between two practiced shapes, a "U" (morph *i*) and a circle (morph *vii*). We defined a category boundary as the morph that was found to elicit trial-by-trial variation between primitives 1 and 2 (morph *v*). We found that drawings for morphs to the left of that boundary (U1) were drawn similarly to primitive 1 (P1), and drawings for morphs to the right (U2) were similar to primitive 2

(P2, **Fig. 2g**). This discrete, as opposed to smoothly varying, distinction across the category boundary was evident as two blocks in a heatmap of the pairwise trajectory distances between all morphs (**Fig. 2h**). This block structure implies high similarity within each block, and low similarity across blocks. To quantify this effect, we calculated a "primitive alignment" score for each trial. Primitive alignment quantifies the relative similarity of a given trial's drawing to primitive 1 (alignment = 0) and primitive 2 (alignment = 1), and is defined as $d_1/(d_1 + d_2)$, where d_1 (or d_2) is that trial's average trajectory distance to primitive 1 (or 2) trials. First, we confirmed that applying this metric to score the images produced a linear relationship with morph number (**Extended Data Fig. 2c**). In contrast, applied to behavior, primitive alignment varied nonlinearly in a sigmoidal manner across morph number (**Fig. 2i**). This analysis also captured the trial-by-trial variation in behavior for morphs at the category boundary (morph *v*, **Fig. 2i**). These effects—nonlinearity and trial-by-trial discrete variation—were consistent across morph sets (**Fig. 2j**), indicating that subjects internally represent primitives as categorical stroke types.

Stroke primitives are recombined into sequences.

If subjects represent stroke primitives as symbols, they should readily recombine primitives to construct multi-stroke drawing sequences. We tested this using two tasks that challenge subjects with complex figures, including novel ones. The first "multi-shape" task presented figures that combined multiple (two to four) disconnected shapes, which could be drawn in any order. We reasoned that one possible strategy was to use a single trajectory that efficiently traces over all shapes with appropriately timed touches and raises to produce strokes and gaps ("Single trajectory", **Extended Data Fig. 3a**). With this strategy, behavior would be biased to minimize the movement, or gaps, between shapes, ignoring whether this leads to reuse of primitives. A second strategy, consistent with a symbolic representation, would be to draw by recombining the learned stroke primitives, at the expense of longer movements during gaps between strokes ("Symbols", **Extended Data Fig. 3a**). We analyzed all pairs of consecutive strokes for which these two strategies predicted different stroke trajectories (**Extended Data Fig. 3b,** top schematic). In virtually all cases, subjects preferred to recombine primitives, at the

expense of longer movements in gaps, consistent with a symbolic representation (**Extended Data Fig. 3b**, bottom).

In the second “characters” task, we presented complex figures called characters, each designed by connecting multiple simple shapes (inspired by logographic writing systems), such that there was ambiguity regarding their components. We considered four possible drawing outcomes (**Fig. 2k**): subjects may fail to accurately draw these novel figures (“Failure”), they may succeed using a single unsegmented trajectory (“Single trajectory”), they may use multiple strokes that are not in their set of learned primitives (“Non-symbolic strokes”), or they may successfully draw by reusing their own primitives (“Symbols”).

We found that drawings were consistent with a symbolic representation. Subjects successfully drew novel characters, and did so using multiple strokes (thus contradicting “Failure” and “Single trajectory” predictions). Critically, stroke trajectories matched each subject’s own primitives, rather than the other subject’s primitives, leading to idiosyncratic differences in the way the same figures were drawn by the two subjects (thus contradicting “Non-symbolic strokes” predictions; **Fig. 2l**, videos in **Supplementary Videos 1-10**). To quantify primitive recombination, we collected all strokes from each subject’s character drawings and quantified how often these strokes matched closely at least one of the subject’s own primitives. To do so, we scored each stroke’s trajectory distance to each of the subject’s own primitives (“ground-truth” strokes from single-shape tasks) and assigned each stroke to its best-matching primitive. We then classified that match as either high- or low-quality, defined so that high-quality means that a stroke falls within “natural” trial-by-trial variation in drawings of its assigned primitive, thus counting as primitive reuse (see Methods, and example resulting matches in **Fig. 2m, n**). We found that a large majority of strokes were classified as high-quality matches to a primitive (>80%, **Fig. 2o**). We further restricted our analysis to just the characters that both subjects had drawn, and found that a given subject’s strokes matched its own primitives, and not those of the other subject (**Fig.**

2p). This finding—that even on identical figures subjects selectively reused their own primitives—indicates that drawings build on prior knowledge of primitives.

Multi-area neuronal recordings across frontal cortex.

Together, the findings that stroke primitives exhibit motor invariance (**Fig. 2a-e**), categorical structure (**Fig. 2f-j**) and recombination (**Fig. 2k-p**) indicate they have an underlying symbolic representation. We next searched for neural correlates of these putative stroke symbols. In order to search broadly, we recorded simultaneously from neurons across multiple areas of frontal cortex, using chronically implanted multi-electrode arrays (sixteen 32-channel arrays per subject, **Fig. 3a, b**, and **Extended Data Fig. 4**). We targeted regions that have been associated with a wide range of motor, planning, and other cognitive functions that may be relevant to this task (see Methods), including (but not limited to) motor control in primary motor (M1), dorsal premotor (PMd), and ventral premotor cortex (PMv)[74]; motor planning and decision-making in PMd and ventral premotor cortex (PMv)[75]; motor abstraction in PMv[44]; abstract reasoning and planning in dorsolateral and ventrolateral prefrontal cortex[76,77] (dlPFC and vlPFC) and frontopolar cortex[78] (FP); and action sequencing in the supplementary motor areas[46] (SMA and preSMA). We recorded 48.4 +/- 19.9 units per area (mean and S.D. across sessions) for subject 1 and 48.0 +/- 16.0 for subject 2 (see details in Methods).

We found clear task-related activity in all recorded areas except FP. This is evident as non-zero baseline-subtracted trial-aligned activity in **Fig. 3c**. A coarse inspection revealed that different areas exhibited grossly different activity patterns in relation to the ongoing trial. For example, many units in prefrontal areas vlPFC and dlPFC and premotor areas PMd and PMv had rapid responses locked to the image onset (**Fig. 3c**, at around 0.1 s post-onset) and varied activity patterns during the planning epoch. In contrast, motor area M1 had relatively weak visual and planning-related activity, but strong movement-related activity during the stroke epoch (**Fig. 3c**). Paralleling the behavioral experiments, we tested whether activity in these areas encodes stroke primitives in a manner that exhibits motor

invariance, categorical structure, and recombination. We found that a single area, PMv, exhibited all of these properties.

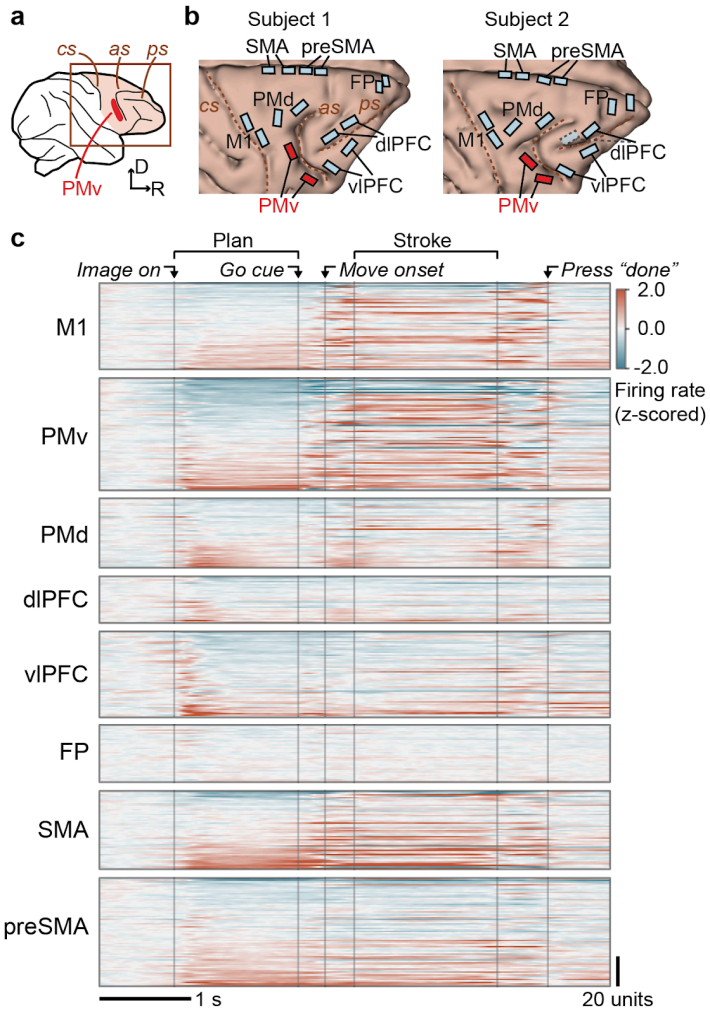


Fig. 3. Multi-area neuronal recordings across frontal cortex.

(a) Recordings were targeted to multiple areas of right frontal cortex, contralateral to the hand used for drawing. Arrays were targeted stereotactically and registered to sulcal landmarks, including central sulcus (cs), arcuate sulcus (as), and principal sulcus (ps). R, rostral. D, dorsal.

(b) Final locations of arrays, depicted to scale on 3D rendering of cortical surface. Note that SMA and preSMA electrodes targeted the medial wall of the hemisphere (**Extended Data Fig. 4**). The caudal dlPFC array for subject 2 was implanted but malfunctioned due to cable damage.

(c) Average activity across trials, grouped by brain area (panels) and split by unit (inner rows). Before averaging, trials were aligned by linear time-warping to a median trial, aligned at "anchor" events that consistently occurred across trials (see Methods). Both single- and multi-units are included. Each unt's activity was z-scored relative to the baseline time window preceding image onset, and sorted within each area from low to high activity in the planning epoch. Sorting was done in a cross-validated manner, by determining sort indices using one subset of data (N = 50 trials) and applying those indices to the remaining subset that is plotted (N = 235). This plot includes from a single session of single-shape trials (subject 2).

PMv encodes stroke primitives in a manner that is invariant over location and size.

In single-shape tasks, where subjects drew primitives that varied in location across trials (**Fig. 2a-e**), we analyzed neural activity during the planning epoch (between image onset and the go cue). PMv activity varied strongly depending on the planned primitive, with relatively small influence of location. For example, the unit shown in **Fig. 4a, b** fired strongest for one specific primitive (red), second strongest for another primitive (orange), and so on, and this pattern was similar across the four locations. To quantify how strongly population activity encodes primitives, we first identified a primitive-encoding

population subspace using targeted principal components analysis (i.e., a linear projection of population activity that maximizes activity variation related to primitives; see Methods). PMv activity in this subspace varied strongly depending on the primitive, and minimally depending on location (**Fig. 4c,d**). For example, in **Fig. 4d**, within each location (subpanel), primitive trajectories (each color) differed from each other after image onset ("x" icon), reflecting strong encoding of primitives; but across locations, trajectories were similar, reflecting minimal encoding of location. In contrast, in a different area known for its role in cognitive control, dlPFC, activity recorded in the same trials, and analyzed the same way, strongly reflected location, with minimal encoding of primitive (**Fig. 4e-h**). Illustrating this point, in **Fig. 4h** primitive trajectories were similar to each other within location, but different across locations.

To quantify the extent to which population activity in each area encodes primitives vs. locations, we devised a "neural distance" metric, which quantifies the activity dissimilarity between any two sets of trials representing two conditions (e.g., different location-primitive conditions). Neural distance is the average pairwise Euclidean distance between all across-condition trials, normalized by average within-condition pairwise Euclidean distance, such that neural distance ranges between 0 and 1 (a ceiling, defined as the 98th percentile of pairwise Euclidean distances). We computed neural distance between each pair of primitive-location conditions. The resulting pairwise neural distances in PMv were consistently low for the same primitive across locations, visible as off-diagonal streaks in a heatmap of pairwise distances (**Fig. 4i**). In contrast, the heatmap for dlPFC had four blocks along the diagonal, indicating similar activity within each location, regardless of primitive (**Fig. 4i**). Summarizing these effects, we found that, in PMv, the neural distances between primitives controlling for location ("primitive encoding") were consistently higher than the distances between locations controlling for primitives ("location encoding") (**Fig. 4j,** left column). In dlPFC the opposite was true: primitive encoding was low and location encoding was high (**Fig. 4j,** left column). A comparison of all areas (**Fig. 4j,** right column) showed that PMv was the only one to exhibit this combination of strong primitive encoding and weak location encoding.

This strong encoding of primitives and weak encoding of location suggests that PMv population activity uses the same population code for primitives across locations. We tested this by measuring the performance of a linear decoder, trained to decode primitives at one location, in generalizing to held-out locations. We found that cross-location decoder generalization was strong in PMv and relatively weak in all other areas (**Fig. 4k**).

In different sessions, using the same methods, we found that PMv population activity encodes primitives in a manner that is also invariant to size (**Extended Data Fig. 5**). Together, these findings indicate that PMv represents primitives in a manner invariant to location and size.

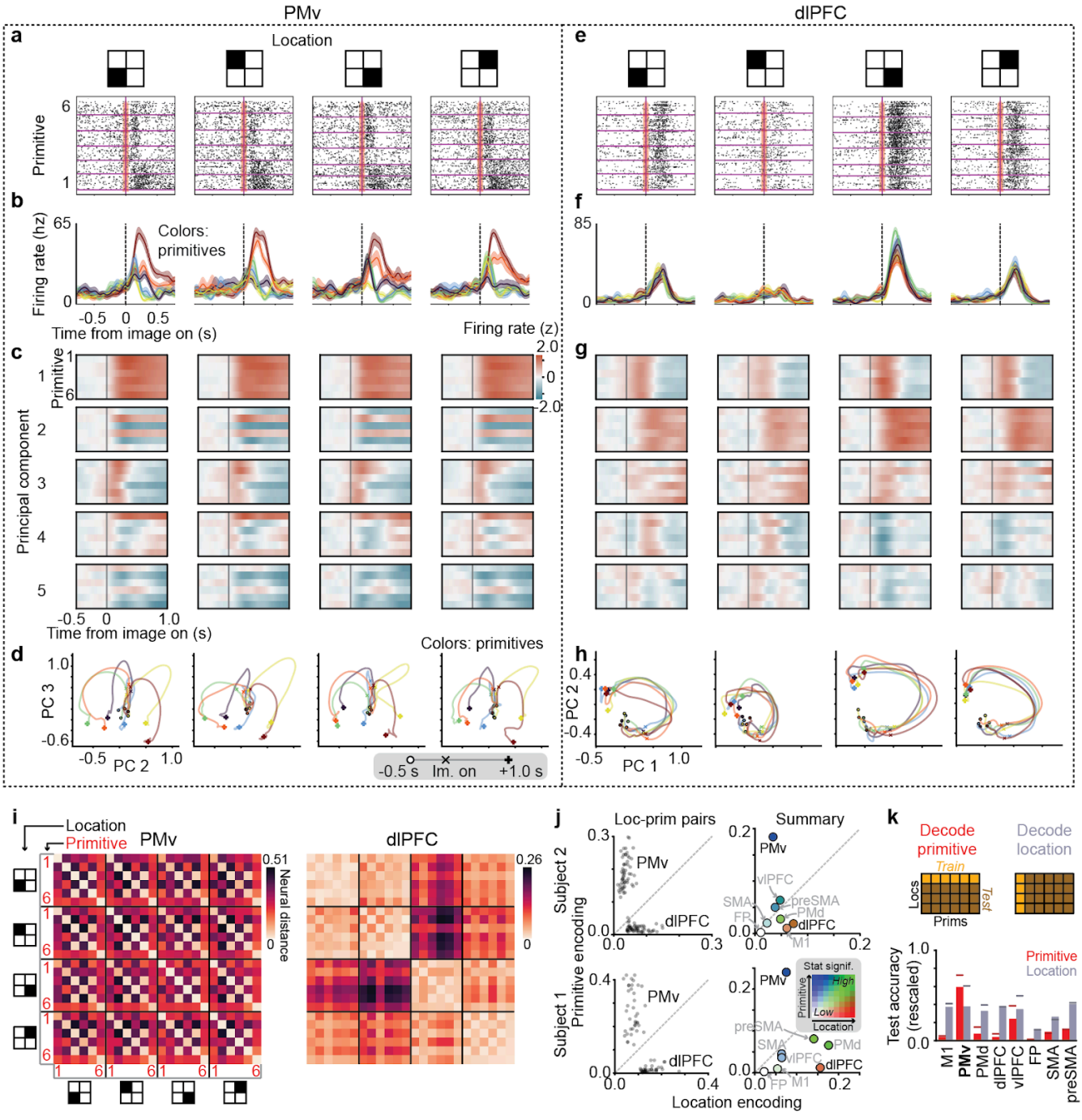


Fig. 4. PMv encodes stroke primitives in a manner that is invariant over location (for invariance over size see Extended Data Fig. 5).

