Neuron

Neural knowledge assembly in humans and neural networks

Highlights

- Humans can rapidly reassemble knowledge into novel relational forms
- Rapid knowledge assembly is problematic for current theories of how we learn
- Vanilla neural networks cannot rapidly assemble knowledge as seen in humans
- By adding a certainty parameter, we can recover this ability in neural networks

Authors

Stephanie Nelli, Lukas Braun, Tsvetomira Dumbalska, Andrew Saxe, Christopher Summerfield

Correspondence

nelli@oxy.edu (S.N.), christopher.summerfield@psy.ox.ac.uk (C.S.)

In brief

Our work addresses a perplexing question: how is human understanding of the world dramatically changed by a single new piece of information? Interestingly, by dialing up and down a "certainty" parameter in a neural network model, we can capture the range of successes and failures shown in human participants.

Nelli et al., 2023, Neuron *111*, 1–13
May 3, 2023
https://doi.org/10.1016/j.neuron.2023.02.014

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Neural knowledge assembly in humans and neural networks

Stephanie Nelli,[1,2,5,6,*] Lukas Braun,[2,5] Tsvetomira Dumbalska,[2] Andrew Saxe,[2,3,4] and Christopher Summerfield[2,*]
[1]Department of Cognitive Science, Occidental College, Los Angeles, CA 90041, USA
[2]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GC, UK
[3]Gatsby Unit & Sainsbury Wellcome Centre, University College London, London W1T 4JG, UK
[4]CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, ON M5G 1M1, Canada
[5]These authors contributed equally
[6]Lead contact
*Correspondence: nelli@oxy.edu (S.N.), christopher.summerfield@psy.ox.ac.uk (C.S.)
https://doi.org/10.1016/j.neuron.2023.02.014

SUMMARY

Human understanding of the world can change rapidly when new information comes to light, such as when a plot twist occurs in a work of fiction. This flexible "knowledge assembly" requires few-shot reorganization of neural codes for relations among objects and events. However, existing computational theories are largely silent about how this could occur. Here, participants learned a transitive ordering among novel objects within two distinct contexts before exposure to new knowledge that revealed how they were linked. Blood-oxygen-level-dependent (BOLD) signals in dorsal frontoparietal cortical areas revealed that objects were rapidly and dramatically rearranged on the neural manifold after minimal exposure to linking information. We then adapt online stochastic gradient descent to permit similar rapid knowledge assembly in a neural network model.

INTRODUCTION

To make sense of the world, we need to know how objects, people, and places relate to one another. Understanding how relational knowledge is acquired, organized, and used for inference has become a frontier topic in both neuroscience and machine learning research.[1–7] Since Tolman, neuroscientists have proposed that when ensembles of states are repeatedly co-experienced, they are mentally organized into cognitive maps whose geometry mirrors the external environment.[8–12] Recently, brain imaging has been used to study how representations change over the course of learning, with a focus on the medial temporal lobe (MTL).[2,13] After learning, the associative distance between objects or locations (i.e., how related they are in space or time) has been found to covary with similarity (or dissimilarity) among neural coding patterns. Some neural signals, especially in MTL structures, may even explicitly encode relational information about how space is structured or how knowledge hierarchies are organized.[14–19]

A striking aspect of cognition is that these knowledge structures can be rapidly reconfigured when a single (or just a few) samples of new information become available.[20] For example, a plot twist in a film might require the viewer to rapidly and dramatically reconsider a protagonist's motives, or an etymological insight might allow a reader to suddenly understand the connection between two words. Here, we dub this process "knowledge assembly" because it requires existing knowledge to be rapidly (re-)assembled on the basis of minimal new information. How do brains rapidly update knowledge structures, selectively adjusting certain relations while keeping others intact? In machine learning research,[21,22] solutions to the general problem of building rich conceptual knowledge structures include graph-based architectures,[23] modular networks,[24] probabilistic programs,[25] and deep generative models.[26] However, while these artificial tools can allow for expressive mental representation or powerful inference, they tend to learn slowly and require dense supervision, making them implausible models of knowledge assembly and limiting their scope as theories of biological learning.

How, then, does rapid, "few-shot" knowledge assembly occur in humans? Here, we designed a task to test this, based on a paradigm in which participants first learn about two independent ordered sets and are then provided with a small number of "list linking" trials that reveal how the two sets are related. We study human behavior and brain activity on this paradigm and offer a theory of how knowledge can be rapidly assembled using a version of an artificial neural network model, providing a computational account of the behavioral and neural results observed in humans.

RESULTS

Human participants (n = 34) performed a computerized task that involved making decisions about novel visual objects. Each object i was randomly assigned a ground truth rank (i_1–i_{12}) on the nonsense dimension of "brispiness" (Figure 1A; where i_1 is the most "brispy" and i_{12} is the least). During initial training (*train short*), the 12 objects were split into two distinct sets (items i_1–i_6

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

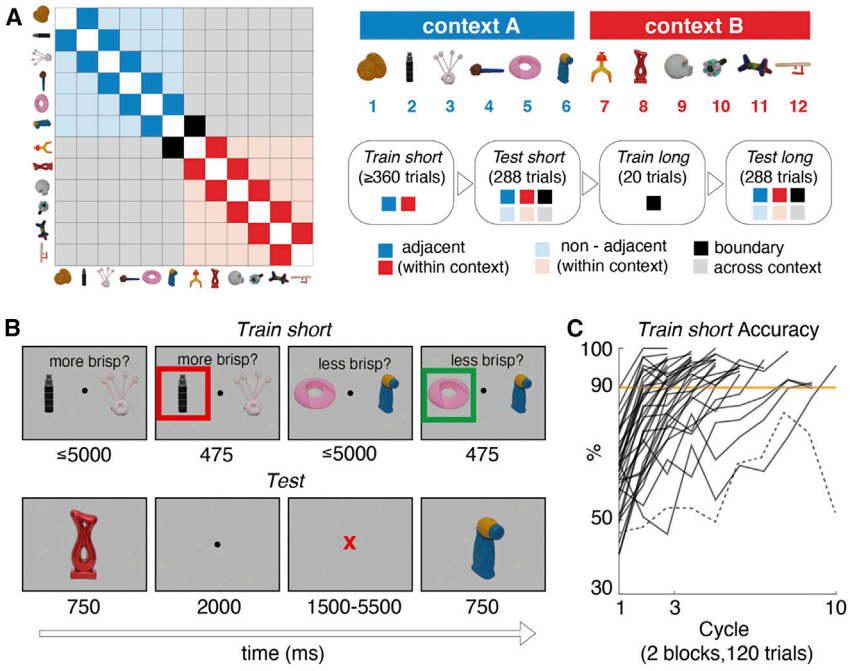

Figure 1. Task and design

(A) Left: matrix illustrating training and testing conditions for an example set of objects ordered by rank (on the x and y axes). Each entry indicates a pair of stimuli defined by their row and column. Colors signal when the pair was trained or tested. Dark blue and red squares are within-context pairs, shown during *train short*. In addition to these, lighter blue and red squares (non-adjacent) and gray (untrained) are within-context pairs not seen during *train short*. The black squares are the pairs shown during boundary training (*train long*). All pairs are tested during *test short* and *test long*. Right: schematic of experimental sequence and legend. Although we use the same set of objects in these figures for display purposes, note that each participant viewed a unique, randomly sampled set of novel objects. The colored squares refer to the pairs trained or tested in each phase, using color conventions from the leftmost panel.

(B) Example trial sequence during training (upper) and test (lower). Numbers below each example screen show the frame duration in ms.

(C) Percentage accuracy during training for each individual (black lines). Stopping criterion is shown as an orange line. The excluded participant is shown as a dashed trace. A training "cycle" (x axis) consists of two blocks (one for each set of items).

and i_7–i_{12}) and presented in alternating blocks of trials (contexts; see STAR Methods). Participants only made comparisons within each context and were asked to indicate with a button press which of two objects with adjacent rank (e.g., i_3 and i_4) was more (or less) brispy, receiving fully informative feedback (Figure 1B, upper panel). Note that this training regime allowed participants to infer ranks within a set (i.e., within i_1–i_6 and i_7–i_{12}) but betrayed no information about the ground truth relation between the two sets (e.g., $i_2 < i_9$). Participants were trained on adjacent relations to a predetermined criterion, with final training accuracy reaching 95.6% ± 2.9% (mean ± SD; Figure 1C; see STAR Methods). The use of novel objects[27] and a nonword label was designed to minimize participants' tendency to use prior information when solving the task.

After training, participants entered the scanner and performed a first test phase (*test short*) in which they viewed objects one by one that were sampled randomly from across the full range (i_1–i_{12}). The task required them to report the brispiness of each object relative to its predecessor (i.e., a 1-back task) with a button press (Figure 1B, lower panel). Therefore, the test phase involved comparisons of trained (adjacent) pairs within context (e.g., i_3 and i_4), untrained (non-adjacent) pairs within context (e.g., i_3 and i_6), and untrained pairs across contexts (e.g., i_3 and i_{10}). Importantly, participants did not receive trialwise feedback on their choices during the test phase.

Our first question was whether humans generalized knowledge about object brispiness both within and between contexts during the *test short* phase. We collapse across the two contexts, as there was no difference in either reaction times (RTs) or accuracy (both p values > 0.3). Participants performed not only above chance both on adjacent pairs on which they had been trained (e.g., i_3 and i_4 or i_9 and i_{10}) (mean accuracy = 86.0 ± 10.4, t test against 50%, $t_{33} = 20.5$, $p < 0.001$) but also on untrained, non-adjacent pairs for which transitive inference was required (e.g., *i*3 and *i*6 or *i*7 and *i*10) (Figure 2A) (mean accuracy 96.7 ± 19.2, $t_{33} = 83.7$, $p < 0.001$). In fact, participants were faster and more accurate for comparisons between non-adjacent than adjacent items (Figure 2B) (accuracy: $t_{33} = 7.8$, $p < 0.001$; RT: $t_{33} = 11.7$, $p < 0.001$). This was driven by an increase in accuracy (and decrease in RT) with growing distance between comparanda (Figures 2A and 2B, right panels) (accuracy: $\beta = 3.4\%$ per rank; $t_{33} = 7.7$, $p < 0.001$. RT: $\beta = 72$ ms faster per rank; $t_{33} = -7.8$, $p < 0.001$; βs obtained with a linear regression model), known as the "symbolic distance" effect.[28,29]

Moreover, behavior also indicated how participants compared ranks between contexts before ground truth was revealed. For example, they tended to infer $i_7 > i_2$ and $i_4 > i_{11}$ (Figure 2A). This implies a natural tendency to match rank orderings between contexts (e.g., that the 3rd item in one set was ranked higher than the 4th in the other) in the absence of information about how objects were related. In line with this, we quantified between-context accuracy relative to an agent that generalizes ranks perfectly between contexts (see STAR Methods and Figure S1A) and found that between-context accuracy was above chance for "adjacent" (75.9 ± 20.9 mean ± SD; $t_{33} = 7.3$, $p < 0.001$) and "non-adjacent" (91.5 ± 9.3 mean ± SD; $t_{33} = 14.2$, $p < 0.001$) trials.