(a) Raster plot for an example PMv unit, showing single-trial spike patterns aligned to image onset. Trials (rows) are grouped by primitive (separated by purple lines) and split by spatial location of the drawing (columns). The go cue occurred 1.1 - 1.5 seconds after image onset; therefore, the subject's finger was held still, pressing the fixation button, throughout the plotted time window.

(b) Smoothed firing rates for the example PMv unit, averaged across trials for each primitive (colors), split by location (columns). Curves show mean and SE (N = 13 - 20 trials per primitive-location combination).

(c) PMv population activity in primitive-encoding principal components (outer rows), showing average activity for each primitive (inner rows), and split by location (columns). Here, activity was z-scored and baseline subtracted relative to activity before image onset.

(d) Average time course of PMv population activity (i.e., state space trajectories) for each primitive (colors), plotted in a subspace spanned by principal components 2 and 3. Data are all from before the "go cue" and therefore before movement onset. No z-scoring or baseline subtraction was performed. Gray legend depicts trajectory, with image onset marked with an "x" icon.

(e) Analogous to panel a, but for an example dlPFC unit recorded simultaneously as the PMv unit.

(f) Analogous to panel f, but for the example dlPFC unit.

(g) Analogous to panel c, but for dlPFC population activity in the same session.

(h) Analogous to panel d, but for dlPFC.

(i) Heatmap of pairwise neural distances between unique primitive-location conditions, for the same session as panels a-h, averaged over time (0.05 to 0.6 s after image onset). N = 13 - 20 trials per primitive-location combination.

(j) Summary of primitive encoding and location encoding across areas and sessions. In the left two panels, each point represents a unique primitive-location condition. Each condition's primitive encoding is its average neural distance to all other conditions that have the same location but different primitive. Location encoding is defined analogously. Only PMv and dlPFC are plotted, to reduce clutter. N = 37 (subject 1, aggregating across 2 sessions), 59 (subject 2, 3 sessions). In the right two panels, each point depicts the mean encoding scores. The color denotes statistical significance. It indicates the number of other brain areas that this area beats in pairwise statistical tests of primitive encoding and location encoding (represented in the inset heatmap). For example, PMv is deep blue because it has higher primitive encoding than every other area, but not higher location encoding. See Methods. For statistics, each data point was a unique pair of primitive-location conditions (trial-averaged). For testing primitive encoding, these pairs were the same location, but different primitive [N = 93 (subject 1), 288 (subject 2)]. For location encoding, the pairs were the same primitive, but different location [N = 114 (S1), 132 (S2)]. Statistical tests were performed on condition pairs pooled across sessions (2 for S1, 3 for S2). See Methods for details.

(k) Across-condition generalization of linear support vector machine decoders for primitive (red) and location (gray). Top: schematic of the train-test split method; the primitive decoder was trained on data from one location and tested on data from all other locations (and analogously for location decoding). Bottom: accuracy is linearly rescaled between 0 (chance) and 1 (100%). The horizontal line represents within-condition decoding accuracy (with train-test splits partitioning data at the level of trials).

PMv activity reflects categorical structure of stroke primitives.

In experiments finding behavioral evidence for categorical structure (**Figs. 2f-j**, **5a**)—sigmoidal variation in behavior even with linear variation in images—we found that PMv activity diverges towards two separate primitive-representing states, in a manner that aligns with the planned primitive on each trial (**Fig. 5**). This is illustrated in a state space plot of PMv population trajectories during the planning epoch (**Fig. 5b**). After image onset ("x" icon), trials with behavior resembling primitive 1 (morphs *i-iv*) separated from trials assigned to primitive 2 (morphs *vi, vii*), such that activity diverged to two states representing primitives 1 and 2 (**Fig. 5b**). Similarly, for the image at the category boundary (morph *v*),

the trajectories diverged towards these two states depending on whether the subject was planning to draw primitive 1 or 2 (A1 or A2, in **Fig. 5c**).

To quantify this neural separation into two primitive-representing states, we first computed the Euclidean distance between each pair of trials, and then scored each trial with its primitive alignment score, representing its relative distance to activity encoding primitives 1 and 2 (P1 and P2). This revealed the same two hallmarks of categorical structure that we saw in behavior: sigmoidal nonlinearity and trial-by-trial switching. Sigmoidal nonlinearity is evident in the heatmap of pairwise neural distances as two main blocks separated by the category boundary (**Fig. 5d**), and in the plot of primitive alignment versus morph number (**Fig. 5e, f**). Trial-by-trial switching between primitive-representing states for ambiguous images was evident in the heatmap of pairwise neural distances (**Fig. 5d**, A1 is closer to morphs *i-iv*, while A2 is closer to morphs *vi-vii*), and in the trial-by-trial variation in primitive alignment scores (**Fig. 5e, f**). Comparing across areas, this effect was strongest in PMv (**Fig. 5g**).

We considered two explanations for the trial-by-trial switching in PMv's activity for the ambiguous images. One explanation is that activity is biased towards whichever primitive-representing state it happens to be closer to at baseline, before image onset. Arguing against this, we found that the relative proximity of baseline activity to states encoding primitive 1 or primitive 2 did not predict the chosen primitive (in **Fig. 5h**, before image onset, A1 and A2 activity is not different). Another possible explanation is competition between states after the image turns on, with the winning state determining what primitive will be drawn. Network modeling suggests that such competition would manifest as slow temporal evolution of activity towards one state or the other[79]. Consistent with this, the separation of activity after image onset was slower for ambiguous images ("A2 - A1" in **Fig. 5h**) than for images unambiguously drawn as primitive 1 or primitive 2 ("P2 - P1" and "U2 - U1" in **Fig. 5h**). In behavior as well, reaction time was slower for ambiguous images (**Extended Data Fig. 6**). This slower temporal

dynamic in neural activity is suggestive of a winner-take-all decision process occurring after image onset.

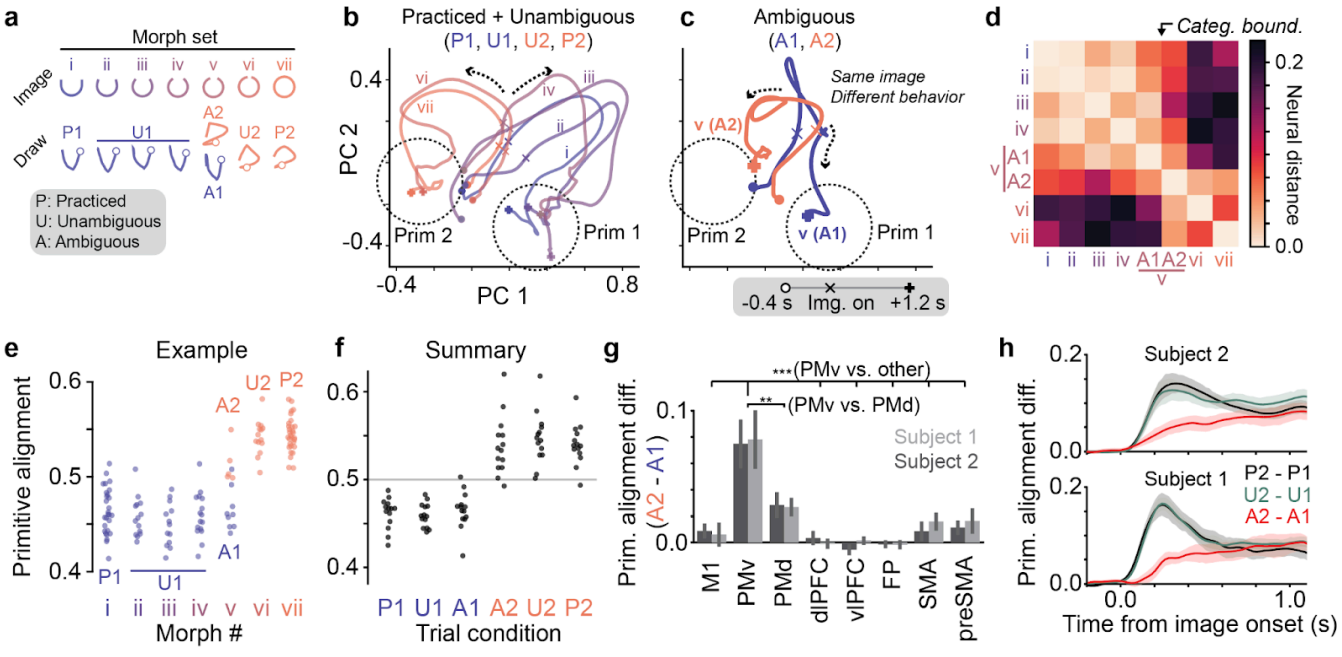


Fig. 5. PMv activity reflects categorical structure of stroke primitives.

(a) Example experiment showing discrete structure in drawing given continuously varying images. See also **Fig. 2f-i**.

(b) PMv mean population trajectories, showing divergence of activity after image onset (x) towards states encoding either primitive 1 or 2, depending on whether the morph was to the left (*i-iv*) or right (*vi-vii*) of the category boundary. Data are all from before the "go cue", which occurred 1.2 - 1.6 seconds after image onset; therefore, the subject's finger was held still (pressing the fixation button) throughout the plotted time window. Principal components were identified using PCA (see Methods). Trajectory legend is under panel c. N = 14 - 30 trials per trajectory.

(c) PMv population trajectories for morph *v*, split by whether the subject will draw primitive 1 (A1) or primitive 2 (A2), overlaid on the same subspace from panel b. N = 9 (A1) and 5 (A2) trials.

(d) Heatmap of neural distance between PMv activity for each pair of morphs. The morph on the category boundary (morph *v*) is split into two groups based on the drawing (A1 or A2). N = 5 - 30 trials (range across morph, considering A1 and A2 separately).

(e) Primitive alignment vs. morph number for the example experiment. Primitive alignment quantifies the relative similarity of each trial's neural activity to the neural activity for primitive 1 and for primitive 2, and is defined as $d_1/(d_1 + d_2)$, where d_1 and d_2 represent that trial's average Euclidean distance to primitive 1 and primitive 2. Each datapoint represents a single trial (N = 5 - 30 trials).

(f) Summary of primitive alignment vs. trial condition across morph sets. Each morph set contributes one point to each condition [N = 14 morph sets (5 from S1, 9 from S2), including only morph sets that have all six trial conditions (14/20)].

(g) Summary of trial-by-trial variation in primitive alignment for ambiguous images (A2 - A1), showing mean +/- SE. **, p<0.005; ***, p<0.0005, paired t-test [N = 20 morph sets, combining subjects 1 (7) and 2 (13)].

(h) Average time course of difference in primitive alignment, split by trial condition, and aligned to image onset (mean +/- SE). [N = 14 morph sets (5 from S1, 9 from S2), including only morph sets that have all six trial conditions (14/20)].

PMv activity reflects recombination of primitives into sequences.

When subjects recombine primitives into a multi-stroke sequence to draw characters (**Fig. 2k-p**), does this reflect a reused primitive-representing state in PMv? For each primitive, we compared PMv activity between when it was used in single-shape vs. character tasks (both from the same session; see Methods). We focused on activity in a time window immediately preceding stroke onset, because, for character tasks, the alternative approach of using the entire planning window leaves uncertain which primitive we would expect to be represented in PMv. For characters, we focused on the first stroke to ensure a fair comparison between the two tasks. This maintained similar pre-stroke gross arm movements between single-shape and character tasks (from the start button towards drawing), ensuring that any differences in activity are unlikely to be due to differences in gross movements. This also helped ensure that activity would not be influenced by the immediately preceding stroke[80].

We found that PMv encodes primitives in a similar manner across single-shape and character tasks. This is evident in the population activity plotted in **Fig. 6a, b**, where each primitive's activity (inner rows in **Fig. 6a**, colored trajectories in **Fig. 6b**) is similar across tasks (columns in **Fig. 6a, b**). We contrasted PMv with preSMA, an area known to encode sequence-related information[46], finding that preSMA activity for the same primitive differed between the two tasks (**Fig. 6c, d**). A quantification of the neural distance between each pair of primitive-task condition (**Fig. 6e**, showing PMv and preSMA) confirmed that PMv encodes primitives similarly across tasks. This is visible in the off-diagonal low-distance streak for PMv in the heatmap of pairwise distances (**Fig. 6e**), and in the summary plot showing that in PMv the neural distance between primitives controlling for task ("primitive encoding") was high while the neural distance between task controlling for primitive ("task encoding") was low (**Fig. 6f**). In contrast, preSMA activity was different when the same primitive was used in single-shape or character tasks. This is evident in the block structure in **Fig. 6e**, and in preSMA having lower primitive encoding and

higher task encoding compared to PMv (**Fig. 6f**). Across all areas, PMv activity most consistently had high primitive encoding and low task encoding (**Fig. 6f**). This finding indicates that representations of primitives in PMv are reused across recombined sequences.

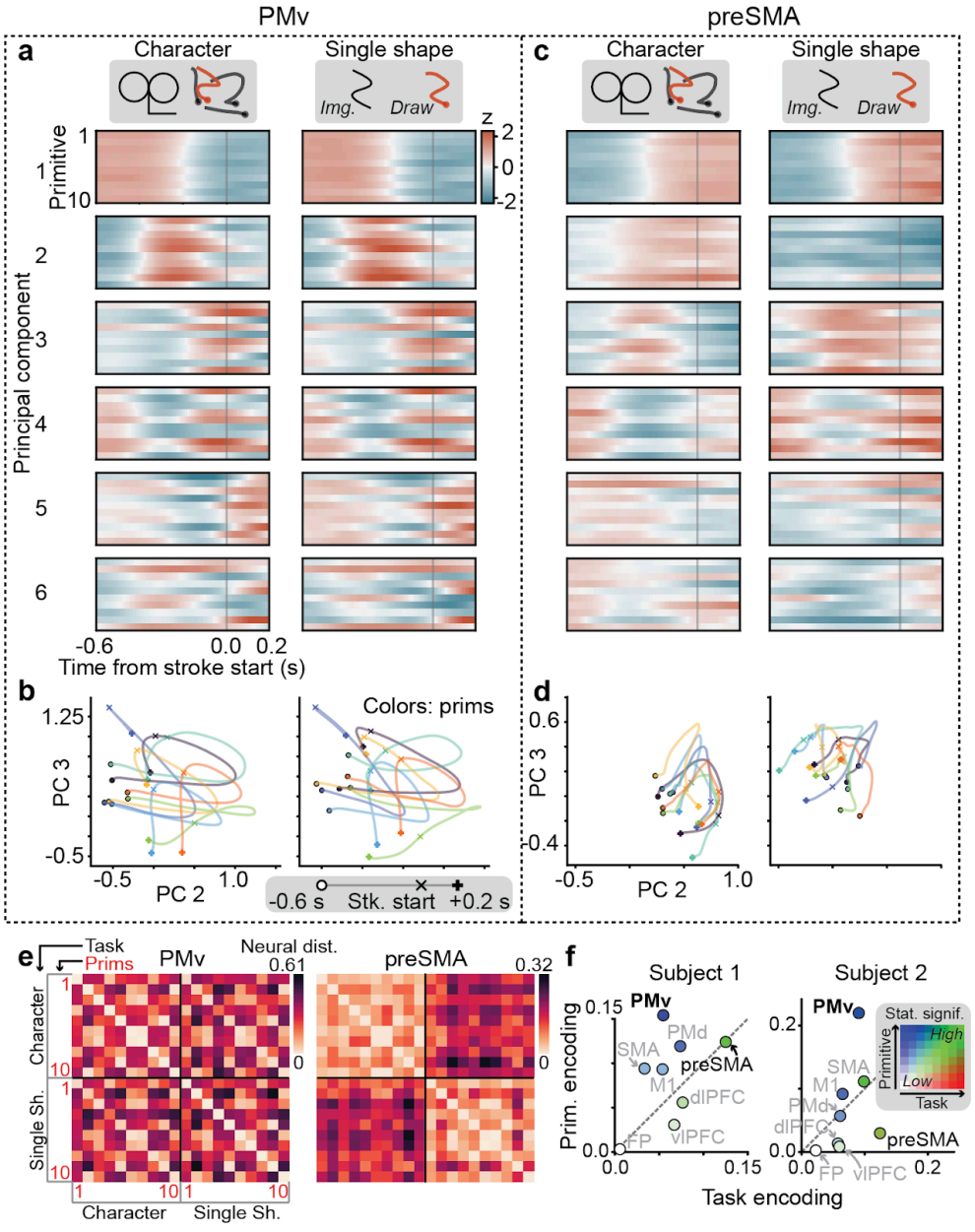


Fig. 6. PMv activity reflects recombination of primitives into sequences.

(a) PMv population activity, projected to its first six principal components using PCA (outer rows), showing average activity for each primitive (inner rows), and split by whether that primitive was performed in character or single-shape tasks (columns). Here, activity was aligned to stroke onset and z-scored over the entire depicted window. Data from a single session for subject 2. N = 4 - 16 (single-shape), 6 - 72 (character) trials.

(b) PMv population trajectories for each primitive (color), plotted in a subspace spanned by principal components 2 and 3, and aligned to stroke onset. Gray legend depicts a trajectory icon, with stroke onset marked with an "x" icon. This plot shows only a subset of primitives (7/10) to reduce clutter.

(c) Analogous to panel a, but for simultaneously recorded activity in preSMA.

(d) Analogous to panel b, but for simultaneously recorded activity in preSMA.

(e) Heatmap of pairwise neural distances between each unique combination of primitive and task kind, for the experiment in panels a-d. N = 4 - 16 (single-shape), 6 - 72 (character) trials.

(f) Summary of primitive encoding and task encoding across areas and sessions. Points show mean encoding scores. The color of each point denotes statistical significance (analogous to **Fig. 4j**). It indicates the number of other brain areas that this area beats in pairwise statistical tests of primitive encoding and task encoding (represented in the inset heatmap). Each data point was a unique pair of primitive-task conditions (trial-averaged). For testing primitive encoding, these pairs were the same task type, but different primitive [N = 350 (subject 1), 380 (subject 2)]. For task encoding, the pairs were the same primitive, but different task type [N = 29 (S1), 48 (S2)]. Statistical tests were performed on condition pairs pooled across sessions (10 for S1, 9 for S2).

Discussion

Identification of a neural substrate of symbols

We identified a localized neural population encoding action symbols. We tested for three essential properties of symbols—invariance, categorical structure, and recombination—in both behavior and recordings. In behavior, we found that monkeys successfully trace complex geometric figures by recombining discrete strokes, which exhibit motor invariance (**Fig. 2a-e**) and categorical structure (**Fig. 2f-j**), into new composed sequences (**Fig. 2k-p**). These properties indicate that strokes reflect internal representations of action symbols. In recordings across motor, premotor, and prefrontal cortex (**Fig. 3**), we found that population neural activity in PMv encodes strokes in a manner that reflects motor invariance (**Fig. 4),** categorical structure (**Fig. 5**), and recombination (**Fig. 6**). PMv activity also shows visual invariance. In two experiments that dissociated image from action, PMv encoded the planned action instead of the image. These were experiments where (i) the same ambiguous image was drawn across trials as one action symbol or another (**Fig. 5f**, “A2” differs from “A1”), and (ii) when character and single-shape tasks presented different images that elicited the same action (**Fig. 6f**, PMv has high primitive encoding and low task encoding). This finding of a localized representation of action symbols in PMv indicates an important role for symbols in producing novel, goal-directed action sequences. More broadly, this finding supports the hypothesis that cognition can be understood, at the algorithmic level, in terms of internal symbolic operations, a hypothesis that so far has been based largely on behavioral and modeling evidence[2–7,9–11,14].

The identification of a symbolic neural representation in monkeys may provide more general insights into the neural basis and evolutionary origins of symbols. Certain abstract properties—notably reversible symbol reference[81], and the ability to recombine using higher-order relations[9,82] and recursive syntax[7,9,83]—are thought to characterize the kinds of symbolic representations used in language, mathematics, and formal reasoning in humans[7,9,81–85]. These abstract properties have not been readily apparent in animal behavior[9,81,82,84,85], suggesting they may be specific to humans. At the same time, animals demonstrate a variety of behaviors that appear consistent with internal symbolic

representations (see Introduction)[11,23–27,29–31,31], raising the possibility that symbols may contribute across cognitive faculties and animal species. Different kinds of symbols may have different properties appropriate for the abilities they contribute to, but they would share the core property of being discrete units that are internally recombined to support generalization[11]. We discovered that at least one kind of symbolic neural representation (action symbols) exists in macaque monkeys. This discovery supports the view that symbolic operations may in some form be common across species and cognitive domains. It remains to be determined whether and how the neural substrates identified here relate to other proposed kinds of symbols, including those with abstract properties hypothesized to be unique to humans[7,9,81–85]. Nonetheless, our finding of a representation of action symbols raises the possibility that a core set of neural mechanisms for generating discrete, invariant representations and recombining them may exist across primates and possibly other taxa.

Abstraction using action symbols in ventral premotor cortex

There is strong evidence for invariant representations in neuronal activity, especially in PFC[28,42,43,47,48,50–52] and medial temporal lobe[41,43,49,54,56,57]; in contrast, we found encoding of action symbols in ventral premotor cortex. This finding is surprising in view of those prior studies, suggesting a reassessment of the brain regions supporting abstraction. In particular, it suggests that PMv is also critical for abstraction: in particular, those kinds of abstractions related to motor behavior. This may be related to the fact that PMv is directly connected with both higher-order frontal areas (vlPFC and preSMA)[86] and motor areas (M1 and spinal cord)[86,87], positioning PMv at the nexus of cognitive and motor circuits. Consistent with a critical role for premotor areas in motor abstraction are findings from human functional imaging and lesion studies linking premotor cortex to the perception, imitation, planning, and imagination of action, including in handwriting[88–90], tool use[91,92], dance[93], and other domains[94,95]. A possibility is that these, and related findings, may have a unified explanation as internal operations on action symbols.

The finding of action symbol representations in PMv is consistent with prior evidence for invariant action representations in PMv, especially in the context of goal-directed hand movements, such as grasping and object manipulation[44]. Recordings have revealed a diversity of motor-invariant firing patterns, including encoding grasp properties independent of the arm being used[96], kinematics independent of muscle dynamics[97], and tool use regardless of hand posture[98]. Perhaps most strikingly, PMv contains "mirror neurons", which fire similarly whether one observes or performs a given action[60]. These abstract firing properties in PMv have been proposed to support a variety of functions, including object-oriented visuo-motor transformation[44,99,100], understanding of others' actions[60] (but see ref[101]), imitation[60], and mental imagery[69]. Perhaps most related to action symbols, PMv has been proposed to encode a "vocabulary" of discrete actions[44,60,99]. All of these proposals differ importantly from action symbols, both conceptually—action symbols place stronger emphasis on categorical structure and recombination—and experimentally—we tested both categorical structure (in a systematic manner not previously done) and recombination into novel sequences (differing from previous studies, which behaviorally tested only a few well-practiced sequences[102,103]). We also recorded more areas than in prior studies, which was essential for revealing specialization in PMv. This action symbol framework may thus provide a unifying understanding of PMv's functions.

The precise neural circuit mechanisms for encoding symbols remain to be determined, but our finding of a neural substrate encoding action symbols provides a starting point to address this question. Importantly, evidence already suggests that the circuit mechanisms underlying action symbol representation in PMv depend on local processes within PMv, as opposed to being driven entirely by image properties in a feedforward manner, or by symbol representations inherited from another region. First, as noted above, we found that PMv was not driven directly by image properties, but instead encoded the planned action. Second, we recorded from two of the higher-order areas, vlPFC and preSMA, that provide input to PMv[86], and did not find activity strongly representing symbols. PMv also receives inputs from the anterior intraparietal area (AIP)[104], but there is evidence that these inputs more strongly reflect visual than action parameters[100]. Together, these lines of evidence suggest that action

symbol representations depend on computations within PMv. These computations may reflect winner-take-all dynamics between discrete attractor states encoding action symbols[79], which may explain the slow activity dynamics for ambiguous images (**Fig. 5h**), and which is consistent with a proposed role for PMv in decision-making[105]. Future studies investigating neural processes in PMv may give insight into circuit mechanisms for generating symbolic representations.

Action sequences as composition of symbols

We introduced a task paradigm testing compositional generalization by recombining action symbols into novel sequences. This task has three critical design features: (i) a large set of possible problems that share common structure [i.e., shape components variably transformed (**Fig. 2a-j**) and connected in composites (**Fig. 2k-p**)]; (ii) a lack of direct instructive cues or strong constraints on how to draw, forcing subjects to decide on their own; and (iii) initial training to impart structured, generalizable prior knowledge (action symbols). Compositional generalization has not, to our knowledge, been the subject of prior studies of neural substrates of motor behavior, which have used tasks that fall largely into the following (not mutually exclusive) classes: automatic, instructed, working-memory-guided, or minimally-restrained (expanding the classification scheme from Mizes et al.[106]). Automatic movements are highly-practiced sequences (typically thousands of practice trials or more) executed with high stereotypy, often as a single “chunk” without the need for moment-by-moment instruction[71,106–117]. Instructed movements are guided by external cues that indicate specific actions, either directly (e.g., the location of a visual cue instructing where to reach)[80,106,118–121] or indirectly from learned rules associated with each cue (e.g., the rule that a red square means “reach left”)[29,47,122]. Working-memory-guided movements are produced from short-term memory, soon (seconds to minutes) after the movement was instructed or learned[28,106,107,121,123]. Minimally-restrained movements are spontaneously produced in a naturalistic setting that, in contrast to common task-based paradigms, is less constrained both physically and in terms of specifically defined problems[124–127]. Previous drawing-like tasks in monkeys can be classified as either automatic[128], instructed[129,130], or unconstrained[130]. Our study therefore complements existing task paradigms by establishing a new

approach to probe the neural mechanisms of action symbols in motor behavior. These mechanisms may turn out to also underlie other forms of motor abstraction described elsewhere, including plans[131], hierarchies[132,133], and categories[134].

We note that action symbols differ importantly from the well-studied concept of motor primitives[135–137]. While both motor primitives and action symbols can be understood as building blocks of movement, they operate at different levels of abstraction and serve different roles. Motor primitives reflect movement parameters at a lower level of abstraction, at the level of spinal cord and muscle activity[136] (although see evidence in M1 activity[138]), and they contribute to movement by simplifying complex (rapidly varying, high-dimensional) movements into simpler (slowly varying, low-dimensional) control parameters, such as in muscle synergies[135]. In contrast, action symbols are higher-level cognitive representations that internally compose to organize action (potentially supporting processes like planning and imagination). Consistent with this, we found encoding of action symbols (i) during the planning phase and therefore not temporally linked to ongoing movements (**Figs. 4, 5**), (ii) exhibiting motor invariance in both behavior (**Fig. 2a-e**) and neural activity (**Fig. 4**), (iii) supporting compositional generalization (**Fig. 6**), and (iv) encoded in a higher-order cortical area, PMv (**Figs. 4-6**).

A foundation for understanding symbolic operations in neural representations and dynamics

This action symbol representation in PMv may provide an entry point to bridge between two paradigms that have dominated the modeling of cognition: one based on symbolic representations and rule-based operations[2–10] and the other on neural network architectures and dynamical systems[33,34,36,37,39]. While these views have sometimes been considered to be at odds, it is recognized that explaining cognition depends on their unification, possibly by implementing symbolic operations in appropriate neural representations and dynamics[25,47,139–144]. One promising approach expresses symbolic computation in terms of symbolic algorithms analogous to computer programs[4,7,10,14–17,47]. These algorithms, in turn, may potentially be understood in terms of neural representations and dynamics[47,65,139–143] (possibly building on insights from task-optimized network models[34,35,39,145,146]). Importantly, implementations of

symbolic algorithms in neural processes can lead to specific predictions for activity. To illustrate, the drawing program “repeat circle three times”—represented symbolically as REPEAT(◯, 3), combining the action symbol ◯ with a syntactic operation REPEAT—could be encoded in a multi-module neural architecture, with one REPEAT module counting up from one to three, and a second DRAW module that, on each increment of that counter, draws a circle. In the brain, such program-like activity may be observed in neural populations in PMv, and perhaps in PFC and preSMA, areas that are directly interconnected with PMv[86], and have been shown to encode abstract sequential information[28,46,147–149], including, in PFC, sequential activity states resembling program operations[47]. Looking forward, an important aspect of our finding an action-symbol-encoding population is that it could provide the substrate for future studies to test for activity dynamics consistent with symbolic operations in programs. Identifying such activity would be a significant step towards a computational explanation of intelligence that spans levels of behavior, cognition, and neural mechanisms, and that could generalize beyond action to other abilities that seem to reflect internal program-like algorithms[7,10,14].

Methods

Subjects and surgical procedures

Data were acquired from two adult male macaques (*Macaca mulatta*, average weights 17 kg (S1) and 10 kg (S2), average ages 9 years (S1) and 7 years (S2)). All animal procedures complied with the NIH Guide for Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee of the Rockefeller University (protocol 24066-H).

After undergoing initial task training in their home cages, subjects underwent two surgeries, the first to implant an acrylic head implant with a headpost, and the second to implant electrode arrays. Both surgeries followed standard protocol, including for anesthetic, aseptic, and postoperative treatment. In the first surgery, a custom-designed MR-compatible Ultem headpost was implanted, surrounded by a bone cement cranial implant, or “headcap” (Metabond, Parkell and Palacos, Heraeus), which was secured to the skull using MR-compatible ceramic screws (Rogue Research). After a six month interval, to allow bone to grow around the screws and for the subject to acclimate to performing the task during head fixation via the headpost, we performed a second surgery to implant 16 floating microelectrode arrays (32-channel FMA, Microprobes), following standard procedures[150]. Briefly, after performing a craniotomy and durotomy over the target area, arrays were inserted one by one stereotactically, held at the end of a stereotaxic arm with a vacuum suction attachment (Microprobes). Using vacuum suction allowed us to release the arrays, after insertion, with minimal mechanical perturbation by turning off the suction. After all arrays had been implanted, the dura mater was loosely sutured and covered with DuraGen (Integra LifeSciences). The craniotomy was closed with bone cement.

We used standard density arrays (1.8 mm x 4 mm) for all areas, except SMA and preSMA, for which we used four high density arrays (1.6 mm x 2.95 mm). Four additional electrodes on each array served as reference and ground. Two arrays were targeted to each of multiple areas of frontal cortex, with locations identified stereotactically, and planned using brain surface reconstructions derived from anatomical MRI scans. Locations were selected based on their published functional and anatomical properties (see below), anatomical sulcal landmarks, and a standard macaque brain atlas[151]. During

surgery, locations were further adjusted based on cortical landmarks, and to avoid visible blood vessels. Arrays were implanted in the right hemisphere (contralateral to the arm used for drawing).