We also observed a *between-context* symbolic distance effect in RTs (Figure 2B) (accuracy: $\beta = 4.9\%$ per rank, $t_{33} = 7.5$, $p < 0.001$. RT: $\beta = 75$ ms per rank; $t_{33} = 7.8$, $p < 0.001$). Participants were slowest when comparing items with equivalent rank across contexts (e.g., i_2 and i_8), responding more slowly than for adjacent items both within ($t_{33} = 3.23$, $p < 0.004$) and between ($t_{33} = 3.66$, $p < 0.001$) contexts. Overall, these results are

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

CellPress
OPEN ACCESS

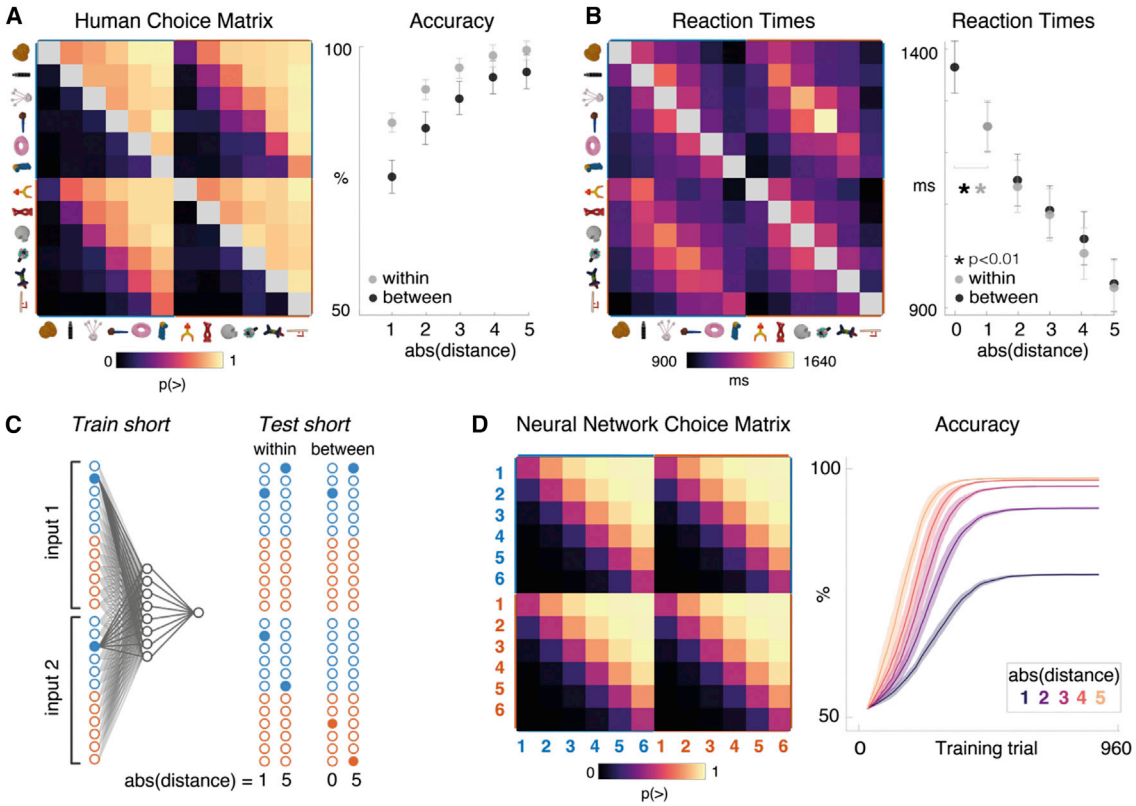

Figure 2. Behavior in humans and neural networks

(A) Left panel: human choice matrix. The color of each entry indicates the probability of responding "greater than" during *test short* for the pair of items defined by the row and column. Color scale is shown below the plot. Object identities are shown for illustration only (and were in fact resampled for each participant). Right panel: accuracy as a function of symbolic distance, shown separately for within-context (e.g., i_3 and i_5; gray dots) and between-context (e.g., i_3 and i_9; black dots) judgments. For between-context judgments, accuracy data are with respect to a ground truth in which ranks are perfectly generalized across contexts (e.g., they infer that $i_2 > i_9$). Errors bars are SEM.

(B) Equivalent data for reaction times. Note that a symbolic distance of zero was possible across contexts (e.g., i_2 vs. i_8) for which there was no "correct" answer, but an RT was measurable. p value indicates significance from paired t tests of RT values.

(C) Left panel: neural network architecture and training scheme. Input nodes are colored red and blue to denote the relevant context. Filled blue dots illustrate an example training trial in which objects i_2 and i_3 are shown. Right panel: example test trials both within and across context, with the symbolic distance signaled below.

(D) Left panel: choice matrix for the neural network, in the same format as (A). Right panel: learning curves (showing accuracy over training epochs) for the neural network, shown separately for trials with different levels of symbolic distance. Shading is 1 SEM over network replicants.

consistent with previous findings in both humans and monkeys and have been taken to imply that participants automatically infer and represent the ordinal position of each item in the set.[30]

Next, to understand the computational underpinnings of this behavior, we trained a neural network to solve an equivalent transitive inference problem. The network had a two-layer feedforward architecture with symmetric input weights and was trained in a supervised fashion using online stochastic gradient descent (SGD) (Figure 2C). For this modeling exercise, we replaced the unrelated object images seen by participants with orthogonal (one-hot) vector inputs. On each trial the network received two inputs, denoting the images shown on the right and left of the screen, and (just like participants) was required to output whether one was "more" or "less." At the point at which we terminated training (960 total trials; see STAR Methods), the network reached an average test accuracy of 97.74% on unseen comparisons (between non-adjacent pairs) and 79.04% for adjacent (trained) pairs (Figure 2D, right panel). Choice matrices for the humans and neural networks were highly correlated ($r = 0.98$, $p < 0.001$ for averaged choice matrices; single participants $r = 0.82 \pm 0.13$ mean $\pm$ SD; all p values < 0.001), and the network showed a qualitatively identical pattern of generalization within and between contexts, such that accuracy grew with rank distance (Figure 2D, left panel).

After training, we examined neural geometry in the neural network by probing it with each (single) item i_1–i_{12} in turn and calculating a representational dissimilarity matrix (RDM) from resultant hidden layer activations (Figure 3A, top row). We then used multidimensional scaling (MDS) to visualize the similarity structure in just two dimensions (Figure 3A, bottom row). As training progressed, the network learned to represent the items in order of brispiness along two overlapping neural lines. We know from recent work that a low-dimensional solution is only guaranteed when the hidden layer weights are initialized from

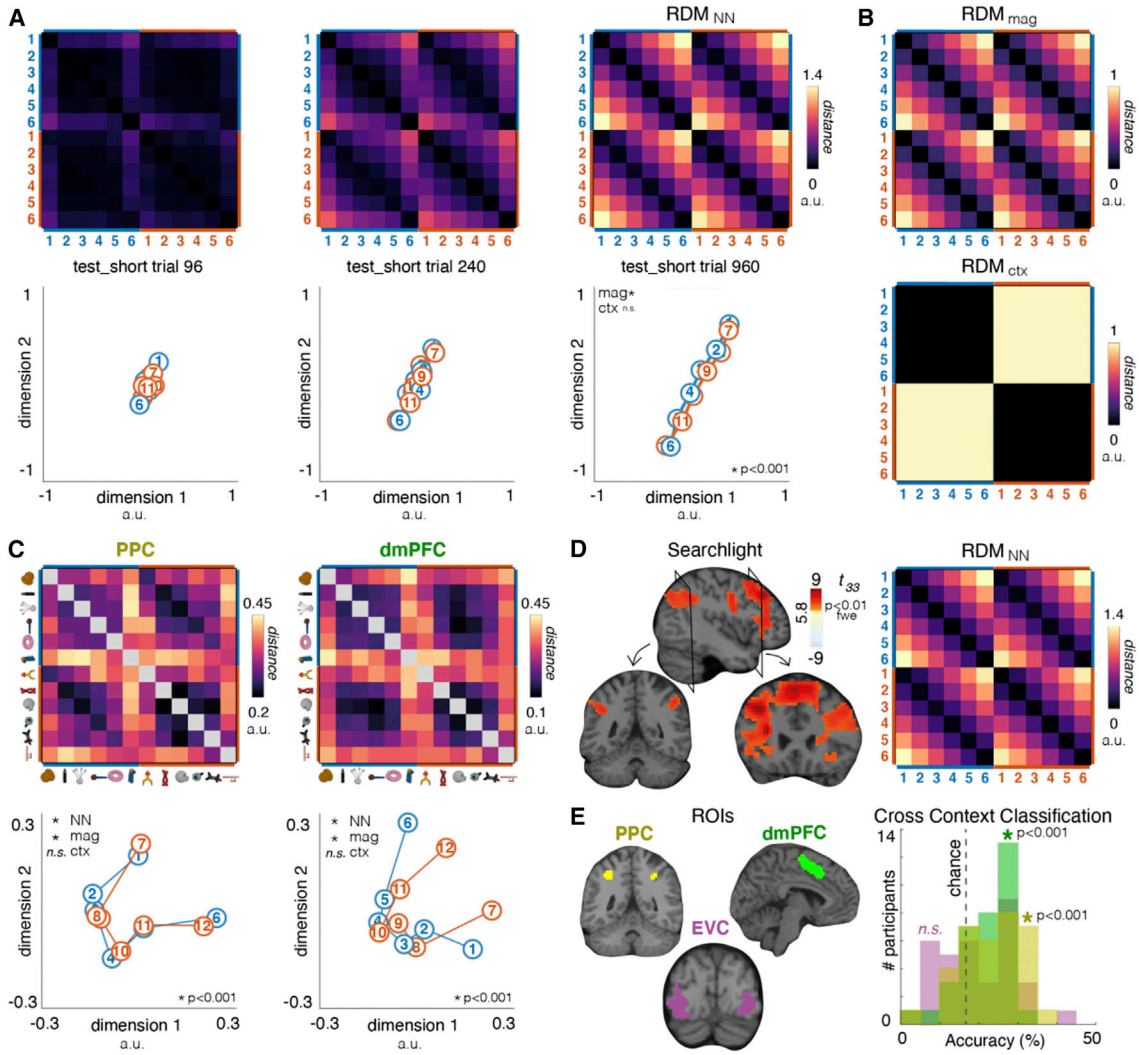

Figure 3. Data from artificial networks and human BOLD signals

(A) Upper panels: RDM for the neural network. Each entry shows the distance between hidden unit activations evoked by a pair of stimuli, for three example time points during training. Lower panels: MDS plot in 2D corresponding to the RDM above. Each circle is a stimulus, colored by its context. Distances between circles conserve similarities in the RDM. Note the emergence of two overlapping lines. p values indicate significance from t tests of Pearson correlation values against zero.

(B) Model RDMs for magnitude (assumes linear spacing between ranks) and context (assumes a fixed distance between contexts).

(C) Upper panels: neural data RDMs from patterns of BOLD in the PPC (left) and dmPFC (right) regions of interest (ROIs). Lower panels: 2D MDS on BOLD data. Red and blue lines denote the two contexts; numbered circles denote items, with their rank signaled by the inset number. p values indicate significance from t tests of Pearson correlation values against zero.