Array locations are depicted in **Fig. 3b** (confirmed with intraoperative photographs). For M1, we targeted hand and arm representations (F1), directly medial to the bend of the central sulcus (which corresponds roughly to the intersection of the central sulcus and the arcuate spur if the latter were extended caudally), based on retrograde labeling from spinal cord and microstimulation of M1[87] and M1 recordings[118]. For PMd, we placed both arrays lateral to the precentral dimple, with one (more caudal) array directly medial to the arcuate spur (the arm representation[87,118,152], F2), and the other was placed more rostrally (straddling F2 and F7). For PMv, we targeted areas caudal to the inferior arm of the arcuate sulcus (F5), which are associated with hand movements based on retrograde labeling from spinal cord[87] and M1[153], microstimulation[154] and functional studies[100,105,155]. These areas contain neurons interconnected with PFC[153]. For SMA (F3) and preSMA (F6), we targeted the medial wall of the hemisphere, with the boundary between SMA and preSMA defined as the anterior-posterior location of the genu of the arcuate sulcus, consistent with prior studies finding significant differences across this boundary in anatomical connectivity (e.g., direct spinal projections in SMA but not preSMA[156]) and function[46,157]. SMA arrays were largely in the arm representation[156]. For dlPFC, we targeted the region immediately dorsal to the principal sulcus (46d), following prior studies of action sequencing[28,47,158] and other cognitive functions[159]. For vlPFC, we targeted the inferior convexity ventral to the principal sulcus, with one (more rostral) array directly ventral to the principal sulcus (46v) and the other rostral to the inferior arm of the arcuate sulcus (45A/B), based on evidence for encoding of abstract concepts in regions broadly spanning these two locations[52,160,161], including a possible heightened role (compared to dlPFC) in encoding abstract concepts in a manner invariant to temporal or spatial parameters[160,162,163]. For FP, we targeted a rostral location similar to prior recording and imaging studies (one array fully in area 10, the other straddling 9 and 10)[164,165]. In general, array locations targeted the cortical convexity immediately next to sulci, instead of within the banks, in order to allow shorter insertion depths that minimize the risk of missing the target. The exceptions were SMA and preSMA in the medial wall, for

which this was not possible. To avoid damaging the superior sagittal sinus, we positioned the arrays laterally (2 mm from midline) and slanted the electrodes medially (**Extended Data Fig. 4**).

The lengths of each electrode were custom-designed to target half-way through the gray matter, and to vary substantially across the array, to maximize sampling of the cortical depth. Electrode lengths spanned 1.5 - 3.5 mm (M1), 1.5 - 3.1 mm (PMd, PMv), 2.8 - 5.8 mm (SMA, preSMA), 1.5 - 2.5 mm (dlPFC, vlPFC), and 1.5 - 2.6 mm (FP) for subject 1, and 1.7 - 3.75 (M1), 1.5 - 3.3 mm (PMv), 1.5 - 3.1 mm (PMd), 2.65 - 5.95 mm (SMA, preSMA), 1.75 - 3.15 mm (dlPFC), 1.35 - 3.2 mm (vlPFC), and 1.6 - 2.9 mm (FP) for subject 2. Reference electrodes were longer (6 mm) to anchor the arrays. 28 electrodes were Pt/Ir (0.5 MΩ) and 4 were Ir (10 kΩ) (for microstimulation). Array connectors (Omnetics A79022) were housed in custom-made Ultem pedestals (Crist), which were secured with bone cement onto the cranial implant. Four pedestals were used per subject, holding 5, 5, 4, and 2 connectors each.

Behavioral task

<u>Task overview</u>

Subjects were seated comfortably in the dark with their head restrained by fixing the headpost to the chair. They faced a touchscreen (Elo 1590L 15” E334335, PCAP, 768 x 1024 pixels, refresh rate 60 Hz, with matte screen protector to reduce finger friction) that presented images and was drawn on. The touchscreen location was optimized to allow each subject to easily draw at all relevant locations on the screen (23 to 26 cm away; see diagram in **Extended Data Fig. 1**). Both subjects decided on their own over the course of learning to perform the task with the left hand. The chairs were designed to minimize movements of the torso and legs (by using a loosely restricting “belly plate”), and of the non-drawing arm (by restricting movement to within the chair). Gravity-delivered reward (water-juice mixture) was controlled by the opening and closing of a solenoid pinch valve (Cole-Parmer, 1/8" ID). Subjects were water-regulated, with careful monitoring that consumption met the minimum requirement per day (and typically exceeded it), and weight was closely monitored to ensure good health. The task was controlled with custom-written software, using the MonkeyLogic behavioral control and data acquisition MATLAB package[166]. (PC: Windows 10 Pro, Intel Core i7-4790K, 32GB RAM; DAQ: National

Instruments PCIe-6343). All stimuli (images of line figures defined as point sets) were also generated with custom-written MATLAB code. Images were presented in a “workspace” area on the screen (16.6 cm x 16.9 cm, corresponding to approximately 37° by 38° visual angle). Shape components in images were on average 4.0 cm (9°) (taking the maximum of width and height).

Each neural recording session consisted of a day’s recording (2 - 3.5 hours). We collected 5-20 trials per condition (i.e., each unique image for **Figs. 4, 5** and single-shape tasks in **Fig. 6**; each primitive stroke for character tasks in **Fig. 6**). All trials were shuffled across all conditions within the session, and presented in a randomly interleaved fashion, except for one case, the experiment in **Fig. 6**, in which character and single-shape tasks were switched in blocks.

Early Training

Before surgery, naive subjects underwent initial training on core task components (i.e., to trace images accurately on a touchscreen, using a sequence of discrete strokes). Early training took place in the home cage, using custom-built rigs that attached to an opening in the cage, using the same hardware and software described above, except for a different computer (Lenovo IdeaPad 14" laptop, Windows 10, AMD Ryzen 5 3500U, 8GB RAM) and DAQ (National Instruments USB-6001). This initial training progressed through the following stages: *(1) Touch circle*. Subjects were rewarded for touching a circle anywhere within its bounds. The circle started large, filling the entire screen, and shrank over trials to enforce more accurate touches. *(2) Touch with a single finger*. We shrank the circle until it was so small that it could only be touched with a single finger. The trial aborted if the subject touched outside the circle, or with multiple fingers simultaneously. *(3) Hold still*. Subjects were rewarded for keeping their fingertip still on a dot, with the duration of this hold increasing across trials (up to a few seconds). *(4) Track moving dot.* Subjects had to track the dot with their finger as it moved (a lag between dot and finger was allowed). *(5) Trace a line.* We increased the speed of the moving dot over trials, until eventually the dot moved so fast that the line it traced appeared immediately. We then positioned the line at locations far from the hold position to train the subject to raise its finger from the hold position and trace lines at arbitrary locations, angles, and lengths. *(6) Trace single shape.* We presented shapes

of increasing difficulty (gradually “morphing” across trials from a straight line), including arcs, “L”-shapes, and squiggles, and circles. Across these stages, the shapes were presented at random locations. We did not enforce any particular tracing strategy for each shape (e.g., which endpoint to start at), allowing subjects to choose on their own. *(7) Trace multiple shapes.* We presented images composed of multiple disconnected shapes. This trained the subjects to understand that they should use multiple strokes to trace multiple shapes. At this point, the subject understood the basic structure of the task—to trace shapes, using multiple strokes if needed. The progression across these stages was not determined by strict quantitative criteria, but instead on a combination of quantitative and qualitative evaluations of how well the subjects understood the task.

After this basic training, subjects practiced various tasks to incentivize the learning of stroke primitives (consistent stroke trajectories for each shape). They practiced single-shape trials using the set of diverse simple shapes in **Fig. 1e**, varying randomly in shape and location across trials. On different days, subjects also practiced multi-shape and character tasks. Below, we describe the trial structure, followed by details for single-shape, multi-shape, and character tasks.

<u>Trial structure</u>

As depicted in **Fig. 1c**, trials began when the subject pressed and held a finger fixation button (blue square) at the bottom of the screen (grey background; note that in this article “button” always means a virtual button; diagram of screen in **Extended Data Fig. 1c**). After a random delay (uniform, 0.2 s window, earliest [0.4, 0.6] s and latest [0.8, 1.0] s across experiments for subject 1 and [0.8, 1.1] s for subject 2) the image appeared (figure colored dark grey). After a random delay (uniform, ranging from [0.6, 1.0] s to [1.2, 1.6] s across experiments for subject 1, and [1.1, 1.5] s to [1.8, 2.4] s for subject 2), a “go” cue (400 Hz tone and image blank for 300 ms) was presented. This delay between image presentation and go cue we call the “planning” epoch. The subject had to keep its finger still on the fixation button during the planning epoch. After the onset of the go cue, the subject was free to raise its finger, move its hand towards the image, and start drawing. During drawing, the image stayed visible, and the finger left a trail of black “ink” on the screen. Immediately after the finger was raised from the

fixation button after the go cue, the “fixation button” disappeared and a “done button” (green square) appeared at its location and stayed on. The drawing epoch ended when the subject signaled it had completed the drawing, by pressing the “done button” (no time limit was imposed). This was followed by a delay (uniformly sampled, 0.65 to 1 s), followed by performance feedback. Feedback spanned four modalities, each signalling performance: (i) screen color, (ii) sound cue, (iii) duration of a delay before getting reward, and (iv) reward magnitude. First, screen color and sound were signaled, followed by the pre-reward delay and then the reward. How performance was scored and converted to feedback is described below. In addition to providing this feedback at the end of the trial, we also provided feedback online by immediately aborting a trial if subjects made serious errors, including (i) if they touched a position far from any image points, and, (ii) for single-shape and multi-shape trials, if they used more than one stroke per shape in the image. These “online abort” modes were turned off for trials testing novel characters.

Screen image changes (including image presentation and trial events) were recorded using photodiodes (Adafruit Light Sensor ALS-PT19) and sounds were recorded using an electret microphone (Adafruit Maxim MAX4466, 20-20KHz). We performed eye tracking (ISCAN), but did not enforce eye fixation.

<u>Scoring behavioral performance</u>

Behavior was scored by aggregating multiple metrics, or “factors”. There were three classes of factors. The primary class measured *image similarity*, or the similarity of the final drawing to the target image (ignoring its temporal trajectory). This factor had the greatest influence on the final aggregate score. Additionally, we computed factors reflecting *behavioral efficiency*, and, in some cases, factors that were *task-specific*. These scores were computed on behavioral data, which was represented as a sequence of touched points (x-y coordinates) with gaps between strokes, and image data, represented as a set of x-y coordinates. First, we describe the factors, and then how they were aggregated into a single score.

Image similarity. This included two factors: drawing-image overlap and Hausdorff distance. Drawing-image overlap was computed as the fraction of the image points that were “touched” (within a

margin of error) by at least one of the drawn points. A subset of the image points were weighted more heavily because they captured unique features of the shape (e.g., the corners and endpoints of an “L” shape). Hausdorff distance is a distance metric between the set of drawn points and the set of image points (see definition below).

Behavioral efficiency. To incentivize efficiency, we included a factor comparing the cumulative distance traveled in the drawing (i.e., the amount of “ink”) to the cumulative distance of the edges of the figure in the image, with its value negatively proportional to the excess of drawn ink over image ink.

Task-specific factors. During practice trials for character tasks (see “Task types”), we also included factors capturing the extent to which drawn strokes matched the shapes used in the image figure. This included two factors: one proportional to the similarity of the number of strokes and the number of image shapes, and the other proportional to the spatial alignment of the drawn strokes to the image shapes. Importantly, these factors were included only for practice images, not for novel test images.

The final score aggregated the image similarity, behavioral efficiency, and task-specific factors, with more weight on image similarity factors. We first rescaled the factors linearly between 0 and 1 (where 1 means good performance), with the dynamic range set by a lower and upper bound on the factor values. These bounds were adaptively updated on every trial based on the distribution of factor values in the last 50 trials (lower bound at 1st percentile and upper bound at 53rd percentile), which ensured that the dynamic range of feedback matched the dynamic range of behavioral performance from recent history. We then weighted each of those factors to tune their relative contributions (using hand-tuned weights for each experiment; generally highest for image similarity), and computed the final scalar score (range 0 to 1) by taking the single factor that was worst after weighting.

$$s_{scal} = \min_i \left(1 - w_i(1 - f_i)\right)$$

where i indexes the factors, w_i are the weights (between 0 and 1) and f_1 are the factor values (rescaled between 0 and 1). We also gave each trial a categorical score:

$$s_{cat} = \begin{cases} \text{great} & s_{scal} > 0.82 \\ \text{good} & 0.65 < s_{scal} \leq 0.82 \\ \text{ok} & 0.15 < s_{scal} \leq 0.65 \\ \text{fail} & s_{scal} \leq 0.15 \end{cases}$$

The scalar and categorical scores were used to determine feedback across four different modalities. The meaning of screen color and sound were learned by the subject, whereas delay and reward had intrinsic value:

(1) Screen color. A linear mapping between two colors, such that a score of 0 was mapped to red (RBG: [1, 0.2, 0]) and a score of 1 was mapped to green ([0.2, 1, 0.2]).

(2) Sound cue. A sound determined by s_{cat}: if *great*, then three pulses (1300 Hz, 0.16 s on and off); if *good*, then a single pulse (1000 Hz, 0.4 s); if *ok*, then no sound; if *fail*, then a single pulse (120 Hz, 0.27 s).

(3) Delay until reward. A nonlinear mapping from score to delay before reward. We first applied a linear mapping to the scalar score, such that a score of 0 was mapped to a long delay (5 sec + random uniform jitter, [0, 2.5] s) and a score of 1 was mapped to 0 s delay. Further, if s_{cat} was *great*, *good*, or *ok*, then this delay was reduced by multiplying by 0.65.

(4) Reward. The open duration of the solenoid gating the juice line was defined as

$$\text{rew} = \text{C} \times m \times a \times s_{scal}$$

where C is a constant in dimensions of time (0.15 - 0.6 s, set manually depending on the difficulty of the task); m is a multiplier that gives a bonus for good performance and further penalizes bad performance, depending on the value of s_{cat}: *great* (1.3), *good* (1.0), *ok* (0.8), *fail* (0); and a is a random variable sampled from the uniform distribution $a \sim 0.75 + 0.5 \times U(0,1)$, and s_{scal} defined as above. On average, including failed trials, subjects received ~0.35 ml reward per trial. The order in which these four feedback signals were delivered is described above ("Trial structure").

<u>Task types</u>

Single-shape tasks. Single-shape tasks presented one of the practiced simple shapes, or, in the "categories" experiments, sometimes a morphed shape. Subjects were allowed only a single stroke

(triggering online abort if >1). In four single-shape sessions for subject 1, the ending of the drawing epoch was triggered by completion of the stroke (i.e., on finger raise), and not on the subject pressing a "done" button, as was the case in all other sessions and experiments.

To test for motor invariance (**Figs. 2a-e, 4**), we presented images of practiced shapes, varying across trials in location, size, or both. For location variation, images spanned a distance in the x- and y-dimensions (measuring from shape centers) of 321 pixels (9.6 cm), which is 2.38 times the average size of shapes (135 pixels, 4.0 cm, maximum across width and diameter). For size variation, the maximum size was 2.5 times larger than the smallest (in diameter), except for two experiments for subject 1, in which the ratio was 2.0.

To test for categorical structure (**Figs. 2f-j, 5**), we constructed morph sets (subject 1: N = 7 morph sets across 3 sessions; subject 2: N = 13 morph sets across 4 sessions), each consisting of two practiced shapes and four to five "morphed" versions of those shapes, which were constructed by linearly interpolating between the two shapes along one image parameter, such as the extent of closure of the top of the U (**Fig. 2f, g**). Across morph sets, we varied different image parameters (see examples in **Extended Data Fig. 2**).

Multi-shape tasks. Each image was composed of two to four shapes positioned at random, non-overlapping, locations spanning the space of the screen (four corners and center). Subjects were allowed to draw the shapes in any order, and to use any trajectory within the shapes, but were constrained to use one stroke per shape and to not trace in the gaps between shapes. In the Results of this manuscript, we present results averaged across two sessions, one from each subject (**Extended Data Fig. 3**). On each trial, an image was constructed by sampling a shape randomly without replacement. This led to N = 531 (S1) and N = 278 (S2) unique images.

Character tasks. Each image was generated by connecting two to six simple shapes into a single character figure. This was done by sampling characters from a generative model. A character with N shapes was defined by randomly sampling N shapes and N-1 relations, where each relation i defines the location of the attachment points on shapes i and $i+1$; these attachment points, in turn, define how these shapes will connect to each other. This is similar to a published generative model for

handwritten characters[16]. Generated characters were only kept if there was minimal crossing of shapes over each other.