(D) Voxels correlating reliably with the terminal RDM from the neural network (RDMNN, see right panel) rendered onto sagittal (upper) and coronal (lower) slices of a standardized brain, at a threshold of FWE $p < 0.01$.

(E) Left panel: regions of interest (ROIs) in the posterior parietal cortex (PPC, yellow), dorsomedial prefrontal cortex (dmPFC, green), and a control region in the visual cortex (EVC, purple). Right panel: frequencies of classification accuracies over participants for support vector machines (SVMs) trained to distinguish item ranks in one context after training on the other. Three histograms are overlaid, one for each ROI; colors correspond to those in the left panel. Dashed line shows chance (16.6%). Significance values in (C) and (E) correspond to FWE multiple comparison correction (2 ROIs). p values indicate significance using paired t tests of classification values against chance.

very small values, sometimes known as the "rich" training regime (see STAR Methods).[31] After training in the rich regime, the Pearson's correlation between the data RDM from the hidden layer of the neural network (RDM_{NN}) and an idealized distance matrix for parallel (and overlapping) lines (RDM_{mag_short}; Figure 3B, top panel) was > 0.99 for all networks trained ($p < 0.001$ for each of 20 networks). However, we observed no correlation between RDMNN and a model RDM coding for the distance between contexts (RDM_{ctx}; Figure 3B, bottom panel) (Pearson $r \leq 0.1$, $p > 0.4$ for all cases), consistent with the observation that the magnitude lines were not just adjacent but fully overlapping by the end of training (Figure 3A).

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

Next, we compared the representational geometry observed in the neural network to that recorded in BOLD signals while human participants judged the brispiness of successive items in the *test short* phase. We initially focus on regions of interest (ROIs) derived from an independent task in which participants judged the magnitude of Arabic digits, localized to the posterior parietal cortex (PPC) and dorsomedial prefrontal cortex (dmPFC; see Figure S1C) and later show the involvement of a larger frontoparietal network using a whole-brain searchlight approach. In both ROIs, we saw a strong correlation between the neural data RDM and RDM_{mag_short} (Figure 3C) (PPC: $t_{33} = 4.2$; dmPFC: $t_{33} = 5.7$; significant at familywise error [FWE] correction level of $p < 0.001$) but no effect of RDM_{ctx} ($t_{33} < 1$, $p > 0.65$ for both regions). This echoes the data from the hidden layer of the neural network (see Figure 3A), and, accordingly, we observed a significant correlation with RDM_{NN} in both regions (PPC: $t_{33} = 5.3$; dmPFC: $t_{33} = 7.2$, both $p < 0.001$). These effects all held when we defined similarity across (rather than within) scanner runs, which each corresponded to one block of trials (see STAR Methods), using a cross-validated RSA approach (RDM_{mag_short}: PPC: $t_{33} = 4.8$, $p < 0.001$; dmPFC: $t_{33} = 4.7$, $p < 0.001$; RDM_{ctx}: PPC: $t_{33} = 0.3$, $p = 0.78$; dmPFC: $t_{33} = 0.4$, $p = 0.69$; RDM_{NN}: PPC: $t_{33} = 5.4$, $p < 0.001$; dmPFC: $t_{33} = 6.0$, $p < 0.001$; see Figure S1D). Finally, these results still held when we included RTs as a nuisance covariate (RDM_{mag_short}: PPC: $t_{33} = 7.1$, $p < 0.001$; dmPFC $t_{33} = 7.5$, $p < 0.001$; competitive regression with average RTs included in design matrix alongside RDM_{mag_short}). This suggests that it is unlikely that the observed neural geometry is driven by differences in choice latencies between stimuli (see also Figure S1E for further analysis).

We visualized the neural geometry of the BOLD signals in both regions after reducing to two dimensions with MDS. In both ROIs, this yielded overlapping neural lines that reflected the rank-order of the novel objects (Figure 3C, bottom row). Restricting our analysis to distances between consecutive objects, we found neural distances involving the end anchors (e.g., i_1 and i_6) tended to be larger than those involving intermediate ranks ($t_{33} = 4.8$, $p < 0.001$; $t_{33} = 4.4$, $p < 0.001$ in PPC and dmPFC, respectively). Unlike in the neural network, manifolds (number lines) obtained from BOLD data were curved. We note that the curvature of these representational manifolds around their midpoint yields approximately orthogonal axes for rank and uncertainty, and that this phenomenon has been previously observed in scalp EEG recordings[32] and in multi-unit activity from lateral intraparietal cortex (area LIP) of the macaque.[33]

This similarity between representations in the neural network and human BOLD was confirmed by a whole-brain searchlight approach for which we report only effects that pass an FWE correction level of $p < 0.01$ with cluster size >10 voxels. This approach revealed a frontoparietal network in which multivoxel patterns resembled those for the trained neural network (RDM_{NN}; Figure 3D), with peaks in dmPFC (−33 −3 33; $t_{33} = 9.3$, $p_{uncorr} < 0.001$) and inferior parietal lobe (right: 51 −45 45; $t_{33} = 6.94$, $p_{uncorr} < 0.001$; left: −39 −51 36; $t_{33} = 6.93$, $p_{uncorr} < 0.001$). As expected, this was driven by an explicit representation of magnitude distance, as correlations with RDM_{mag_short} (Figure 3B, top panel) peaked in the same regions (dmPFC: −33 −3 33; $t_{33} = 9.4$, $p < 0.001$; inferior parietal lobe right: 48 −45 42; $t_{33} = 7.01$, $p_{uncorr} < 0.001$; left: −39 −51 36; $t_{33} = 7.15$, $p_{uncorr} < 0.001$). Notably, we did not observe an effect of RDM_{ctx} (Figure 3B, bottom panel) (no clusters survived FWE correction, $t_{33} < 2.65$, $p_{uncorr} > 0.012$), indicating that neural representations for similarly ranked items within each of the two contexts were effectively superimposed, as in the neural network.

This representational format, whereby ranked items are represented on overlapping manifolds, lends itself to generalization across contexts, i.e., between items with distinct identity but equivalent brispiness.[34,35] To test this, we trained a support vector machine (SVM) on binary classifications among ranks for context A and evaluated it on the (physically dissimilar) objects from context B. We found above-chance classification in PPC and dmPFC (Figure 3E) ($t_{33} = 4.39$, $p < 0.001$; $t_{33} = 4.01$, $p < 0.001$), but not in an extrastriate visual cortex (EVC) ROI that also showed significant activation during the independent localizer ($t_{33} = 1.75$, $p > 0.08$). These analyses not only cross-validated across runs but also counterbalanced response contingencies and so are unlikely to be due to any spurious effect of motor control. Together, these results show that neural patterns indexed a concept of brispiness divorced from the physical properties of the objects themselves.

Next, we turned to our central question of how neural representations are reconfigured following a single piece of new information about the overall knowledge structure. After *test short*, participants performed a brief "boundary training" (*train long*) phase in the scanner in which they learned that object i_7 (the most brispy object in context B) was less brispy then object i_6 (the least brispy object in context A). This information was acquired over just 20 trials in which participants repeatedly judged whether item i_6 or i_7 was more or less brispy. Following this boundary training, participants performed a new phase *test long* which was identical in every respect to *test short*. Our main question was whether and how the boundary training reshaped both behavior and neural coding for the full set of objects. The average choice and RT matrices observed during *test long* are shown in Figure 4A. As can be seen, on aggregate participants used knowledge of relations between items i_6 and i_7 to correctly infer that all objects lay on a single long axis of brispiness (ranked 1–12). We confirmed this in two ways. First, unlike in *test short*, items in context A (i_1–i_6) were mostly ranked as more brispy than items in context B (i_7–i_{12}), and the symbolic distance effect now spanned the whole range of items 1–12 (with a "dip" near the boundary between contexts; Figure 4B, left panel) (accuracy: $\beta = 2.1\%$ per rank; $t_{33} = 7.8$, $p < 0.001$. RT: $\beta = -29$ ms per rank; $t_{33} = -10.4$, $p < 0.001$). Next, we directly quantified the full pattern of responses seen in Figure 4A by constructing idealized ground truth choice and RT matrices (Figure S1A). These matrices reflected the assumption that the items either lay on two parallel short axes (as most participants inferred in *test short*) or a single long axis (as was correct in *test long*). Fitting these to human behavioral matrices using competitive regressions, we found that while the long axis matrix fits the *test long* behavioral data (choice: $t_{33} = 9.2$, $p < 0.001$; RTs: $t_{33} = 7.9$, $p < 0.001$) there remained a strong residual fit to the short axis choice and RT patterns (choice: $t_{33} = 5.0$, $p < 0.001$; RTs: $t_{33} = 3.5$, $p < 0.01$).

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

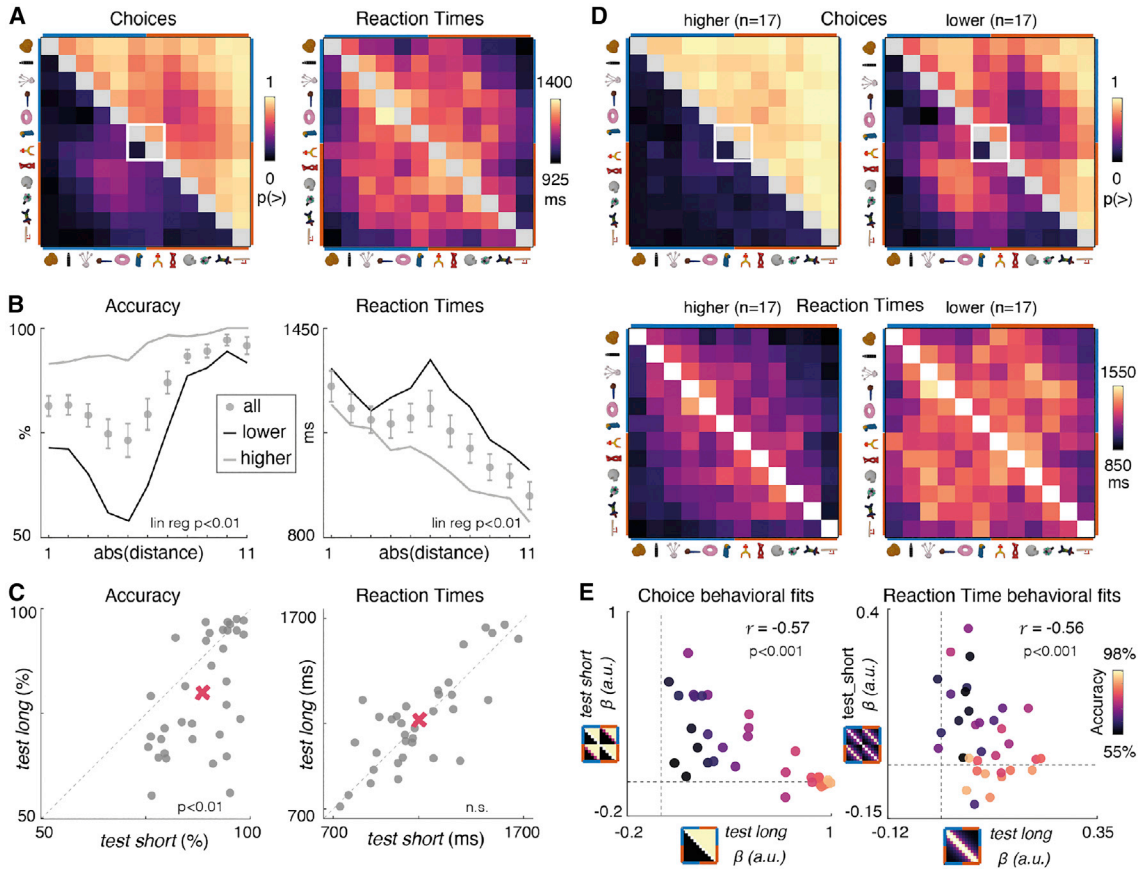

Figure 4. *Test long* behavior

(A) Right panel: choice matrices from human participants after boundary training. Format as for Figure 2A. The white box highlights the two items viewed during boundary training. Note that, on average, choices respect the ground truth rank ($i_1 - i_{12}$). Left panel: same for RTs.