For experiments testing behavioral generalization to novel characters (**Fig. 2k-p**), we mixed practiced and novel characters (practiced characters: N = 189 (mean, range 22 to 491) per day; novel characters: 48 (mean, range 0 to 155) per day). For analyses, we label as “novel” only the very first trial for a given character. Because of random sampling in generating characters, it would in principle be possible that characters generated on different days are in fact identical; to avoid this, we ensured that all characters we called “novel” indeed did not match any images the subject previously encountered. We did this by ensuring that each novel character was different from every previously encountered character across all days (quantified with the Hausdorff distance).

For neural experiments comparing single-shape and character tasks (**Fig. 6**), we analysed the sessions for which we collected data from both single-shape tasks and character tasks [subject 1: N = 10 sessions, median N matching primitives between single-shape and character tasks = 9 (range 5 - 12); subject 2: N = 9 sessions, median N matching primitives = 10 (range 2 - 14)]. We switched between single-shape and character tasks using a block design (2 - 5 blocks each), except one session for subject 2, in which they were randomly interleaved across trials.

Behavioral data analysis

<u>Preprocessing of touchscreen data</u>

Touchscreen data were represented as time series of (x, y) coordinates in units of pixels (conversion: 33.6 pixels/cm) and sampled at 60 Hz, which we upsampled to 500 Hz and low-pass filtered (15 Hz). Strokes were segmented based on the time the finger first touched the screen (stroke onset) and the last time before raising off the screen (stroke offset) with 500 Hz resolution.

To compute stroke instantaneous velocity and speed—in **Fig. 1f** and as input to the “trajectory distance” below—we first further low-pass filtered the data (12.5 Hz), and downsampled to 25 Hz. We then used the five-point stencil method to compute a finite difference approximation of the derivative:

$$f'[n] = \frac{f[n-2] - 8f[n-1] + 8f[n+1] - f[n+2]}{12\mathrm{h}}$$

where $f[n]$ is a discrete time series (i.e, the x- or y-coordinates) indexed by integer n, and h is the sampling period in seconds. This differentiation was performed separately for the x and y coordinates. The resulting velocity time series was upsampled to the original 500 Hz sampling rate with a cubic spline. Speed was computed as the norm of the (x, y) velocity at each timepoint.

<u>Computing “trajectory distance”</u>

To quantify the similarity between two strokes, in a way that compares their spatio-temporal trajectories, while ignoring their relative size (or scale) and location (on the screen), we devised a “trajectory distance” metric. This metric is a scalar dissimilarity score, based on the dynamic time warping distance between two strokes represented as velocity time series $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. To compute trajectory distance between two strokes, we (1) spatially rescaled each stroke (while maintaining its x-y aspect ratio), to make the diagonal of its bounding box unit length 1. (2) We then linearly interpolated each stroke to the same number of points (70) to allow direct point-by-point comparison between strokes. This was done spatially by interpolating based on the fraction of cumulative distance traveled (so that the distances between successive points were the same value over the entire stroke), in order to capture the spatio-temporal trajectory, as in a previous study modeling strokes in handwriting[16]. (3) Interpolated trajectories were then converted to velocity time series as above. (4) We then computed the dynamic time warping distance between these velocities $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.:

$$D_{\mathrm{DTW}}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\min_\pi \sum_{(i,j)\in\pi} d(i,j)}{\mathrm{N}}$$

where i and j index the two velocity trajectories, N is the number of points (70), π is a set of (i,j) pairs representing a contiguous path from (0, 0) to (N, N). The local distance metric $d(i,j)$ is the Euclidean distance plus a regularization factor that discourages excessive warping:

$$d(i,j) = \|\mathbf{v}_1[i] - \mathbf{v}_2[j]\| + \lambda \times |i-j|$$

$$\lambda = 0.045 \left\langle |\mathbf{v}_n[i]| \right\rangle_{i,n}$$

where $\langle \cdot \rangle$ is the average. For the regularization parameter, λ, the purpose of the summation term was to rescale lambda to match the magnitude of velocities. The resulting distance D_{DTW} was then rescaled to 0 and 1 to return the trajectory distance:

$$D_{traj}(\mathbf{v_1}, \mathbf{v_2}) = 1 - \frac{1}{D_{DTW}(\mathbf{v_1}, \mathbf{v_2}) + 1}$$

Computing “image distance”

To compare the similarity of two images—each a figure represented as a set of (x, y) points, with no temporal or stroke-related information—we used a modified version of the Hausdorff distance, a distance metric that has been commonly used in machine vision for comparing the similarity between two point sets based on their shape attributes[167]. There are, in principle, at least 24 variants of the Hausdorff distance based on possible variations in the formula[167]; we used a variant that is minimally susceptible to outlier points (because it is based on taking means instead of minima and maxima; variant #23 in the referenced study[167]). Image distance was computed as follows: (1) Each image was first centered so that its center of mass becomes (0,0). (2) Image distance was then computed. First, we define the distance between two points, $d(a, b)$, as the Euclidean distance. We also define the distance between a point and a set of points, $d(a, B)$, and the distance from set A to set B, $d(A, B)$, as:

$$d(a, B) = \min_{b \in B} d(a, b)$$

$$d(A, B) = \frac{1}{|A|} \sum_{a \in A} d(a, B)$$

Image distance was then defined as:

$$D_{image}(A, B) = \frac{d(A, B) + d(B, A)}{2}$$

Computing “primitive alignment score”

For experiments on categorical structure, we generated a set of images, with each set containing four to five novel images that morph between one primitive (P1) and another primitive (P2): a “morph set”. Across morph sets, different image parameters were morphed (see **Fig. 2g** and **Extended Data Fig. 2**). Each trial presents a single image from one morph set.We sought to quantify the relative similarity between a given trial’s data—either its behavioral, image, or neural data (see below)—and data for the two primitives, P1 and P2, in its morph set. To do so, we devised a “primitive alignment” score defined, for each individual trial, as:

$$a = \frac{d_1}{d_1 + d_2}$$

where d_1 is the average of the distances between that trial and each of the P1 trials, and d_2 the average of the distances between that trial and each of the P2 trials. A score closer to 0 implies similarity to P1, and a score close to 1 implies similarity to P2. The particular metric used for these distances depended on the analysis. For images, we used the image distance. For drawings, we used the trajectory distance. For neural activity, we used the Euclidean distance between population vectors. We confirmed that primitive alignment scores for image data varied linearly with morph number (Fig. 2j and Extended Data Fig. 2c), ensuring that any deviation from linearity in behavioral or neural data could not trivially be the consequence of how the score is defined.

<u>Classifying strokes from character tasks</u>

To assess whether subjects drew characters by reusing their own stroke primitives, we scored the fraction of character strokes that were high-quality matches to one of the subject’s own primitives, and the fraction that were high-quality matches to the other subject’s primitives. If the fraction of matches to a subject’s own primitives was high, and to the other subject’s primitives was low, then we interpreted this as evidence that subjects recombined their own primitives.

This was performed by assigning each stroke the label of its nearest primitive using the trajectory distance, and then further defining this as a high-quality match only if the trajectory distance

was sufficiently low; in particular, if the distance was within the 95% CI of the expected distribution of trajectory distances caused by trial-by-trial variation in behavior.

First, each stroke was assigned its best-matching primitive, p^* from a set of primitives (the choice of primitive set—same or different subject—depending on the analysis; see below):

$$p^* = \arg\min_p d(\mathbf{s}, \mu_p)$$

Where p indexes the primitives, s is the stroke trajectory, μ_p is the mean stroke for primitive p (averaged over trials from single-shape tasks), and $d(\cdot,\cdot)$ is the trajectory distance.

Second, the quality of the stroke's assignment to its nearest primitive was scored:

$$quality = \begin{cases} \text{high} & \text{if } d(s, \mu_{p^*}) < \mathrm{D}_{\max,p^*} \\ \text{low} & \text{if } d(s, \mu_{p^*}) \geq \mathrm{D}_{\max,p^*} \end{cases}$$

Where $\mathrm{D}_{\max,p^*}$ is an upper-bound on trajectory distances that would be expected from trial-by-trial variation in primitive p^*. It is the 97.5th percentile of the distribution of trajectory distances from single-shape trials, determined separately for each primitive, which we consider as a good—if anything, conservative—estimate of trial-by-trial variation, because single-shape tasks present no ambiguity as to what primitive needs to be drawn.

These steps assigned each stroke a class tuple $(p^*, quality)$. In summary analyses, we pooled all cases of high-quality matches into a single "high-quality match" class (no matter the assigned primitive), and all low-quality matches into a single "no match" class (**Fig. 2o, p**). In summary analyses testing whether a given subject's character strokes aligned better with its own primitives vs. the other subject's primitives (**Fig. 2p**) we performed the above analysis separately for all four combinations of stroke data (2 subjects) x primitives (2 subjects), using only images performed by both subjects.

Neural recordings

Recordings were acquired using a Tucker-Davis Technologies (TDT) system, including headstage (Z-Series 32 Channel Omnetics, LP32CH-Z), amplifier (PZ5M-256), processor (RZ2), and storage (RS4), sampled at 25 kHz (local reference mode), controlled with TDT Synapse software run on a

Windows 10 PC (Intel Core i7-3770, 32GB RAM), and saved to disk. Analog and digital task-related signals, including behavioral events (photodiode, audio, and trial event markers) and eye tracking (ISCAN, 125 Hz), were synchronized external triggers recorded by the neural data acquisition system.

Neural data preprocessing

Spike sorting

We extracted for later analysis both isolated single-unit (SU) and multi-unit (MU) spike clusters from the stored broadband signal. MU clusters consisted of threshold crossings that were clearly spikes, but which were not isolatable into distinct SU clusters. We used a three-step approach for extracting and sorting spikes into these clusters, with a first pass using Kilosort[168] (v2.5) to extract putative spike clusters, a second pass using a custom-written program to label these clusters as either SU, MU, or noise, and a final manual curation step. Note that Kilosort classifies clusters, but we did not use those labels.

For Kilosort, we used the default parameters, except *AUCsplit* (0.90), *Th* ([6, 4]), and *lam* (10), which we optimized using parameter sweeps on data from representative sessions and by manual evaluation of results.

We next refined these cluster labels. For each cluster, we first removed outlier waveforms (any exceeding a 3 x interquartile-range threshold for any of the minima, maxima, or sum-of-squares). Waveforms were then shifted slightly in time (<1 ms) to improve their alignment by peaks (or toughs, for negative-going waveforms). For each cluster, we computed two features. *(1) Signal-to-noise ratio (SNR)*, defined as the ratio of the peak-to-trough difference (of the average spike waveform) divided by standard deviation (averaged across time bins). Before computing SNR, we checked whether the cluster contained both positive- and negative-going waveforms. If so, SNR was computed separately for these two subsets of data and then averaged. *(2) Inter-spike-interval violations (ISIV)*, defined as the fraction of inter-spike intervals less than a refractory period (1.5 ms). Based on these SNR and ISIV, we provisionally classified clusters as SU [if either (SNR > 9.6 and ISIV < 0.05) or (SNR > 6.9 and ISIV < 0.01)], noise (SNR < 3.9), or MU (the remaining clusters).

We then manually curated these clusters. We visualized every cluster to either confirm its label (MU, SU, noise) or to manually re-assign it to a different label (including “artifact”), using a custom-written MATLAB GUI. We also manually checked whether multiple SU clusters on a single channel should be merged into a single SU cluster, if they have high waveform similarity, inversely correlated spike count frequency over the course of the session, or a negative peak close to zero lag in a cross-correlogram of spike times. Finally, for each channel, all MU clusters were merged into a single MU cluster. Combining SU and MU, this yielded the following total number of units per area (mean +/- S.D. across sessions). For subject 1: M1 (59.9 +/- 12.5), PMd (44.1 +/- 6.2), PMv (34.2 +/- 7.3), SMA (63.0 +/- 7.9), preSMA (75.4 +/- 17.7), dlPFC (47.8 +/- 17.2), vlPFC (43.3 +/- 9.9), FP (19.2 +/- 3.8). For subject 2: M1 (40.7 +/- 13.1), PMd (54.7 +/- 5.5), PMv (71.1 +/- 6.6), SMA (53.4 +/- 7.2), preSMA (57.9 +/- 11.1), dlPFC (24.6 +/- 4.8), vlPFC (38.6 +/- 13.7), FP (42.6 +/- 5.0).

<u>Converting spike times to firing rates</u>

Single-trial spike trains were converted to firing rate functions by smoothing with a 0.025 s Gaussian kernel (0.01 s slide). We removed units with very low firing rates (if the 80th percentile of their firing rates across all trials and time bins was less than 1 Hz). We square-root transformed activity to normalize its variance. Following a common approach in analyses of population firing rates[148], we “soft” z-scored each unit’s activity to ensure that neurons with very different firing rates contributed similarity to population analyses, but with higher-firing-rate neurons still contributing relatively more:

$$x_i^{\mathrm{norm}}(t) = \frac{x_i(t) - \mu_x}{\sigma_x + C}$$

Where $x_i(t)$ is firing rate for trial i at time bin t, μ_x and σ_x are the mean and standard deviation, respectively (across trials and time bins), and C is an additive factor to ensure “softness”: $C = \min(\mathbf{m}) + 3\mathrm{Hz}$ where $\mathbf{m}$ is a vector of mean firing rates, one for each unit. All subsequent analyses used this normalized firing rate representation of the data.

<u>Time-warping neural activity to a common trial template</u>

For the figure showing the average firing rates over the entire trial (**Fig. 3**), we first time-warped each trial to a common trial template. We defined a set of events that occur across trials as “anchors” (fixation touch, image onset, go cue, finger raise off fixation, stroke onset, stroke offset, touch done button, reward). We included only single-stroke trials. We first generated a “median trial”. For each segment (i.e., time window between a pair of successive anchor events), we found its median duration, and then concatenated these median segments to construct a median trial. We then aligned each trial to this median trial at the anchor events, warping time linearly within each segment. To avoid sharp discontinuities at anchor points, we smoothed the final firing rates at the times of the anchor points (2.5 ms Gaussian kernel). This warping did not change the firing rate values, just their timing.

Neural data analyses

<u>Dimensionality reduction of population activity</u>

We performed dimensionality reduction on the neural population activity, in general because high-dimensional noise can reduce the interpretability of the Euclidean distance[126], and in one case in order to identify a potential linear projection of population activity (i.e., a subspace) that preferentially encodes primitives, a standard approach[47,146]. We represent a single area’s data from a single session and within a specific within-trial time window as a matrix $\mathbf{X}$, of size N x KT, where N, K, and T are the number of units, trials, and time bins, constructed by concatenating time bins from all trials along the second dimension. Data were first binned in time (0.15 s window, 0.02 s slide) before constructing this data matrix. We used principal components analysis (PCA), but instead of applying PCA on single-trial data $\mathbf{X}$, we applied PCA on trial-averaged data $\mathbf{X}_C$, in order to minimize the influence of trial-by-trial variation (noise). $\mathbf{X}_C$ holds the mean activity for each trial condition, of size N_C x KT, where N_C is the number of unique conditions, and where the specific conditions depended on the experiment (see below). We performed PCA on $\mathbf{X}_C$, and retained the top eight principal components (PCs). The specific trial-averaged conditions used for identifying PCs were the following. For analysis of motor invariance (**Fig. 4**), the conditions were each unique primitive (averaging over location/size), which resulted in identifying PCs that preferentially encoded primitives if they exist in the dataset. For analysis

of categorical structure (**Fig. 5**), PCA was performed separately for each morph set, and the conditions were the unique images (i.e., the two endpoint shapes plus the morphed shapes in between). For the analysis of primitive representational reuse in characters (**Fig. 6**), the conditions were each combination of primitive and task kind (i.e., resulting in *num_primitives* x *2* conditions).

We performed PCA in a cross-validated manner, to ensure that it was not overfitting to noise. We partitioned trials into two subsets (in a stratified manner), one “training” set that was used only for identifying the PCs, and a “test” set that was projected onto these PCs and then used for all subsequent analyses. We performed 8 randomized train-test splits (including all downstream analyses), and averaged their results.