(B) Mean accuracy and response times as a function of ground truth symbolic distance, now defined in the long axis space ($i_1 - i_{12}$). p values indicate the significance of the symbolic distance effect as measured with a linear regression model, lines show means within the "lower" and "higher" performing participant partitions.

(C) Accuracy and RT for each participant (gray dots) in *test long* (y axis) and *test short* (x axis). Diagonal dashed line is the identity line. Red cross is the mean in each condition.

(D) Choice matrices (upper panels) and RTs (lower panels) separately for the two groups. The lower-performing group exhibits choice matrices that resemble those observed after short axis training, as if they failed to update relational knowledge after boundary training.

(E) Left panel: regression coefficients (βs) obtained by fitting idealized the *test short* choice matrix to its corresponding human choice matrix, plotted against coefficients obtained by fitting the idealized *test long* matrix against human choices in *test long*, and the dashed lines indicate zero for either measure. Each dot is a participant colored by their accuracy. Right panel: the same plot obtained regressing idealized RT against human RT matrices. P values indicate significance using Pearson correlation between regression coefficients.

There was substantial variability in performance among participants on *test long*, and median accuracy dropped to 79.8% compared with 88.3% in *test short* (Figure 4C, left panel). As average RTs did not differ between *test short* (1,153 ± 41 ms) and *test long* (1,166 ± 43 ms) (t_{33} = 0.49, p = 0.63), this difference was probably not attributable to a decrement in attention between the two conditions (Figure 4C, right panel). Instead, we reasoned that some participants might have failed to restructure their knowledge of the transitive series, retaining the belief that the two sets were still independent and treating the relative brispiness of item $i_6 < i_7$ as an exception. Indeed, participants who performed more poorly (defined by a median split; Figure 4D, right panels) behaved as if they were still in *test short* (Figure 4E, left panel), whereas those who performed better generalized the few-shot information to correctly infer the rank of all other items (Figure 4D, left panels). Moreover, there was a negative correlation across the cohort between the extent to which participant choices ($r = -0.57$) and RTs ($r = -0.56$) were captured by the idealized *test short* and *test long* matrices, implying that participants that performed poorly on *test long* continued to exhibit behavior that was optimized for *test short* (Figure 4E). We ruled out the possibility that these participants simply failed to learn from the boundary training phase, as they reported the newly trained object relation ($i_6 < i_7$) on 86% ± 15% of *test long* trials, compared with 5.8% ± 19% in *test short* (mean ± SD; t_{33} = 18.0, p < 0.001;

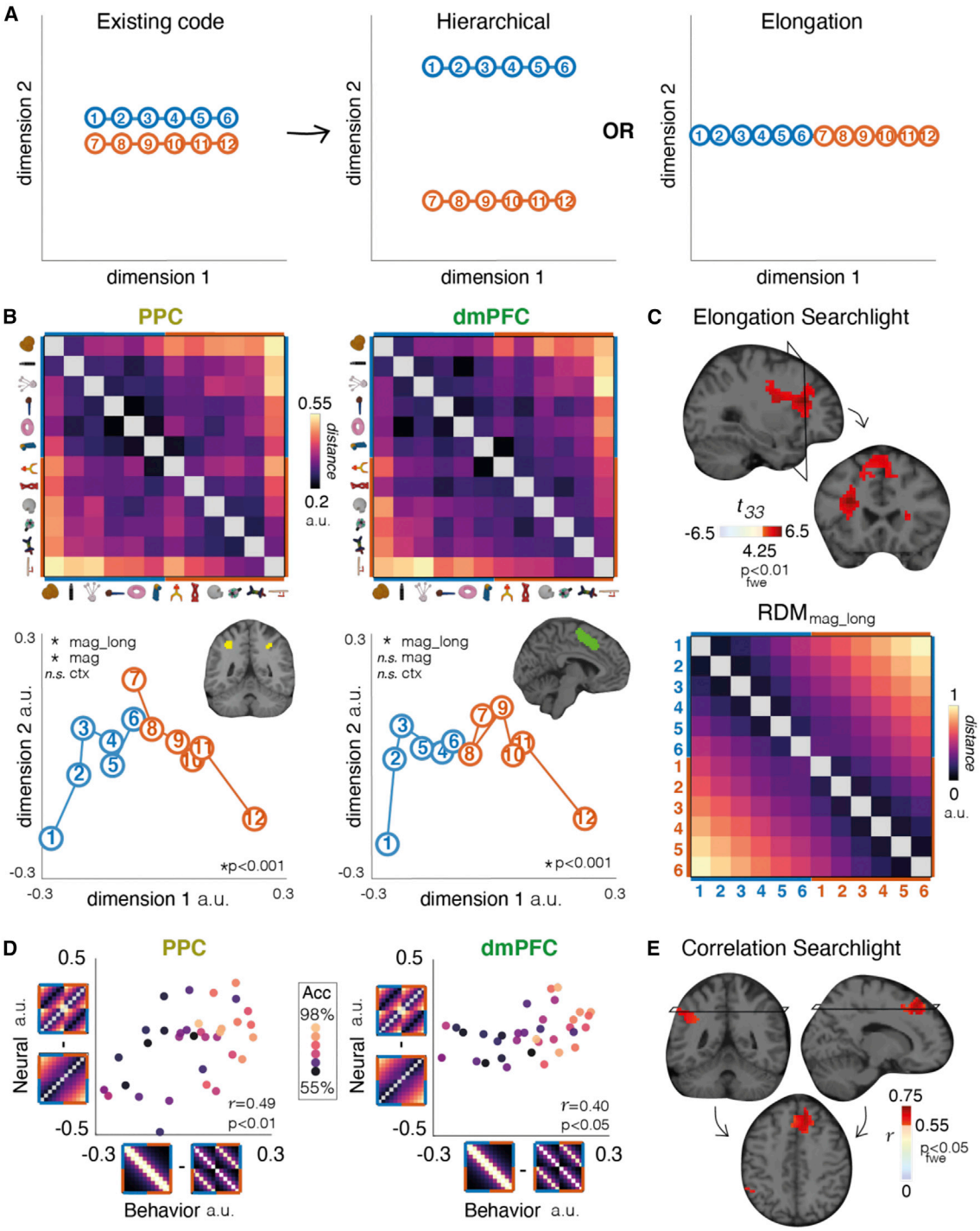

Figure 5. Neural data from *test long*

(A) Schematic illustration of hypotheses about how the extant neural code (after *test short*, left panel) might be transformed in *test long*. The hierarchical hypothesis (middle panel) proposes that magnitude and context are represented on factorized (orthogonal) neural axes. Under the elongation hypothesis (right panel), the items are rearranged on a one-dimensional neural manifold (or magnitude line).

(B) Neural RDMs in the PPC and dMPFC after *test long* (upper panels), and MDS projection of each item in the two contexts (red and blue dots; lower panels), with PPC (left) and dmPFC (right) ROIs inset. p values in all panels indicate significance from t tests of Pearson correlation values against zero.

(C) Model RDM for magnitude in *test long*, and regions correlating with this RDM in a searchlight analysis, rendered onto sagittal and coronal slices of a template brain at a threshold of FWE $p < 0.01$.

(legend continued on next page)

see Figure 4D). Thus, although average participant choices suggested knowledge of a long axis, there was a sizable cohort that only partially integrated the new relation into their knowledge structure.

Next, we turned to the geometry of neural representations in BOLD during the *test long* phase. We considered two hypotheses for how neural representations might adjust following boundary training to permit successful performance on *test long* (Figure 5A). First, under a *hierarchical* coding scheme, the parallel lines observed in *test short* (RDM_{mag_short}) might separate along a direction perpendicular to the within-context magnitude axis, so that one dimension codes for a "superordinate" rank given by context (i.e., $[i_1–i_6] > [i_7–i_{12}]$) and the other for rank within each context (e.g., $i_2 > i_3$ and $i_9 > i_{10}$).[35] This effectively implements a place-value (or "dimension-value") representational scheme (akin to numbers in base 6; Figure 5A, central panel). Note that this hierarchical coding scheme would not require altering the learned neural representation of within-context magnitude but would instead just incorporate contextual information along a perpendicular dimension, thus predicting increased coefficients for RDM_{ctx}. Alternatively, under the *elongation* scheme, objects could be neurally rearranged on a single line stretching from i_1 to i_{12} to represent the objects on a single dimension (RDM_{mag_long}; Figure 5A, right panel). Such few-shot knowledge assembly would alter the relative geometry of the contexts along the dimension encoding magnitude. We thus constructed a new RDM_{mag_long} that encoded the predictions of this elongation model (Figure 5C, bottom panel). We note that nothing in our training protocol privileges one model over the other; both schemes could allow learning from the boundary training (for items i_6 and i_7) to be rapidly generalized to other items in each context, by either shifting the context perpendicularly to the magnitude line (hierarchical) or by sliding each context along the magnitude line (elongation). We first compared these schemes empirically by fitting model RDMs to multivoxel pattern data in PPC and dmPFC (Figure 5B, top row). We compared two regression models, one in which the model RDM was generated under the elongation scheme and one under the hierarchical scheme. We found that neural data were better fit by the elongation model in both PPC and dmPFC ($t_{33} = 4.2$; $t_{33} = 5.2$; paired t test on residual sum of squared error; significant at FWE threshold $p < 0.001$). Indeed, in both PPC and dmPFC, we observed positive correlations with RDM_{mag_long} (PPC: $t_{33} = 4.3$, $p < 0.001$; dmPFC: $t_{33} = 6.0$, $p < 0.001$) and when we plotted the neural geometry it can be seen that they lay on a single (curved) line, consistent with the elongation scheme (Figure 5B, bottom row). The curvature of the line is such that the manifolds for the two contexts lie almost perpendicular to each other while still permitting a one-dimensional readout (V test for nonuniformity of circular data with a known mean direction of 1.57 radians [90°]: $v = 33.65$, $p < 0.001$ in PPC; $v = 33.89$, $p < 0.001$ in dmPFC); the average angle between the projections was 1.45 ± 0.08 radians in PPC and 1.51 ± 0.05 radians in dmPFC (mean ± SEM). By contrast, we found no evidence for the hierarchical coding scheme, and in particular no effect of context in our ROIs (PPC: $t_{33} = 0.2$, dmPFC: $t_{33} = 0.6$, both p values > 0.5). Finally, we confirmed the fit of the elongation model in frontal regions using a searchlight approach (Figure 5C, top panel) (peak in left frontal gyrus: −30 23 23, $t_{33} = 6.60$, significant at FWE threshold $p < 0.01$) and continue to see a better fit for the elongation model when RTs were included as a nuisance covariate (PPC: $t_{33} = 3.5$, $p < 0.002$; dmPFC: $t_{33} = 5.3$, $p < 0.001$; paired t tests on residual sum of squared error; Figure S1D).