<u>Computing "neural distance”</u>

To quantify the similarity of population activity between two sets of trials, such as trials for conditions A and B (where A and B are specific values of task-relevant variables), we devised a “neural distance” metric. This metric has the useful property of being unbiased (so that the expected value of the neural distance between two sets of trials sampled from the same distribution is zero). Inspired by the “normalized distance” in Liu et al[169], it is the average pairwise Euclidean distance across conditions A and B, minus the within-condition distances. This subtraction step ensures that the distance is unbiased (unlike the mean Euclidean distance, which is biased upwards[45]). In addition, the resulting distance is normalized by dividing by an upper-bound distance to normalize it between 0 and 1. Neural distance is defined as:

$$D^*_{AB} = D_{AB} - D_{AA} - D_{BB}$$

where the normalized Euclidean distance between sets of trial indices in conditions A and B is:

$$D_{AB} = \left\langle \frac{1}{d_{\max}(t)} \left\langle \left\| \boldsymbol{x}_i(t) - \boldsymbol{x}_j(t) \right\| \right\rangle_{i \in A, j \in B} \right\rangle_{t \in \{\mathrm{t}_1, \ldots, \mathrm{t}_\mathrm{n}\}}$$

Here, $\mathbf{x}_i(t)$ is the population activity vector at time t (in a window between times t_1 and t_n), and $d_{\max}$ is a normalization factor that is an upper-bound (98th percentile) of the distances between all pairs of different trials combined across all conditions.

Computing a variable's encoding strength

To compute how strongly a given variable is encoded in population activity (e.g., "primitive encoding" in **Fig. 4j**), we computed the mean effect of that variable on population activity, in terms of neural distance, while controlling for the other relevant variables. Consider an experiment varying two variables, primitive and location, such that a condition is represented by the tuple (p, l), where p and l index the primitives and locations. Primitive encoding is the average neural distance across all pairs of conditions that have different primitives but same locations:

$$\text{primitive encoding} = \left\langle D^*_{(p,l),(p',l')} \right\rangle_{p \neq p', l = l'}$$

Location encoding is defined analogously:

$$\text{location encoding} = \left\langle D^*_{(p,l),(p',l')} \right\rangle_{p = p', l \neq l'}$$

This approach generalizes to any pair of variables, such as primitive and task kind in **Fig. 6**.

Statistically comparing brain regions in strength of variable encoding

In analyses that compare the encoding strengths of a particular pair of variables (e.g. primitive vs. location in **Fig. 4j**), we performed the following statistical tests to compare each brain region with every other brain region, in terms of how strongly they encode these two variables. We used the following procedure. (1) For each variable and pair of brain regions, we performed a statistical test comparing how strongly these two regions encoded that variable. This involved first extracting a dataset of neural distances between each pair of trial conditions for each of the two brain regions. For example, if the variable was "primitive", then each of the two brain regions would contribute a dataset consisting of neural distance scores between all pairs of trial conditions that have different primitives but the same location. The datasets for these two regions would be combined into a single dataset, to which we fit a linear model to test for an effect of brain region on neural distance y, controlling for trial-condition pair:

$$y = \beta_0 + \beta_r X_r + \sum_{j=1}^{N_c} \gamma_j Z_j + \epsilon$$

where X_r is 0 or 1 depending on brain region, and Z_j is an indicator variable for trial-pair condition, with γ_j as their coefficients. Finally, we extracted the p-value for β_r (two-sided t-test), which represents the significance of the difference between this region pair in how strongly they encode the variable being tested. (2) This procedure was performed once for each combination of variable (2) and brain region pair (28), resulting in 56 p-values. We corrected the p-values for multiple comparisons using the Bonferroni method (28 brain region pairs x 2 variables = 56 comparisons). (3) Using these 56 values, we then summarized each region with two numbers representing the number of regions for which this region more strongly encodes these two variables. For example, for primitive x location experiments, each region was scored with a tuple (N_{prim}, N_{loc}), where N_{prim} is the number of other regions that this region beats in the pairwise statistical tests of primitive encoding (and analogous for N_{loc}, except that it scores location encoding). In summary plots (**Figs. 4j, 6f**), the results of this entire procedure were represented by mapping each region's resulting tuple to a color based on a 2D color map.

<u>Specific analyses</u>

Analysis of motor invariance in neural activity. Dimensionality reduction was performed as described above, using a time window of 0.05 s to 0.6 s after image onset for fitting the PCs and for analyses that involve time-averaging (**Fig. 4i-k**).

To test cross-condition decoder generalization (**Fig. 4k**), we used a linear support vector machine classifier (SVC), using a one-vs-the-rest scheme for multi-class classification (LinearSVC, scikit-learn, regularization parameter C set to 0.1). We report test accuracy linearly rescaled so that chance level (inverse of the number of classes) and 1 were mapped to 0 and 1. Because decoders were trained and tested on different conditions (with non-overlapping sets of trials), there was no concern of overfitting. Decoding was performed separately for, and averaged across, time bins (0.05 to 0.6 s relative to image onset).

Analysis of categorical structure in neural activity. Dimensionality reduction was performed as described above, using a time window from 0.05 s to 0.9 s after image onset for fitting the PCs. For analyses involving time-averaging (**Fig. 5d-g**), we used a window late in the planning period (0.6 to 1.0

s), when separation for A1 and A2 trials was the greatest (**Fig. 5h**). Primitive bias index was computed as above, using the Euclidean distance.

Analysis for recombination of primitive representations in character tasks. For each session, we presented both single-shape and character tasks using a block-interleaved design (except one session using a trial-interleaved design). We analyzed primitives that were performed in both single-shape (instructed by the shape image) and character trials (the subject's choice). For one experiment, the subject performed multi-shape but not single-shape tasks—we therefore used the first stroke from multi-shape trials instead of single-shape trials. For character trials, we used only strokes that were scored as high-quality primitive matches. Dimensionality reduction was performed as above, using a time window of -0.8 to 0.3 s relative to stroke onset for fitting the PCs. For analyses involving time-averaging, we used a window -0.5 to -0.05 s relative to stroke onset. Neural distance, primitive encoding, and task encoding were computed as above, using primitive and task kind (instead of location or size) as the two relevant variables.

Data and code availability

The data and code used in this study are available from the corresponding author (??) upon reasonable request.

References

1. Boden, M. A. *The Creative Mind: Myths and Mechanisms*. (Routledge, 2004).
2. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. (MIT Press, 2003).
3. Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**, 3–71 (1988).
4. Newell, A. & Simon, H. A. Computer Science as Empirical Inquiry: Symbols and Search. in (ed. Haugeland, J.) 81–110 (ACM, 1976). doi:10.7551/mitpress/4626.003.0004.
5. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
6. Frankland, S. M. & Greene, J. D. Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
7. Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S. & Sablé-Meyer, M. Symbols and mental programs: a hypothesis about human singularity. *Trends Cogn. Sci.* **26**, 751–766 (2022).
8. Do, Q. & Hasselmo, M. E. Neural circuits and symbolic processing. *Neurobiol. Learn. Mem.* **186**, 107552 (2021).
9. Nieder, A. Prefrontal cortex and the evolution of symbolic reference. *Curr. Opin. Neurobiol.* **19**, 99–108 (2009).
10. Piantadosi, S. T. The Computational Origin of Representation. *Minds Mach.* **31**, 1–58 (2021).
11. Quilty-Dunn, J., Porot, N. & Mandelbaum, E. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behav. Brain Sci.* **46**, e261 (2023).
12. Karmiloff-Smith, A. Constraints on representational change: Evidence from children's drawing. *Cognition* **34**, 57–83 (1990).
13. Mahowald, K. *et al.* Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517–540 (2024).
14. Rule, J. S., Tenenbaum, J. B. & Piantadosi, S. T. The Child as Hacker. *Trends Cogn. Sci.* **24**, 900–915 (2020).

15. Amalric, M. *et al.* The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLOS Comput. Biol.* **13**, e1005273 (2017).
16. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
17. Tian, L. Y., Ellis, K., Kryven, M. & Tenenbaum, J. B. Learning abstract structure for drawing by efficient motor program induction. *Adv. Neural Inf. Process. Syst. NeurIPS* (2020).
18. Fan, J. E., Bainbridge, W. A., Chamberlain, R. & Wammes, J. D. Drawing as a versatile cognitive tool. *Nat. Rev. Psychol.* **2**, 556–568 (2023).
19. Bläsing, B., Tenenbaum, G. & Schack, T. The cognitive structure of movements in classical dance. *Psychol. Sport Exerc.* **10**, 350–360 (2009).
20. Zuidema, W. & De Boer, B. The evolution of combinatorial structure in language. *Curr. Opin. Behav. Sci.* **21**, 138–144 (2018).
21. Studdert-Kennedy, M. & Goldstein, L. Launching Language: The Gestural Origin of Discrete Infinity. in *Language Evolution* (eds. Christiansen, M. H. & Kirby, S.) 235–254 (Oxford University Press, 2003). doi:10.1093/acprof:oso/9780199244843.003.0013.
22. Grosenick, L., Clement, T. S. & Fernald, R. D. Fish can infer social rank by observation alone. *Nature* **445**, 429–432 (2007).
23. Ferrigno, S., Huang, Y. & Cantlon, J. F. Reasoning Through the Disjunctive Syllogism in Monkeys. *Psychol. Sci.* **32**, 292–300 (2021).
24. Cheney, D. L. & Seyfarth, R. M. *Baboon Metaphysics: The Evolution of a Social Mind*. (University of Chicago Press, Chicago, IL, 2008).
25. Kazanina, N. & Poeppel, D. The neural ingredients for a language of thought are available. *Trends Cogn. Sci.* **27**, 996–1007 (2023).
26. Gallistel, C. R. Prelinguistic Thought. *Lang. Learn. Dev.* **7**, 253–262 (2011).
27. Ferrigno, S., Cheyette, S. J., Piantadosi, S. T. & Cantlon, J. F. Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Sci. Adv.* **6**, eaaz1002 (2020).
28. Shima, K., Isoda, M., Mushiake, H. & Tanji, J. Categorization of behavioural sequences in the

prefrontal cortex. *Nature* **445**, 315–318 (2007).

29. Jiang, X. *et al.* Production of Supra-regular Spatial Sequences by Macaque Monkeys. *Curr. Biol.* **28**, 1851-1859.e4 (2018).
30. Herman, L. M., Uyeyama, R. K. & Pack, A. A. Bottlenose dolphins understand relationships between concepts. *Behav. Brain Sci.* **31**, 139–140 (2008).
31. Livingstone, M. S. *et al.* Symbol addition by monkeys provides evidence for normalized quantity coding. *Proc. Natl. Acad. Sci.* **111**, 6822–6827 (2014).
32. Jordan, K. E., MacLean, E. L. & Brannon, E. M. Monkeys match and tally quantities across senses. *Cognition* **108**, 617–625 (2008).
33. McClelland, J. L. *et al.* Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* **14**, 348–356 (2010).
34. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
35. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
36. Gelder, T. van. The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* **21**, 615–628 (1998).
37. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation Through Neural Population Dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
38. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian Computation through Cortical Latent Dynamics. *Neuron* **103**, 934-947.e5 (2019).
39. Driscoll, L. N., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat. Neurosci.* **27**, 1349–1363 (2024).
40. Camp, E. THINKING WITH MAPS*. *Philos. Perspect.* **21**, 145–182 (2007).
41. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).
42. Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract

rules. *Nature* **411**, 953–956 (2001).

43. Bernardi, S. *et al. The Geometry of Abstraction in Hippocampus and Pre-Frontal Cortex*. http://biorxiv.org/lookup/doi/10.1101/408633 (2018) doi:10.1101/408633.

44. Rizzolatti, G., Luppino, G. & di Parma, U. The Cortical Motor System Review. (2001).

45. Willett, F. R. *et al.* Hand Knob Area of Premotor Cortex Represents the Whole Body in a Compositional Way. *Cell* **181**, 396-409.e26 (2020).

46. Tanji, J. Sequential Organization of Multiple Movements: Involvement of Cortical Motor Areas. *Annu. Rev. Neurosci.* **24**, 631–651 (2001).

47. Tian, Z. *et al.* Mental programming of spatial sequences in working memory in the macaque frontal cortex. *Science* **385**, eadp6091 (2024).

48. Diester, I. & Nieder, A. Semantic Associations between Signs and Numerical Categories in the Prefrontal Cortex. *PLOS Biol.* **5**, e294 (2007).

49. Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F. & Nieder, A. Single Neurons in the Human Brain Encode Numbers. *Neuron* **100**, 753-761.e4 (2018).

50. Bongard, S. & Nieder, A. Basic mathematical rules are encoded by primate prefrontal cortex neurons. *Proc. Natl. Acad. Sci.* **107**, 2277–2282 (2010).

51. Reinert, S., Hübener, M., Bonhoeffer, T. & Goltstein, P. M. Mouse prefrontal cortex represents learned rules for categorization. *Nature* **593**, 411–417 (2021).

52. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science* **291**, 312–316 (2001).

53. Freiwald, W. A. & Tsao, D. Y. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science* **330**, 845–851 (2010).

54. Neupane, S., Fiete, I. & Jazayeri, M. Mental navigation in the primate entorhinal cortex. *Nature* **630**, 704–711 (2024).

55. El-Gaby, M. *et al.* A cellular basis for mapping behavioural structure. *Nature* 1–10 (2024) doi:10.1038/s41586-024-08145-x.

56. Dijksterhuis, D. E. *et al.* Pronouns reactivate conceptual representations in human hippocampal

neurons. *Science* **385**, 1478–1484 (2024).

57. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).

58. Gupta, A. S., Meer, M. A. A. van der, Touretzky, D. S. & Redish, A. D. Hippocampal Replay Is Not a Simple Function of Experience. *Neuron* **65**, 695–705 (2010).

59. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).

60. Rizzolatti, G. & Craighero, L. THE MIRROR-NEURON SYSTEM. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).

61. Jiang, G. *et al.* Finding structure in logographic writing with library learning. Preprint at https://doi.org/10.48550/arXiv.2405.06906 (2024).

62. Tiedemann, H., Morgenstern, Y., Schmidt, F. & Fleming, R. W. One-shot generalization in humans revealed through a drawing task. *eLife* **11**, e75485 (2022).

63. Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M. & Tenenbaum, J. Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. Preprint at https://doi.org/10.48550/arXiv.1707.08212 (2017).

64. Toussaint, M., Allen, K., Smith, K. & Tenenbaum, J. Differentiable Physics and Stable Modes for Tool-Use and Manipulation Planning. in *Robotics: Science and Systems XIV* (Robotics: Science and Systems Foundation, 2018). doi:10.15607/RSS.2018.XIV.044.

65. Arbib, M. A. Perceptual Structures and Distributed Motor Control. in *Comprehensive Physiology* (ed. Terjung, R.) 1449–1480 (Wiley, 1981). doi:10.1002/cphy.cp010233.

66. Pastra, K. & Aloimonos, Y. The minimalist grammar of action. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 103–117 (2012).

67. Krüger, V., Kragic, D., Ude, A. & Geib, C. The meaning of action: a review on action recognition and mapping. *Adv. Robot.* **21**, 1473–1501 (2007).

68. Stout, D., Chaminade, T., Apel, J., Shafti, A. & Faisal, A. A. The measurement, evolution, and neural representation of action grammars of human behavior. *Sci. Rep.* **11**, 13720 (2021).

69. Jeannerod, M. The representing brain: Neural correlates of motor intention and imagery. *Behav. Brain Sci.* **17**, 187–202 (1994).

70. Raibert, M. H. Motor control and learning by the state space model. (Massachusetts Institute of Technology, 1977).

71. Kondapavulur, S. *et al.* Transition from predictable to variable motor cortex and striatal ensemble patterning during behavioral exploration. *Nat. Commun.* **13**, 2450 (2022).

72. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science* **291**, 312–316 (2001).

73. Prather, J. F., Nowicki, S., Anderson, R. C., Peters, S. & Mooney, R. Neural correlates of categorical perception in learned vocal communication. *Nat. Neurosci.* **12**, 221–228 (2009).

74. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).

75. Hoshi, E. & Tanji, J. Distinctions between dorsal and ventral premotor areas: anatomical connectivity and functional properties. *Curr. Opin. Neurobiol.* **17**, 234–242 (2007).