Interestingly, although neural codes in the dmPFC no longer correlated with RDM_{mag_short} at *test long* ($t_{33} = 1.9$, $p = 0.06$; t test on *Z* scored Pearson correlations with RDM_{mag_short}), the PPC continued to residually code for two overlapping neural lines ($t_{33} = 3.1$, $p = 0.004$; regression with design matrix [RDM_{mag_short}, RDM_{mag_long}], t test on *Z* scored RDM_{mag_short} beta weights). We speculated that this residual coding for the *test short* geometry (i.e., parallel lines) may predict the inability of some participants to integrate new knowledge (see Figure 4E). Indeed, we found that participants with a tendency to respond with latencies expected in *test* short also displayed a neural geometry more reminiscent of *test short* in PPC ($r = 0.50$, $p = 0.002$; Pearson correlation between behavioral fits to *test short* RT matrix and neural fit to RDM_{mag_short}). We summarized this relationship by relating the degree of neural elongation (difference in fit for RDMmag_long − RDM_{mag_short}) to the degree of behavioral integration (difference in the fit of idealized RT matrices, *test long* − *test short*) (Figure 5D, left panel) ($r = 0.49$, $p < 0.01$). We also saw this relationship in dmPFC (Figure 5D, right panel) ($r = 0.40$, $p < 0.05$), but it did not reach the threshold in the visual cortex ($r = 0.33$, $p = 0.06$). We note that this effect was not significant when choice was used as a measure of behavior rather than RT (Figures S1F and S1G), which could be due to a correlation between the idealized *test long* and *test short* accuracy matrices ($r = 0.48$, $p < 0.01$). Using a searchlight approach within the frontoparietal network that coded for RDM_{mag_short} during *test short* (see Figure 3D), we found that this neural-behavioral relationship was expressed most strongly in the right superior frontal gyrus (Figure 5E; significant at FWE threshold $p < 0.05$, $r \geq 0.55$, $p_{uncorr} < 0.001$) (peak correlation: 17 34 54; $r = 0.71$, $p_{uncorr} < 0.001$; significant at FWE threshold $p < 0.01$) along with being evident in the left parietal cortex (Figure 5E) (peak correlation: −48 45 42; $r = 0.65$, $p_{uncorr} < 0.001$; significant at FWE threshold $p < 0.05$). Finally, we conducted extensive analyses to explore the possibilities that neural results obtained in both *test short* and *test long* could be driven by "difficulty" or that they encode the extreme items (an "end-anchoring" effect) (see Figure S2) and found that magnitude is expressed alongside an end-anchoring effect such that unambiguous items are

(D) Neural-behavioral correlations in PPC (left) and dmPFC (right) ROIs. The x and y axis show relative behavioral model fits (*test long* − *test short* RT matrices) vs. neural fits ($RDM_{mag\_long} - RDM_{mag\_short}$). The axes display the relative RDMs (y axis) and relative RT matrices (x axis) from which the neural and behavioral scores were calculated. Each dot is a participant, colored by their accuracy during *test long*. p values in all panels indicate significance from t tests of Pearson correlation values against zero.

(E) Voxels showing a significant neural-behavioral correlation. Significance values shown in (B) and (D) correspond to FWE multiple comparison correction thresholds (2 ROIs and 2 behavioral metrics, respectively) at a threshold of FWE $p < 0.01$.

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

CellPress
OPEN ACCESS

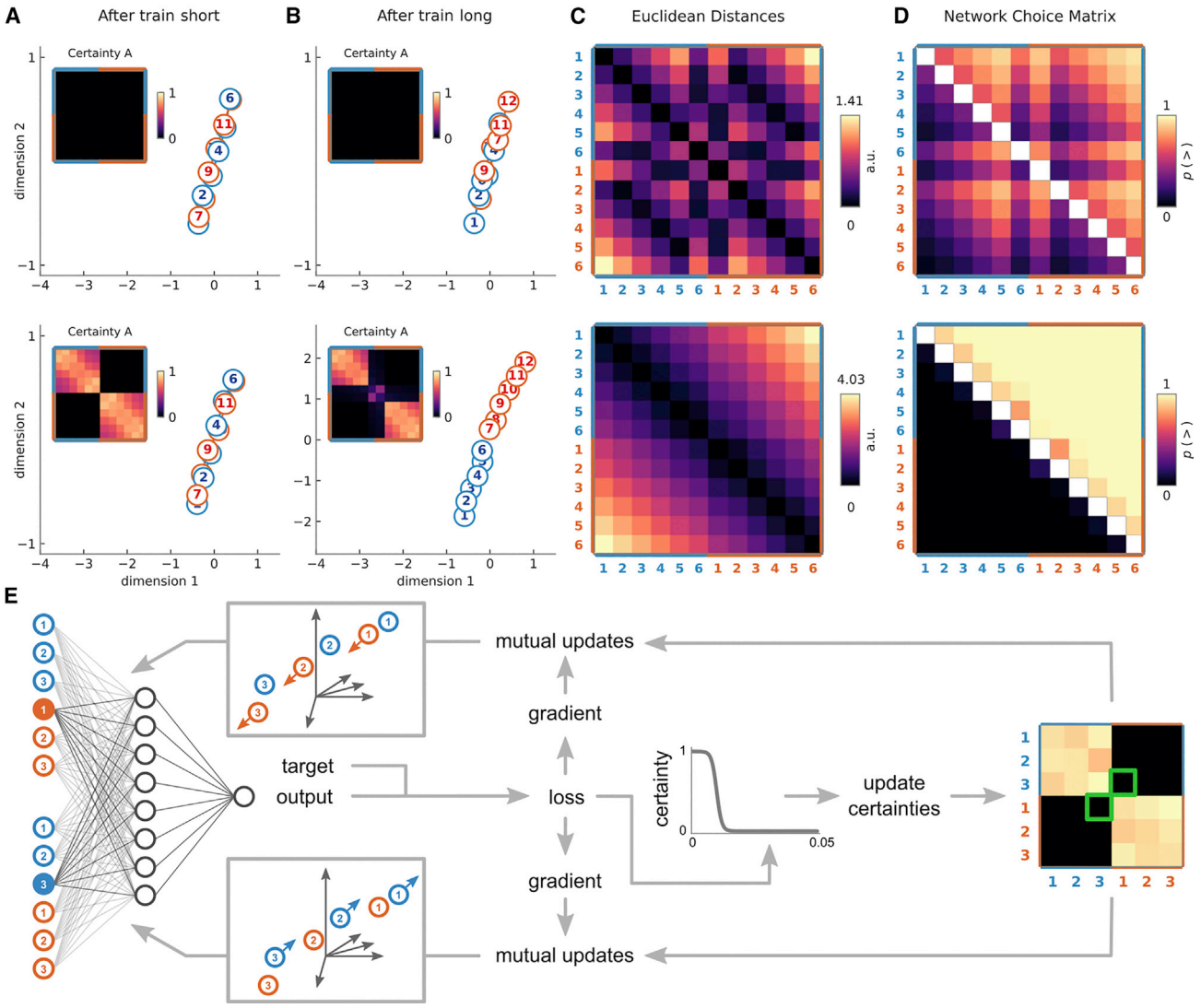

Figure 6. Knowledge assembly in artificial neural networks

(A–D) Fits of neural network to lower (top row) and higher (bottom row) human performance on *test long*. Fits were generated with $\gamma = 2e^{-3}$, which leads to SGD-like training, and $\gamma = 0.11$ for lower and higher performers, respectively. (A and B) show two-dimensional MDS of hidden layer representations after *train short* and *train long* (boundary training) with accompanying certainty matrices (inset). Note the lack of certainty acquisition for the fit to lower performers (top row), suggesting that there is no relational representation among items. After *train short*, embeddings for the two contexts (1–6 blue and 1–7 red) lie on two overlapping lines for both fits. After *train long,* these lines are only slightly elongated (learning items 6 and 7 as an exception) for the fit to lower performers but show full elongation for the fit to higher performers (separation between red and blue dots). Note that as fits were generated with $\gamma = 2e^{-3}$ and $\gamma = 0.11$ for lower and higher performers, respectively, this minimal change allows the network to behave like the best human performers. (C and D) RDMs of hidden layer representations and fitted network choice matrix after *train long*. Equivalent matrices for humans are shown in Figure 4D.

(E) Schematic of the update rule. The network output is calculated for two input items (filled dots) and then compared with the target value to calculate the loss. From the loss, the certainty is calculated and the corresponding entries in the certainty matrix updated (green squares). Then, for each of the two items, the gradients are combined with the respective column of the certainty matrix to calculate mutual updates (see STAR Methods). Note that only 3 items per context are shown for simplicity, without the loss of generality.

represented as more distal in neural space. This is also visible in the MDS plots in Figures 3C and 5B.

How, then, might knowledge assembly occur on the computational level? Training the neural network with vanilla SGD (as in Figures 2 and 3) allowed us to capture initial human learning and the emergence of two overlapping neural magnitude lines in the brain. However, it did not allow the rapid few-shot knowledge assembly that is characteristic of human behavior. In fact, even after prolonged boundary training on $i_6 < i_7$, the network learns this comparison as an exception (Figures 6A–6D, top row), thus failing to generalize the greater (lesser) brispiness to other items in context A (B). One possibility is that human participants store and mentally replay pairwise associations learned during previous training, a contention we test in

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

Figures S3A–S3D, finding results that are quite discrepant with both the behavior and neural geometry seen in human data (Figures S3F–S3H). Thus, in the final part of our report, we describe an adaptation of SGD that can account for the behavior and neural coding patterns exhibited by human participants, including the rapid reassembly of knowledge and its expression on a fast-changing neural manifold.

We reasoned that a simple computational innovation within the neural network could account for the pattern of knowledge assembly observed at *test long* (Figure 6E). We can think of a neural network as learning to embed inputs on a manifold with maximum potential dimensionality of d, equal to the number of hidden units. As we have seen, during training, the network learns to represent stimuli on a manifold with low intrinsic dimensionality (a single neural axis) that represents the transitive series from either context with overlapping embeddings (e.g., Figure 3A). The assumption we make now is that the network retains a certainty estimate regarding each relation in the embedding space (we call this the *certainty matrix* A). For example, as the relation between items i_3 and i_4 is acquired by the network and the loss consequently decreases, the respective certainty value $A3,4 = A4,3$ increases. With this assumption, new updates can propagate to conserve more certain relations in embedding space while allowing less certain relations to change (see STAR Methods). This model generalizes vanilla SGD, which is the special case where γ, a free parameter that determines how quickly the certainty matrix is updated, equals 0. When training the neural network to solve the transitive inference task, it is possible to recover both successful and less successful knowledge assembly observed in humans by varying γ (Figures 6A–6D and S4).

We gave the network approximately the same number of *train short* and *train long* trials as human participants had experienced (960 trials and 20 trials respectively). Over *test short* training, the network learned to represent items on two magnitude lines, regardless of the value of gamma (Figure 6A). Interestingly, performance and training dynamics were indistinguishable across γ values after *train short* (Figure S4A). However, the values of the certainty matrix A for within-context relations still depended on γ (Figure 6A, inset). SGD-like training dynamics ($\gamma \approx 0$, Figures 6A and 6B upper panel) failed to encode relations within-contexts, and so boundary items were treated as an exception while magnitude lines for each context continued to overlap. By contrast, networks with $\gamma \approx 0.1$ (Figures 6A and 6B lower panel) learned with high certainty that objects were related within contexts (i_1–i_6 and i_7–i_{12}). As a result, this allowed mutual parameter updates to conserve these relations even with the limited information provided during boundary training (Figure 6B, lower panel). These updates pushed the contexts in opposing directions, qualitatively consistent with the elongation scheme observed in humans. In fact, we found that we could capture both the high-performing human participants and the participants who learned $i_6 > i_7$ as an exception by varying γ (Figure 6D). This exercise revealed good fits for low performers at both $\gamma = 2e^{-3}$ and $\gamma = 0.87$, while only one minimum around $\gamma = 0.11$ fit the participants who correctly assembled the knowledge structures (see STAR Methods).

DISCUSSION

We report behavioral and neural evidence for rapid knowledge assembly in human participants. Just 20 boundary training trials were enough for most participants to learn how two sets of related objects were linked. Strikingly, neural representations in multivoxel BOLD patterns rapidly reconfigured into a novel geometry that reflects this knowledge, especially in dorsal stream structures such as PPC and dmPFC. However, neural networks fail at this few-shot knowledge assembly problem. Our main questions thus concerned the mechanisms by which humans achieve this and how they might be modeled in a neural network.

List linking requires participants to make inferences that go beyond the training data. In this case, after boundary training it is parsimonious to assume that because item $i_6 > i_7$, then item $i_7 < i_{1-6}$, and $i_6 > i_{7-12}$. How are these inferences made? Our simulations show that a vanilla neural network does not naturally show this behavior, nor would we expect more advanced architectures for relational reasoning, which typically learn from many thousands of training examples, to show few-shot learning. One possibility is that a dynamic process occurs at the time of inference, proposed by models of transitive inference based on the hippocampus in which updates spread across items via online recurrence.[36] However, because it takes more cycles to bridge the associative distance between disparate items (e.g., i_1 and i_6), this scheme predicts that these comparisons would garner longer RTs and lower accuracy rates—the opposite of the symbolic distance effect we see here.

An alternative is that periods of sleep or quiet resting may allow for replay events, such as those associated with sharp wave ripples in rodents and humans, which might facilitate planning and inference,[37] as well as spontaneous reorganization of mental representations during statistical learning.[38] We acknowledge that it is possible that replay occurring during *train long* leads to a readjustment of the item ranks and contributes to the few-shot learning we measure. However, our paradigm allowed very little time for rehearsal or replay—boundary training was few-shot. Our neural network modeling approach allowed us to estimate that prohibitively many full-batch replay events would be required for replay to permit successful knowledge assembly. Moreover, we failed to observe any increment in performance for item pairs that should benefit from offline rehearsal (i.e., those whose rank reversed), as predicted by replay models. Taken together, we think that it is unlikely that replay is the main driver of participants' behavior on *test long*.

Instead, our model proposes that items are earmarked during initial learning in a way that might help future knowledge restructuring, by coding certainty about relations among items (here, a trained transitive ordering). We describe such a mechanism and show that it can account for our data. Our model is agnostic about how precisely certainty is encoded, but one idea is that in neural systems connections may become tagged in ways that render them less labile. On a conceptual level, this resembles previously proposed solutions to continual learning that freeze synapses to protect existing knowledge from overwriting.[39,40] Thus, notwithstanding a recent interest in replay as a basis for structure memory—including in humans[41–43]—our model has implications for the understanding of other

phenomena that involve retrospective revaluation or representational reorganization, such as sensory preconditioning.[44]

One curiosity of our findings is that unlike for the neural network models, neural manifolds for the transitive series were not straight but inflected around the midpoint (ranks 3/4 in *test short* or 6/7 in *test long*), forming a horseshoe shape in low-dimensional space. We have previously observed this pattern in geometric analysis of whole-brain scalp EEG signals evoked by transitively ordered images,[32] and a recent report has emphasized a similar phenomenon in macaque PPC and medial PFC during discrimination of both faces and dot motion patterns.[33] The reasons for this form of manifold shape is unclear. One possibility is that the axis coding for choice certainty is driven by the engagement of additional control processes and that these processes are currently missing from our neural network model.[45] Another possibility is that this geometry conserves independence between contexts through orthogonalization[46] while aligning them along a single optimal readout dimension. We further note that the curvature occurs naturally in a model in which inputs are subject to Gaussian noise (rather than one-hot, as in our neural network model; Figure S2A). Resolving this issue is likely to be an important goal for future studies.

In sum, we observed rapid reorganization of neural codes for object relations in dorsal stream structures, including the PPC and dmPFC. This is consistent with a longstanding view that dorsal structures, and especially the parietal cortex, encode an abstract representation of magnitude or a mental "number line."[7,47,48] Recently, many studies have emphasized instead that the MTL, and especially the hippocampus and entorhinal cortex, may be important for learning about the structure of the world.[3,6,49,50] One important difference between our work and many studies reporting MTL structures is that our study involved an active decision task, whereas previous studies have used passive viewing or implicit tasks to measure neural structure learning. We do not doubt that both regions are important for coding relational knowledge, but their precise contributions remain to be defined.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants
- METHOD DETAILS
  - Stimulus and task
  - fMRI data acquisition
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Idealised behavioural matrices
  - fMRI data analysis
  - Representational similarity analysis
  - Support vector machine decoding
  - Brain behavior correlations
  - Difficulty and end anchoring simulations
  - Replay-based methods
  - REMERGE model
  - Cross task normalization
  - Neural network simulations
  - Learning relational certainty
  - Fitting human choice matrices
  - Neural network representations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.02.014.

ACKNOWLEDGMENTS

We are grateful to Thomas Blennerhassett for their valuable input on neural network simulations and Keno Juchems for their comments on the manuscript. We thank Palmoa Diaz, José M. Colino Jiménez, and Félix A. Navas Moya at Universidad de Granada for their assistance with fMRI data acquisition.

This work was supported by generous funding from the European Research Council (ERC award no. 725937) to C.S. and by Special Grant Agreement no. 945539 (Human Brain Project SGA) to C.S. L.B. was supported by the Woodward Scholarship awarded by Wadham College, Oxford and the Medical Research Council (MR/N013468/1). A.S. was supported by the Gatsby Charitable Foundation (GAT3755), by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z), and the Sainsbury Wellcome Centre Core Grant (219627/Z/19/Z). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program. T.D. was supported by an Economic and Social Research Council (ESRC) doctoral studentship and a University of Oxford Scatherd Scholarship.

AUTHOR CONTRIBUTIONS

S.N. and C.S. conceived human experiments. S.N. implemented the experiments. S.N. and T.D. collected human behavioral and fMRI data. S.N., C.S., and A.S. conceived the analyses. S.N. implemented the analyses. L.B., S.N., C.S., and A.S. conceived neural network simulations. L.B. and S.N. implemented the neural network simulations. S.N., L.B., and C.S. drafted the paper. S.N., L.B., and C.S. edited and revised the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 29, 2022
Revised: December 21, 2022
Accepted: February 9, 2023
Published: March 9, 2023

REFERENCES

1. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253. https://doi.org/10.1017/S0140525X16001837.
2. Morton, N.W., and Preston, A.R. (2021). Concept formation as a computational cognitive process. Curr. Opin. Behav. Sci. *38*, 83–89. https://doi.org/10.1016/j.cobeha.2020.12.005.
3. Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. Neuron *100*, 490–509. https://doi.org/10.1016/j.neuron.2018.10.002.
4. Lynn, C.W., and Bassett, D.S. (2020). How humans learn and represent networks. Proc. Natl. Acad. Sci. USA *117*, 29407–29415. https://doi.org/10.1073/pnas.1912328117.

5. Tervo, D.G.R., Tenenbaum, J.B., and Gershman, S.J. (2016). Toward the neural implementation of structure learning. Curr. Opin. Neurobiol. *37*, 99–105. https://doi.org/10.1016/j.conb.2016.01.014.

6. Bellmund, J.L.S., Gärdenfors, P., Moser, E.I., and Doeller, C.F. (2018). Navigating cognition: spatial codes for human thinking. Science *362*, eaat6766. https://doi.org/10.1126/science.aat6766.

7. Summerfield, C., Luyckx, F., and Sheahan, H. (2020). Structure learning and the posterior parietal cortex. Prog. Neurobiol. *184*, 101717. https://doi.org/10.1016/j.pneurobio.2019.101717.

8. Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208. https://doi.org/10.1037/h0061626.

9. Schapiro, A.C., Kustner, L.V., and Turk-Browne, N.B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr. Biol. *22*, 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056.

10. Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. Nat. Neurosci. *16*, 486–492. https://doi.org/10.1038/nn.3331.

11. Garvert, M.M., Dolan, R.J., and Behrens, T.E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife *6*, e17086. https://doi.org/10.7554/eLife.17086.

12. Zeithamova, D., and Preston, A.R. (2017). Temporal proximity promotes integration of overlapping events. J. Cogn. Neurosci. *29*, 1311–1323. https://doi.org/10.1162/jocn_a_01116.

13. Horner, A.J., and Doeller, C.F. (2017). Plasticity of hippocampal memories in humans. Curr. Opin. Neurobiol. *43*, 102–109. https://doi.org/10.1016/j.conb.2017.02.004.

14. Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Cell *183*. 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024.

15. Dordek, Y., Soudry, D., Meir, R., and Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. eLife *5*, e10094. https://doi.org/10.7554/eLife.10094.

16. Klukas, M., Lewis, M., and Fiete, I. (2020). Efficient and flexible representation of higher-dimensional cognitive variables with grid cells. PLoS Comput. Biol. *16*, e1007796. https://doi.org/10.1371/journal.pcbi.1007796.

17. Collin, S.H., Milivojevic, B., and Doeller, C.F. (2017). Hippocampal hierarchical networks for space, time, and memory. Curr. Opin. Behav. Sci. *17*, 71–76. https://doi.org/10.1016/j.cobeha.2017.06.007.

18. Theves, S., Neville, D.A., Fernández, G., and Doeller, C.F. (2021). Learning and representation of hierarchical concepts in hippocampus and prefrontal cortex. J. Neurosci. *41*, 7675–7686. https://doi.org/10.1523/JNEUROSCI.0657-21.2021.

19. Collin, S.H.P., Milivojevic, B., and Doeller, C.F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. Nat. Neurosci. *18*, 1562–1564. https://doi.org/10.1038/nn.4138.