76. Mansouri, F. A., Freedman, D. J. & Buckley, M. J. Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* **21**, 595–610 (2020).

77. Tanji, J. & Hoshi, E. Role of the Lateral Prefrontal Cortex in Executive Behavioral Control. *Physiol. Rev.* **88**, 37–57 (2008).

78. Mansouri, F. A., Koechlin, E., Rosa, M. G. P. & Buckley, M. J. Managing competing goals — a key role for the frontopolar cortex. *Nat. Rev. Neurosci.* **18**, 645–657 (2017).

79. Wong, K.-F. & Wang, X.-J. A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *J. Neurosci.* **26**, 1314–1328 (2006).

80. Zimnik, A. J. & Churchland, M. M. Independent generation of sequence elements by motor cortex. *Nat. Neurosci.* **24**, 412–424 (2021).

81. van Kerkoerle, T. *et al.* Brain areas for reversible symbolic reference, a potential singularity of the human brain. *eLife* **12**, RP87380 (2025).

82. Penn, D. C., Holyoak, K. J. & Povinelli, D. J. Darwin's mistake: Explaining the discontinuity between

human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130 (2008).

83. Hauser, M. D., Chomsky, N. & Fitch, W. T. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science* **298**, 1569–1579 (2002).

84. Sablé-Meyer, M. *et al.* Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proc. Natl. Acad. Sci.* **118**, e2023123118 (2021).

85. Yang, C. Ontogeny and phylogeny of language. *Proc. Natl. Acad. Sci.* **110**, 6324–6327 (2013).

86. Barbas, H. & Pandya, D. N. Architecture and frontal cortical connections of the premotor cortex (area 6) in the rhesus monkey. *J. Comp. Neurol.* **256**, 211–228 (1987).

87. He, S., Dum, R. & Strick, P. Topographic organization of corticospinal projections from the frontal lobe: motor areas on the lateral surface of the hemisphere. *J. Neurosci.* **13**, 952–980 (1993).

88. Anderson, S. W., Damasio, A. R. & Damasio, H. TROUBLED LETTERS BUT NOT NUMBERS: DOMAIN SPECIFIC COGNITIVE IMPAIRMENTS FOLLOWING FOCAL DAMAGE IN FRONTAL CORTEX. *Brain* **113**, 749–766 (1990).

89. Longcamp, M., Anton, J.-L., Roth, M. & Velay, J.-L. Visual presentation of single letters activates a premotor area involved in writing. *NeuroImage* **19**, 1492–1500 (2003).

90. James, K. H. & Gauthier, I. Letter processing automatically recruits a sensory–motor brain network. *Neuropsychologia* **44**, 2937–2949 (2006).

91. Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L. & Ungerleider, L. G. Discrete Cortical Regions Associated with Knowledge of Color and Knowledge of Action. *Science* **270**, 102–105 (1995).

92. Grafton, S. T., Fadiga, L., Arbib, M. A. & Rizzolatti, G. Premotor Cortex Activation during Observation and Naming of Familiar Tools. *NeuroImage* **6**, 231–236 (1997).

93. Cross, E. S., Hamilton, A. F. D. C. & Grafton, S. T. Building a motor simulation de novo: Observation of dance by dancers. *NeuroImage* **31**, 1257–1267 (2006).

94. Iacoboni, M. *et al.* Cortical Mechanisms of Human Imitation. *Science* **286**, 2526–2528 (1999).

95. Decety, J. *et al.* Mapping motor representations with positron emission tomography. *Nature* **371**, 600–602 (1994).

96. Rizzolatti, G. *et al.* Functional organization of inferior area 6 in the macaque monkey. (1988).

97. Kakei, S., Hoffman, D. S. & Strick, P. L. Direction of action is represented in the ventral premotor cortex. *Nat. Neurosci.* **4**, 1020–1025 (2001).

98. Umiltà, M. A. *et al.* When pliers become fingers in the monkey motor system. *Proc. Natl. Acad. Sci.* **105**, 2209–2213 (2008).

99. Fagg, A. H. & Arbib, M. A. Modeling parietal–premotor interactions in primate control of grasping. *Neural Netw.* **11**, 1277–1303 (1998).

100. Schaffelhofer, S. & Scherberger, H. Object vision to hand action in macaque parietal, premotor, and motor cortices. *eLife* **5**, e15278 (2016).

101. Hickok, G. Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans. *J. Cogn. Neurosci.* **21**, 1229–1243 (2009).

102. Bonini, L. *et al.* Ventral Premotor and Inferior Parietal Cortices Make Distinct Contribution to Action Organization and Intention Understanding. *Cereb. Cortex* **20**, 1372–1385 (2010).

103. Bonini, L. *et al.* Grasping Neurons of Monkey Parietal and Premotor Cortices Encode Action Goals at Distinct Levels of Abstraction during Complex Action Sequences. *J. Neurosci.* **31**, 5876–5886 (2011).

104. Matelli, M., Camarda, R., Glickstein, M. & Rizzolatti, G. Afferent and efferent projections of the inferior area 6 in the macaque monkey. *J. Comp. Neurol.* **251**, 281–298 (1986).

105. Romo, R., Hernández, A. & Zainos, A. Neuronal Correlates of a Perceptual Decision in Ventral Premotor Cortex. *Neuron* **41**, 165–173 (2004).

106. Mizes, K. G. C., Lindsey, J., Escola, G. S. & Ölveczky, B. P. Dissociating the contributions of sensorimotor striatum to automatic and visually guided motor sequences. *Nat. Neurosci.* **26**, 1791–1804 (2023).

107. Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z. & Rand, M. K. Differential roles of monkey striatum in learning of sequential hand movement: *Exp. Brain Res.* **115**, 1–5 (1997).

108. Matsuzaka, Y., Picard, N. & Strick, P. L. Skill Representation in the Primary Motor Cortex After Long-Term Practice. *J. Neurophysiol.* **97**, 1819–1832 (2007).

109. Russo, A. A. *et al.* Motor Cortex Embeds Muscle-like Commands in an Untangled Population

Response. *Neuron* **97**, 953-966.e8 (2018).

110. Lu, X. & Ashe, J. Anticipatory Activity in Primary Motor Cortex Codes Memorized Movement Sequences. *Neuron* **45**, 967–973 (2005).

111. Yu, A. C. & Margoliash, D. Temporal Hierarchical Control of Singing in Birds. *Science* **273**, 1871–1875 (1996).

112. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).

113. Kawai, R. *et al.* Motor cortex is required for learning but not for executing a motor skill. *Neuron* **86**, 800–812 (2015).

114. Geddes, C. E., Li, H. & Jin, X. Optogenetic Editing Reveals the Hierarchical Organization of Learned Action Sequences. *Cell* **174**, 32-43.e15 (2018).

115. Rueda-Orozco, P. E. & Robbe, D. The striatum multiplexes contextual and kinematic information to constrain motor habits execution. *Nat. Neurosci.* **18**, 453–460 (2015).

116. Kornysheva, K. & Diedrichsen, J. Human premotor areas parse sequences into their spatial and temporal features. *eLife* **3**, e03043 (2014).

117. Ramkumar, P. *et al.* Chunking as the result of an efficiency computation trade-off. *Nat. Commun.* **7**, 12176 (2016).

118. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

119. Baldauf, D., Cui, H. & Andersen, R. A. The Posterior Parietal Cortex Encodes in Parallel Both Goals for Double-Reach Sequences. *J. Neurosci.* **28**, 10081–10089 (2008).

120. Kadmon Harpaz, N., Ungarish, D., Hatsopoulos, N. G. & Flash, T. Movement Decomposition in the Primary Motor Cortex. *Cereb. Cortex* **29**, 1619–1633 (2019).

121. Mushiake, H., Inase, M. & Tanji, J. Neuronal activity in the primate premotor, supplementary, and precentral motor cortex during visually guided and internally determined sequential movements. *J. Neurophysiol.* **66**, 705–718 (1991).

122. Ohbayashi, M., Ohki, K. & Miyashita, Y. Conversion of Working Memory to Motor Sequence in

the Monkey Premotor Cortex. *Science* **301**, 233–236 (2003).

123. Tanji, J. & Shima, K. Role for supplementary motor area cells in planning several movements ahead. *Nature* **371**, 413–416 (1994).

124. Markowitz, J. E. *et al.* The Striatum Organizes 3D Behavior via Moment-to-Moment Action Selection. *Cell* **174**, 44-58.e17 (2018).

125. Lanzarini, F. *et al.* Neuroethology of natural actions in freely moving monkeys. *Science* **387**, 214–220 (2025).

126. Testard, C. *et al.* Neural signatures of natural behaviour in socializing macaques. *Nature* **628**, 381–390 (2024).

127. Voloh, B. *et al.* Hierarchical action encoding in prefrontal cortex of freely moving macaques. *Cell Rep.* **42**, 113091 (2023).

128. Averbeck, B. B., Chafee, M. V., Crowe, D. A. & Georgopoulos, A. P. Neural activity in prefrontal cortex during copying geometrical shapes: I. Single cells encode shape, sequence, and metric parameters. *Exp. Brain Res.* **150**, 127–141 (2003).

129. Schwartz, A. B. Direct Cortical Representation of Drawing. *Science* **265**, 540–542 (1994).

130. Abeles, M. *et al.* Compositionality in neural control: an interdisciplinary study of scribbling movements in primates. *Front. Comput. Neurosci.* **7**, (2013).

131. Lashley, K. S. The Problem of Serial Order in Behavior. (1951).

132. Bernstein, N. A. *Dexterity and Its Development*. x, 460 (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1996).

133. Merel, J., Botvinick, M. & Wayne, G. Hierarchical motor control in mammals and machines. *Nat. Commun.* **10**, 5489 (2019).

134. Cesanek, E., Zhang, Z., Ingram, J. N., Wolpert, D. M. & Flanagan, J. R. Motor memories of object dynamics are categorically organized. *eLife* **10**, e71627 (2021).

135. d'Avella, A., Giese, M., Ivanenko, Y. P., Schack, T. & Flash, T. Editorial: Modularity in motor control: from muscle synergies to cognitive action representation. *Front. Comput. Neurosci.* **9**, (2015).

136. Bizzi, E., Cheung, V. C. K., d'Avella, A., Saltiel, P. & Tresch, M. Combining modules for movement. *Brain Res. Rev.* **57**, 125–133 (2008).

137. Tresch, M. C. & Jarc, A. The case for and against muscle synergies. *Curr. Opin. Neurobiol.* **19**, 601–607 (2009).

138. Overduin, S. A., d'Avella, A., Carmena, J. M. & Bizzi, E. Microstimulation Activates a Handful of Muscle Synergies. *Neuron* **76**, 1071–1077 (2012).

139. Piantadosi, S. T. *et al.* Why concepts are (probably) vectors. *Trends Cogn. Sci.* **28**, 844–856 (2024).

140. Zylberberg, A., Dehaene, S., Roelfsema, P. R. & Sigman, M. The human Turing machine: a neural framework for mental programs. *Trends Cogn. Sci.* S136466131100088X (2011) doi:10.1016/j.tics.2011.05.007.

141. Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M. & Maass, W. Brain computation by assemblies of neurons. *Proc. Natl. Acad. Sci.* **117**, 14464–14472 (2020).

142. Kriete, T., Noelle, D. C., Cohen, J. D. & O'Reilly, R. C. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl. Acad. Sci.* **110**, 16390–16395 (2013).

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

144. Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* **46**, 159–216 (1990).

145. Riveland, R. & Pouget, A. Natural language instructions induce compositional generalization in networks of neurons. *Nat. Neurosci.* **27**, 988–999 (2024).

146. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

147. Xie, Y. *et al.* Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).

148. Russo, A. A. *et al.* Neural Trajectories in the Supplementary Motor Area and Motor Cortex Exhibit Distinct Geometries, Compatible with Different Classes of Computation. *Neuron* **107**,

745-758.e6 (2020).

149. Averbeck, B. B., Sohn, J.-W. & Lee, D. Activity in prefrontal cortex during dynamic selection of action sequences. *Nat. Neurosci.* **9**, 276–282 (2006).

150. Mollazadeh, M. *et al.* Spatiotemporal Variation of Multiple Neurophysiological Signals in the Primary Motor Cortex during Dexterous Reach-to-Grasp Movements. *J. Neurosci.* **31**, 15531–15543 (2011).

151. Saleem, K. S. & Logothetis, N. K. *A Combined MRI and Histology Atlas of the Rhesus Monkey Brain in Stereotaxic Coordinates*. (Academic Press, 2012).

152. Berger, M., Agha, N. S. & Gail, A. Wireless recording from unrestrained monkeys reveals motor goal encoding beyond immediate reach in frontoparietal cortex. *eLife* **9**, e51322 (2020).

153. Lu, M.-T., Preston, J. B. & Strick, P. L. Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. *J. Comp. Neurol.* **341**, 375–392 (1994).

154. Gentilucci, M. *et al.* Functional organization of inferior area 6 in the macaque monkey.

155. Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* **3**, 131–141 (1996).

156. He, S., Dum, R. & Strick, P. Topographic organization of corticospinal projections from the frontal lobe: motor areas on the medial surface of the hemisphere. *J. Neurosci.* **15**, 3284–3306 (1995).

157. Nakamura, K., Sakai, K. & Hikosaka, O. Effects of Local Inactivation of Monkey Medial Frontal Cortex in Learning of Sequential Procedures. *J. Neurophysiol.* **82**, 1063–1068 (1999).

158. Petrides, M. Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *J. Neurosci.* **15**, 359–375 (1995).

159. Mansouri, F. A., Freedman, D. J. & Buckley, M. J. Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* **21**, 595–610 (2020).

160. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).

161. Nieder, A., Freedman, D. J. & Miller, E. K. Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex. *Science* **297**, 1708–1711 (2002).

162. Wilson, F. A. W., Scalaidhe, S. P. Ó. & Goldman-Rakic, P. S. Dissociation of Object and Spatial Processing Domains in Primate Prefrontal Cortex. *Science* **260**, 1955–1958 (1993).

163. Ninokura, Y., Mushiake, H. & Tanji, J. Integration of Temporal Order and Object Information in the Monkey Lateral Prefrontal Cortex. *J. Neurophysiol.* **91**, 555–560 (2004).

164. Tsujimoto, S., Genovesio, A. & Wise, S. P. Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nat. Neurosci.* **13**, 120–126 (2010).

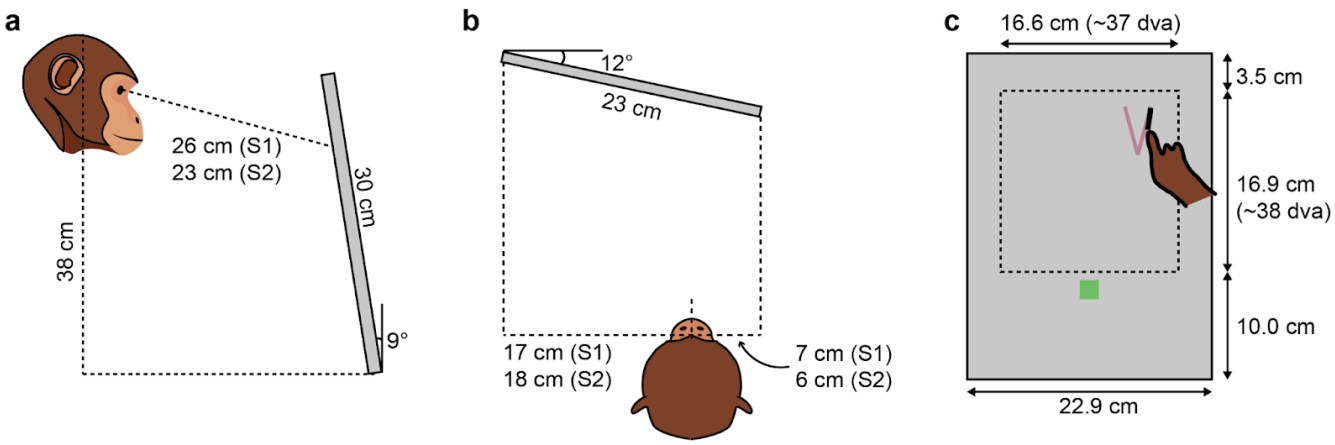
165. Miyamoto, K., Setsuie, R., Osada, T. & Miyashita, Y. Reversible Silencing of the Frontopolar Cortex Selectively Impairs Metacognitive Judgment on Non-experience in Primates. *Neuron* **97**, 980-989.e6 (2018).