20. Lee, S.W., O'Doherty, J.P., and Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. PLOS Biol. *13*, e1002137. https://doi.org/10.1371/journal.pbio.1002137.

21. Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? Nat. Rev. Neurosci. *22*, 55–67. https://doi.org/10.1038/s41583-020-00395-8.

22. Lindsay, G.W. (2021). Convolutional neural networks as a model of the visual system: past, present, and future. J. Cogn. Neurosci. *33*, 2017–2031. https://doi.org/10.1162/jocn_a_01544.

23. Barrett, D.G.T., Hill, F., Santoro, A., Morcos, A.S., and Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. https://doi.org/10.48550/arXiv.1807.04225.

24. Chang, M.B., Gupta, A., Levine, S., and Griffiths, T.L. (2019). Automatically composing representation transformations as a means for generalization. https://doi.org/10.48550/arXiv.1807.04640.

25. Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science *350*, 1332–1338. https://doi.org/10.1126/science.aab3050.

26. Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C.P., Bosnjak, M., Shanahan, M., Botvinick, M., Hassabis, D., and Lerchner, A. (2018). SCAN: learning hierarchical compositional visual concepts. International Conference on Learning Representations (ICLR).

27. Horst, J.S., and Hout, M.C. (2016). The Novel Object and Unusual Name (NOUN) Database: a collection of novel images for use in experimental research. Behav. Res. Methods *48*, 1393–1409. https://doi.org/10.3758/s13428-015-0647-3.

28. Woocher, F.D., Glass, A.L., and Holyoak, K.J. (1978). Positional discriminability in linear orderings. Mem. Cogn. *6*, 165–173. https://doi.org/10.3758/BF03197442.

29. D'Amato, M.R., and Colombo, M. (1990). The symbolic distance effect in monkeys (Cebus apella). Anim. Learn. Behav. *18*, 133–140. https://doi.org/10.3758/BF03205250.

30. Chen, S., Swartz, K.B., and Terrace, H.S. (1997). Knowledge of the ordinal position of list items in rhesus monkeys. Psychol. Sci. *8*, 80–86. https://doi.org/10.1111/j.1467-9280.1997.tb00687.x.

31. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., and Summerfield, C. (2021). Rich and lazy learning of task representations in brains and neural networks. Neuron. https://doi.org/10.1101/2021.04.23.441128.

32. Luyckx, F., Nili, H., Spitzer, B., and Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. eLife *8*, e42816. https://doi.org/10.7554/eLife.42816.

33. Okazawa, G., Hatch, C.E., Mancoo, A., Machens, C.K., and Kiani, R. (2021). Representational geometry of perceptual decisions in the monkey parietal cortex. Cell *184*. 3748.e18–3761.e18. https://doi.org/10.1016/j.cell.2021.05.022.

34. Bernardi, S., Benna, M.K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C.D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. Cell *183*. 954–967.e21. https://doi.org/10.1016/j.cell.2020.09.031.

35. Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., and Summerfield, C. (2021). Neural state space alignment for magnitude generalization in humans and recurrent networks. Neuron *109*. 1214–1226.e8. https://doi.org/10.1016/j.neuron.2021.02.004.

36. Kumaran, D., and McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. Psychol. Rev. *119*, 573–616. https://doi.org/10.1037/a0028681.

37. Hunt, L.T., Daw, N.D., Kaanders, P., MacIver, M.A., Mugan, U., Procyk, E., Redish, A.D., Russo, E., Scholl, J., Stachenfeld, K., et al. (2021). Formalizing planning and information search in naturalistic decision-making. Nat. Neurosci. *24*, 1051–1064. https://doi.org/10.1038/s41593-021-00866-w.

38. Liu, Y., Dolan, R.J., Kurth-Nelson, Z., and Behrens, T.E.J. (2019). Human replay spontaneously reorganizes experience. Cell *178*. 640–652.e14. https://doi.org/10.1016/j.cell.2019.06.012.

39. Zenke, F., Poole, B., and Ganguli, S. (2017). Continual learning through synaptic intelligence. Proceedings of the 34th International Conference on Machine Learning*70* (PMLR), pp. 3987–3995.

40. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA *114*, 3521–3526. https://doi.org/10.1073/pnas.1611835114.

41. Kurth-Nelson, Z., Economides, M., Dolan, R.J., and Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. Neuron *91*, 194–204. https://doi.org/10.1016/j.neuron.2016.05.028.

42. Wimmer, G.E., Liu, Y., Vehar, N., Behrens, T.E.J., and Dolan, R.J. (2020). Episodic memory retrieval success is associated with rapid replay of episode content. Nat. Neurosci. *23*, 1025–1033. https://doi.org/10.1038/s41593-020-0649-z.

43. Nour, M.M., Liu, Y., Arumuham, A., Kurth-Nelson, Z., and Dolan, R.J. (2021). Impaired neural replay of inferred relationships in schizophrenia. Cell *184*. 4315–4328.e17. https://doi.org/10.1016/j.cell.2021.06.012.

44. Wimmer, G.E., Daw, N.D., and Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. Eur. J. Neurosci. *35*, 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x.

45. Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. Psychol. Rev. *108*, 624–652. https://doi.org/10.1037/0033-295X.108.3.624.

46. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., and Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. Neuron *110*. 1258–1270. e11. https://doi.org/10.1016/j.neuron.2022.01.005.

47. Hubbard, E.M., Piazza, M., Pinel, P., and Dehaene, S. (2005). Interactions between number and space in parietal cortex. Nat. Rev. Neurosci. *6*, 435–448. https://doi.org/10.1038/nrn1684.

48. Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. Trends Cogn. Sci. *7*, 483–488. https://doi.org/10.1016/j.tics.2003.09.002.

49. Morton, N.W., Sherrill, K.R., and Preston, A.R. (2017). Memory integration constructs maps of space, time, and concepts. Curr. Opin. Behav. Sci. *17*, 161–168. https://doi.org/10.1016/j.cobeha.2017.08.007.

50. Yu, L.Q., Park, S.A., Sweigart, S.C., Boorman, E.D., and Nassar, M.R. (2021). Do grid codes afford generalization and flexible decision-making?. Conference on Cognitive Computational Neuroscience. [q-bio]. https://doi.org/10.48550/arXiv:2106.16219.

51. Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. *35*, 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015.

52. Flandin, G., and Friston, K.J. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. Hum. Brain Mapp. *40*, 2052–2054. https://doi.org/10.1002/hbm.23839.

53. Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn. Sci. *20*, 512–534. https://doi.org/10.1016/j.tics.2016.05.004.

54. McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419–457. https://doi.org/10.1037/0033-295X.102.3.419.

55. Vapnik, V.N. (1995). Setting of the learning problem. In The Nature of Statistical Learning Theory (Springer), pp. 15–32. https://doi.org/10.1007/978-1-4757-2440-0_2.

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MATLAB 2022	Mathworks	https://www.mathworks.com
Python version 2.7	Python Software Foundation	https://www.python.org
Custom Code for simulating Neural Networks	Lukas Braun	https://github.com/lukas-braun/neural-knowledge-assembly https://doi.org/10.5281/zenodo.7579723
Neural Data for Figures 3 and 5	Stephanie Nelli	https://osf.io/wkx5h/ https://doi.org/10.17605/OSF.IO/X3B9M

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Stephanie Nelli (nelli@oxy.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

De-identified human fMRI data for Figures 3 and 5 have been deposited at Open Science Foundation. They are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

All original code utilized for neural network simulations has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

Thirty-seven healthy adult participants were recruited for this study. One was excluded for failure to reach performance threshold (see below), and two more for practical reasons (failure to show up for the experiment; discomfort in scanner leading to early termination of the experiment). This left n = 34 total (19 males, mean ± SD age: 23.3 ± 3.4 years). Participants reported no history of psychiatric or neurological disorders, and gave informed consent prior to scanning. The study was approved by the ethics committee of the University of Granada and informed consent was obtained from all participants. Participants' base compensation was 35 Euros, and also could receive a performance-based bonus for an average payment of 40.93 ± 2.57 (mean ± SD) Euros. After the experiment, participants were given a voluntary anonymous debrief concerning their insight into the *test long* phase, which we display in Table S2.

METHOD DETAILS

Stimulus and task

Stimuli were novel objects drawn from the NOUN database.[27] Out of the 60 possible images in this database, objects that were rated as most similar to the others (e.g., an average similarity rating within 1 standard deviation of the maximum) and objects that were rated as most familiar (e.g., scoring less than 50% on an inverse familiarity score) were excluded, leaving 41 possible objects. For each participant, 12 of these 41 objects were randomly selected and arbitrarily assigned a rank from 1-12, with ranks 1-6 belonging to context A and 7-12 to context B. We denote these i_1–i_{12} in the text.

Before entering the scanner, participants performed a computerised training phase which we call *train short*. This training phase consisted of between 3 and 10 cycles of 120 trials. Each cycle consisted of two blocks of 60 trials. On each block, objects were sampled from a single context for 60 trials (A or B), and then the alternate context was presented for another block of 60 trials.

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

Each trial began with the presentation of two objects drawn from adjacent ranks within a single context (e.g. *i*3 and *i*4 or *i*8 and *i*9). These objects were shown either side of a central fixation point, and above the point the words "more brispy?" or "less brispy?" appeared in Spanish (i.e. "mas brispo?" or "menos brispo?"). Participants were instructed to select the corresponding object (i.e. that which was more or less brispy) using either the "F" (left object) or "J" (right object) keys. These objects remained on screen for 5000 ms or until response, whichever was shorter. Once a response was recorded, a green (correct) or red (incorrect) box would appear around the selected object to indicate whether it was the correct selection, and this response feedback box persisted for 475 ms. If participants did not respond within 5000 ms of stimulus presentation, the trial was considered incorrect and was not repeated. After feedback there was a blank screen for a variable delay of up to 50 ms before the next trial. Critically, participants were only trained to compare 5 object pairs from each context, e.g., i_1–i_2, i_2–i_3, i_3–i_4, i_4–i_5, and i_5–i_6, from context A and i_7–i_8, i_8–i_9, i_9–i_{10}, i_{10}–i_{11}, and i_{11}–i_{12}, from context B.

Whether participants were asked to select the more or less brispy object, and hemifield presentation of the objects, were randomised on each trial. Additionally, the trial-order of each object pair was randomly shuffled. Participants performed this task for a minimum of 3 cycles (3 blocks per context, for 6 total blocks), and were trained until they reached a criterion of at least 90% correct on each block of the last cycle (e.g., more than 90% accurate on each of the two previously performed blocks, where one was from each context). Additionally, participants had to select the correct answer on the final 20% of trials (12) within each context. On average, it took participants 5 ± 2 cycles (mean ± SD) to learn the task, and one participant was excluded for failing to reach criterion after 10 cycles of training (see Figure 1C).