166. Hwang, J., Mitz, A. R. & Murray, E. A. NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. *J. Neurosci. Methods* **323**, 13–21 (2019).

167. Dubuisson, M.-P. & Jain, A. K. A modified Hausdorff distance for object matching. in *Proceedings of 12th International Conference on Pattern Recognition* vol. 1 566–568 vol.1 (1994).

168. Pachitariu, M., Sridhar, S., Pennington, J. & Stringer, C. Spike sorting with Kilosort4. *Nat. Methods* **21**, 914–921 (2024).

169. Liu, Y., Brincat, S. L., Miller, E. K. & Hasselmo, M. E. A Geometric Characterization of Population Coding in the Prefrontal Cortex and Hippocampus during a Paired-Associate Learning Task. *J. Cogn. Neurosci.* **32**, 1455–1465 (2020).

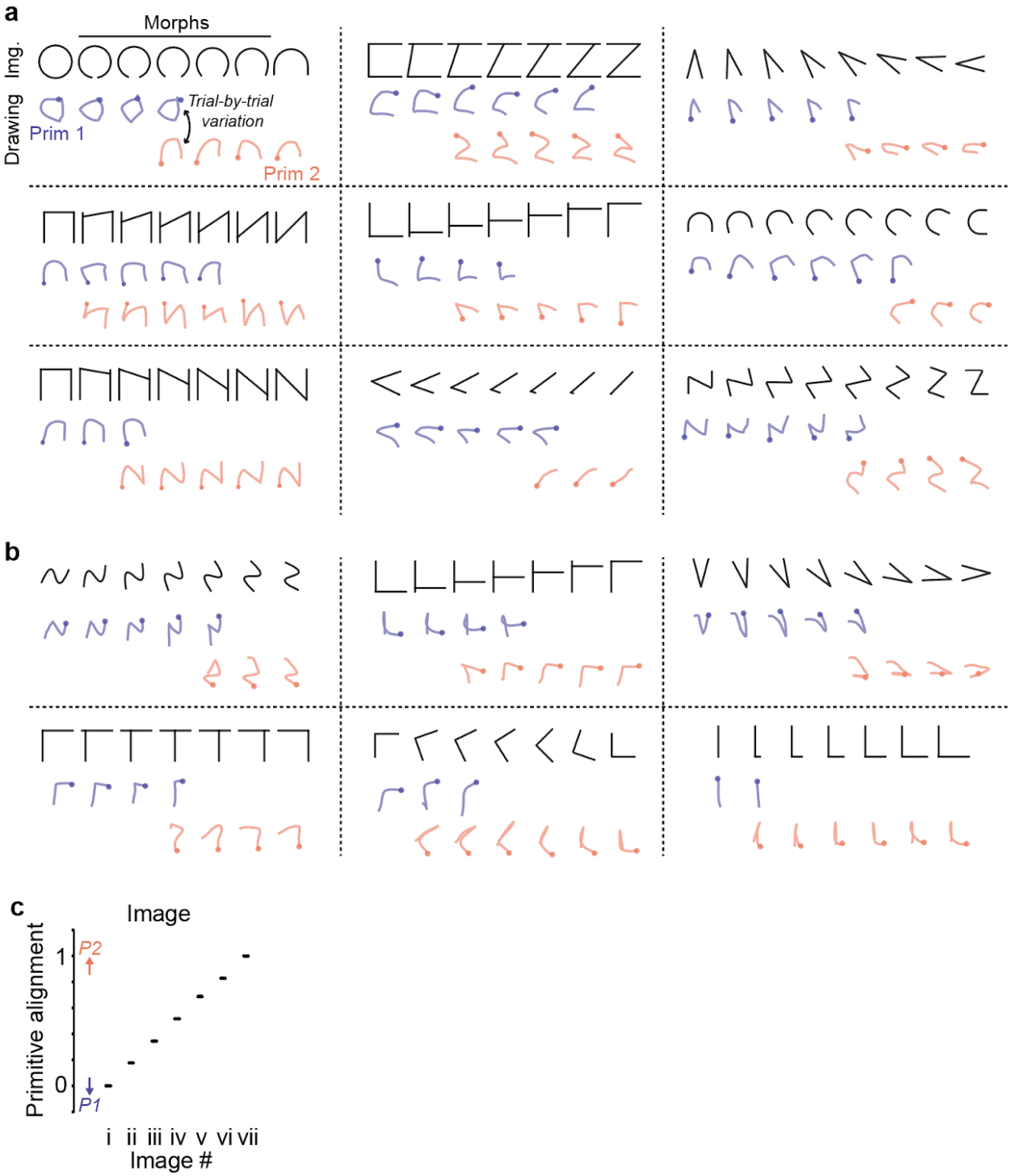


Extended Data Fig. 1. Behavioral task setup.

(a) Schematic of subject relative to screen, profile view. Subjects 1 (S1) and 2 (S2) were positioned at different distances to accommodate their individual anatomies and postures. The screen was slanted slightly to optimize the ability to see and reach to the same part of the screen (the workspace at the top of the screen). Monkey schematic here and in panel b by D. Hanuska.

(b) Schematic of subject position relative to screen, top view. Subjects were positioned to the right, to accommodate reaching to the screen with the hand they used for drawing (left).

(c) Schematic of screen during trial, with component locations and sizes to scale. The finger is tracing over the figure (purple-gray “V”), leaving a black trail of “ink” behind. The thickness and color of the figure and “ink” are to scale. The “done button” is visible (green square), and the subject can press it at any time to report completion. The dashed line indicates the workspace (not visible to the subject).



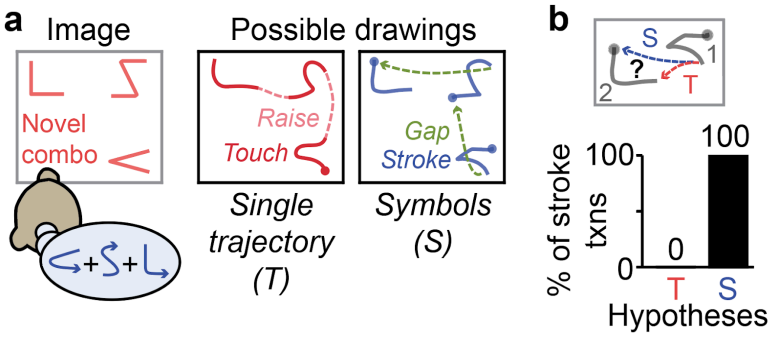
Extended Data Fig. 2. Categorical structure in behavior: more examples and quantification of primitive alignment for image data.

(a) Example single-trial drawings across nine different morph sets, for subject 2. Drawings are colored by whether they reflect use of primitive 1 (blue) or primitive 2 (orange). Images morph between two

well-practiced shapes. The examples here are depicted in a similar manner to the example in **Fig. 2.**

(b) Same as panel a, but for subject 1.

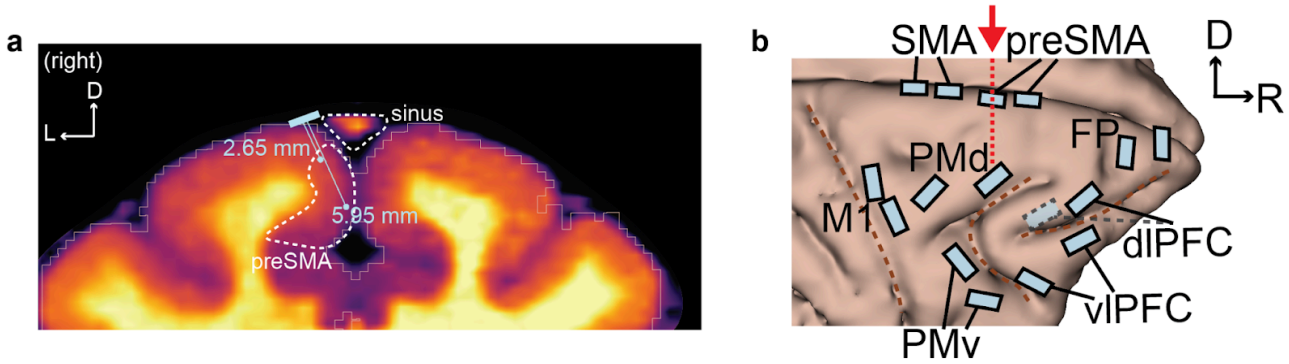
(c) Primitive alignment vs. trial condition for the example experiment in panel **Fig. 2b,** performed on image data (using the image distance), and not on behavioral data as in **Fig. 2i**.



Extended Data Fig. 3. Recombination of stroke primitives into sequences, in "multi-shape" tasks.

(a) Experiment testing for recombination of primitives into sequences, using "multi-shape" tasks. Given images composed of multiple disconnected shapes, two possible drawing responses are shown, consistent with either the "Single trajectory" (T) or "Symbols" (S) hypotheses (see main text).

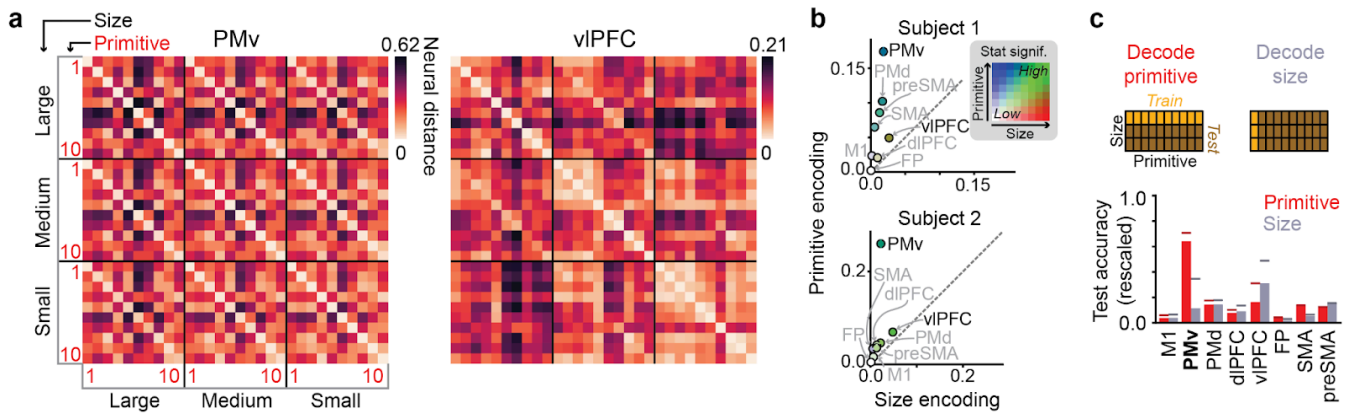
(b) Fraction of stroke-to-stroke transitions in which the second stroke is drawn in a manner consistent with the Single trajectory (T) or Symbols (S) strategies, restricted to transitions where the behavioral predictions of these two strategies differed (bottom bar plot). Top, schematic of an example transition between two strokes labeled 1 and 2. Stroke 2 can be drawn either starting from the top-left, consistent with primitive reuse ("Symbols"), or from the bottom right, which would reflect the taking of the shorter of the two gap distances ("Single trajectory").



Extended Data Fig. 4. Angled implantation of SMA and preSMA arrays to avoid the sinus and target the medial wall.

(a) Coronal MRI section showing the location and angle of a preSMA array relative to the superior sagittal sinus ("sinus"). SMA and preSMA arrays were implanted ~2 mm lateral to the midline, and angled medially, in order to target the medial wall (preSMA), while avoiding the superior sagittal sinus. The array is depicted as a rectangular blue surface and two lines representing the shortest (2.65 mm) and longest (5.95 mm) electrodes. This location and angle is our best estimate based on the stereotactic coordinates and intra-operative photographs. The other two SMA and one preSMA arrays were located and angled similarly. D, dorsal. L, lateral.

(b) Rostral-caudal location of the coronal section in panel a (red arrow and dashed line), overlaid on a brain surface model. D, dorsal. R, rostral.

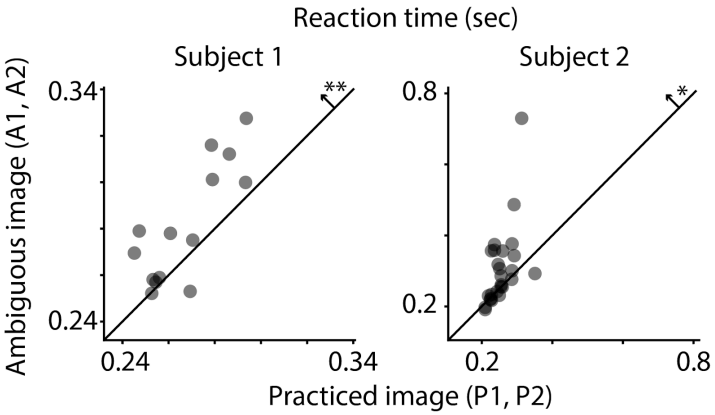


Extended Data Fig. 5. PMv encodes stroke primitives in a manner that is invariant to size.

(a) Heatmap of pairwise neural distances between each unique combination of primitive and size, averaged over the planning epoch (0.05 to 0.6 s relative to image onset), for PMv and vlPFC. Shown are data from a single session for subject 2. N = 17 - 23 trials per primitive-size combination.

(b) Summary of primitive encoding and size encoding across areas and sessions, with each point depicting the encoding scores for a given area. Each point's color denotes statistical significance, in terms of the number of other brain areas that this area beats in pairwise statistical tests of primitive encoding and size encoding (represented in the inset heatmap); see **Fig. 4j** and Methods for details. Each data point was a unique pair of primitive-size conditions (trial-averaged). For testing primitive encoding, these pairs were the same size, but different primitive [N = 313 (subject 1), 253 (subject 2)]. For size encoding, the pairs were the same primitive, but different size [N = 60 (S1), 56 (S2)]. Statistical tests were performed on condition pairs pooled across sessions (3 for S1, 2 for S2).

(c) Across-condition generalization of linear SVM decoders for primitive (red) and size (gray). See **Fig. 4k**.



Extended Data Fig. 6. Slower reaction time for ambiguous images in tasks testing categorical structure.

Average reaction time (between go cue and stroke onset), comparing ambiguous images to practiced images. Each data point represents a single primitive from one morph set, its y-value indicating reaction time when the primitive was drawn in response to an ambiguous image, and its x-value indicating reaction time when it was drawn in response to the practiced image. Each morph set has two primitives and thus contributed two data points. *, $p < 0.05$; **, $p < 0.005$, two-sided Wilcoxon signed-rank test. N = 14 primitives (7 morph sets) for subject 1; N = 26 primitives (13 morph sets) for subject 2.

Supplementary Video 1. Behavior in character task (example character 1, subject 1). This depicts the trial in **Fig. 2l**, 4th column from the right.

Supplementary Video 2. Behavior in character task (example character 1, subject 2). This depicts the trial in **Fig. 2l**, 4th column from the right.

Supplementary Video 3. Behavior in character task (example character 2, subject 1). This depicts the trial in **Fig. 2l**, 7th column from the left.

Supplementary Video 4. Behavior in character task (example character 2, subject 2). This depicts the trial in **Fig. 2l**, 7th column from the left.

Supplementary Video 5. Behavior in character task (example character 3, subject 1). This depicts the trial in **Fig. 2l**, 3rd column from the left.

Supplementary Video 6. Behavior in character task (example character 3, subject 2). This depicts the trial in **Fig. 2l**, 3rd column from the left.

Supplementary Video 7. Behavior in character task (example character 4, subject 1). This depicts the trial in **Fig. 2l**, 2nd column from the left.

Supplementary Video 8. Behavior in character task (example character 4, subject 2). This depicts the trial in **Fig. 2l**, 2nd column from the left.

Supplementary Video 9. Behavior in character task (example character 5, subject 1). This depicts the trial in **Fig. 2l**, 1st column from the left.

Supplementary Video 10. Behavior in character task (example character 5, subject 2). This depicts the trial in **Fig. 2l**, 1st column from the left.