The format of the test phases *test short* and *test long* was identical (but different to *train short*). Test phases occurred in the scanner, and consisted of 288 trials (4 blocks of 72 trials each) in which lone objects were presented in a random sequence, with the constraint that each combination of 12 (current trial) x 12 (previous trial) ranked objects occurred exactly once in the first half (144 trials) and once in the second half (144 trials) of the test phase. Each object was presented centrally for 750 ms, after which participants had 2000 ms to respond whether it was more or less brispy than the previous object. Participants were instructed about the mapping from more/less to left/right buttons (held in either hand) before each block of trials, and this mapping stayed consistent for 2 blocks, and then was switched midway through the test phase. After participants responded, there was a pseudo-randomly jittered interval of 1500-5500 ms, during which the fixation dot turned blue if a response was recorded within this deadline, while a red letter X appeared if the response was missed. Critically, participants did not receive trial-wise feedback and instead were rewarded bonus points at the end of each block. These bonus points were proportional to their accuracy on that block and were translated into additional monetary reward at the end of the experiment. Timings for test were chosen to assist with BOLD modelling.

The boundary training phase (*train long*) occurred between *test short* and *test long*. It was similar to *train* short except that it lasted just 20 trials, and the only items presented were the objects i_6 and i_7. These could occur on either side of the screen, with "more" or "less" randomised over trials as in *train short*. Participants viewed the objects for 3000 ms after which a feedback screen stayed up for 1500 ms, be it a green (correct) or red (incorrect) bounding box, or a red X at fixation if no response was recorded. There was then a variable intertrial interval from 1400-5000 ms before the next trial.

Finally, after completing boundary training (*train long*) and *test long*, participants remained in the scanner and performed a number localiser phase, which was identical to *test short* / *test long* with the exception that objects were replaced with Arabic digits 1–6 and participants responded "more" or "less" according to whether each number was greater or less than the previous. Note that we compare the representations elicited by each phase of the experiment (*test short*, *test long*, and the number localiser), and find evidence for a normalised coding scheme across the experimental phases in both PPC and dmPFC (Figure S6; Table S1).

fMRI data acquisition

MRI data were acquired on a 3T Siemens scanner. T1 weighted structural images were recorded directly prior to the task using an MPRAGE sequence: 1x1x1 mm^3 voxel resolution, 176x256x256 grid, TR = 2530 ms, TE = 2.36 ms, TI = 1100ms. Each fMRI image contained 72 axial echo-planar images (EPI) acquired at a multiband acceleration factor of 4 in interleaved sequence. Voxel resolution was 2 mm^3 isotropic, slice spacing of 1.6 mm, TR = 1355 ms, flip angle = 8, and TE of 32.4 ms. 560 EPI images were recorded for the number localiser and 1220 EPI images for each of the *test short* and *test long* runs. This resulted in 3000 EPI images per participant with a scanning time of about 100 min. Scans were realigned using an affine rigid body transformation minimizing the sum of squared differences between each scan and the mean scan within each run. All images were resliced, including the mean image. The anatomical scan was co-registered to the mean of all functional images using SPM's default mutual information matching criterion. Anatomical scans were normalized to the standard MNI152 template brain using SPM defaults and 2 x 2 x 2 mm^3 voxel resolution. This entails first a linear transformation to account for major differences in head shape and position, and then a non-linear warping transformation that uses deformations consisting of linear combinations of low frequency periodic basis functions to account for smaller differences in anatomy. We did not correct for susceptibility artefacts, meaning signal dropout could impact signal quality in areas like OFC. The functional EPI images were then normalized and smoothed with a full width half maximum Gaussian kernel of 8mm. Images were then downsampled by reslicing to 3 x 3 x 3 mm^3 voxel resolution before performing analyses.

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

QUANTIFICATION AND STATISTICAL ANALYSIS

Idealised behavioural matrices

We analyse behavioural data by plotting accuracies and RTs for each combination of 12 objects shown at test, and/or as a function of symbolic distance (i.e. the distance in rank between the current and previous item). We constructed idealised reaction time and choice matrices (Figure S1A) under the assumption that choices were noiseless triangular matrices and that RTs depended linearly on symbolic distance. Our accuracy matrices were created by setting the upper triangle of each quadrant (in *test short*) or the entire matrix (in *test long*) to 1 and the lower triangle to zero. In *test short*, this model collapses across context, meaning between-context comparisons are treated as if they came from the same context contexts (e.g. the 3rd item in one set should be ranked higher than the 4th in the other). Thus, this quantifies participant's tendency to perfectly match rank orderings between contexts (Figure S1A, top left panel). Specifically, our idealised RT matrices were constructed as $1/(1 + dist(v^T, v))$ where $v = [1{:}\,6\ 1{:}\,6]$ for *test short* and [1:12] for test_long, and v^T is the transpose of v. Note that as participants did not compare objects to themselves, diagonal elements of our design matrices were excluded from analyses.

fMRI data analysis

Scanner runs were concatenated and delta functions convolved with the canonical haemodynamic response function (HRF) and time-locked to trial events. All GLMs included an intercept term for each scanner run to account for differences in mean activation, and we also included the 6 head motion parameters derived from pre-processing as nuisance regressors [translation in x, y, z; yaw, pitch, roll]. Additionally, we used the default SPM high pass filter setting of 128 seconds, meaning event related variance slower than ~0.008 Hz was removed from the data to account for scanner drift. Automatic orthogonalization was switched off. Data were analyzed with SPM12 and in-house scripts. All contrasts were constructed as simple t-contrasts with first-level t-maps as input. Unless otherwise noted, we only report clusters that fell below an FWE-corrected p value of 0.01, which corresponded to a voxel-wise uncorrected threshold of $p < 0.001$ (as in Niv et al.[51]). These clusters were determined using random field theory, as is default in SPM 12, with the minimum cluster extent set to 10 voxels.[52] Data were visualized using the XjView toolbox (http://www.alivelearn.net/xjview).

We fit our data with 4 different general linear models (GLMs). The first GLM was used to define ROIs from the number localiser. The design matrix for GLM1 included parametric modulators time-locked to stimulus onset for each number (1-6), as well as 6 nuisance head motion regressors. We considered clusters of voxels that passed a FWE threshold of $p < 0.01$ in response to the stimulus regressor. Although several regions passed this threshold, we focussed on ROIs in dmPFC and PPC, chosen on the basis of previously stated predictions.[7] We show searchlight results in addition to ROI analyses, which seem to justify this choice. The second and third GLMs were used to estimate neural patterns associated with each object within *test short* and *test long*. The design matrix for these models each included 12 regressors, one for each of the objects locked to stimulus onset, as well as 6 additional nuisance regressors for head motion. In one case (GLM2) we estimated this regression separately for each block of 72 trials, where there were n = 4 blocks within both *test short* and *test long*. Notably, fMRI data corresponding to each experimental block of 72 trials was acquired within a distinct scanner run, allowing us to conduct analyses that required between-run cross-validation (e.g. SVM analysis). In the other case (GLM3) we simultaneously modelled all trials within *test short* or *test long*. This latter GLM was used for calculating RDMs in ROI and searchlight analyses of fMRI data. In a fourth GLM, we additionally included either 11 (*test short*) or 22 (*test long*) regressors coding for the distance from the current to previous image. Fits from GLM4 were used to generate data for multidimensional scaling visualization aids.

Representational similarity analysis

BOLD RDMs were constructed by taking the correlation distance between multivoxel patterns elicited by each of the objects in *test short* and *test long*, yielding a 12 x 12 RDM. For searchlight analyses, we used all voxels within a radius of 12mm of the center voxel. For each searchlight sphere or ROI, we computed the neural RDMs from the condition-by-voxel matrix of estimated neural responses using Pearson correlation distance between pairs of conditions.

These were compared to model RDMs which were created from linear distances between item ranks within context (1-6 and 7-12; RDM_{mag_short}), distances between ranks across contexts (1-12; RDM_{mag_long}) and between contexts themselves (i.e. 0 within context, 1 between context). All model RDMs were standardised and comparisons to neural data were conducted with tests of correlation (Pearson's *r*), or regression. The additional RDM reported here (RDM_{NN}) was obtained by taking the Euclidean distance between the 12 hidden-unit activations elicited by probing the network with one-hot inputs corresponding to i_1–i_{12}). Z-scored neural RDMs were regressed, or correlated, with z-scored (neural network) model RDMs. All statistics reported for RSA analyses were obtained by evaluating RDMs at the single subject level and conducting group- level (random effects) inference on the resulting coefficients, using FWE correction where appropriate. All reported probability values (p-values) are two-sided, and statistics reported at uncorrected significance levels are denoted p_{uncorr}. No further methods were employed to determine if data met assumptions of our statistical tests.

To visualise neural state spaces, we used multidimensional scaling with metric stress (equivalent to plotting the first principal components of the data) in two dimensions, using GLM4 (see above).

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

Support vector machine decoding

All classification analyses utilised a multiclass support vector machine model in MatLab. GLM2 generated 4 beta values for each object (one for each run), and we trained binary SVM classifiers on this data from objects in context 1, and then tested the model on objects in context 2. The classifier used a "one versus one" coding design, meaning that each learner *l* was trained on observations in 2 classes, treating one as the positive class and the other as the negative class and ignoring the rest. To exhaust all combinations of class-pair assignments, we fit $K * \frac{K-1}{2}$ binary SVM models where *k* are the unique classes (ranks 1-6 here). Specifically, let *M* be the ground truth coding design matrix with elements *mkl*, and *sl* be the predicted classification score for the positive class of learner *l* (without loss of generality). Then, the algorithm assigns a new observation (from context 2) to the class $\widehat{k}$ that minimizes the aggregate loss for the *L* binary learners.

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{l=1}^{L} |m_{kl}| G(m_{kl}, s_l)}{\sum_{l-1}^{L} |m_{kl}|}$$

We used MatLab default settings for the learners in the SVM model. Thus, each of these binary learners used a linear kernel function $G(m_{kl}, s_l) = m_{kl}' s_l$, and learned according to the default 'classiferror' loss function, which is simply the rate of misclassification in decimal format. We note that our SVM classification accuracy was not significantly higher in PPC [t_{33} = 1.48, p = 0.15] and dmPFC [t_{33} = 0.97, p = 0.34] than in visual cortex (Figure 3).

Brain behavior correlations

We performed a correlation analysis to quantify the extent to which the elongation of neural representations predicted integrated behavioural responses. We analyzed human choice patterns by computing idealised choice matrices (described above) and inputting both the *test short* and *test long* patterns into a competitive regression model. The degree of behavioural integration was defined as the relative fit of each of these matrices. Similarly, we constructed neural model RDMs describing the ground truth symbolic distance between each pair of items in *test short* and *test long* (see above). We then defined the degree of neural elongation as the relative fit to each of these RDMs in a competitive regression model. We then tested at the group level the extent to which the degree of neural elongation predicted the degree of behavioural integration using Pearson correlation. All statistics involving these metrics control for multiple comparisons using FWE correction.

Difficulty and end anchoring simulations

We also considered alternative explanations for the neural results obtained in both *test short* and *test long*. One possible concern is that they are driven by "difficulty", that is, the computational demand incurred by comparing each item to its predecessor in the *test short* and *test long* phases, which might vary with their rank distance. Another possibility is that PPC and dmPFC encode the extreme items (for which comparisons are unambiguous) and middle items (for which they are not) with distinct neural codes (an "end-anchoring" effect), but do not continuously represent rank or "magnitude" in either *test short* or *test long*, as implied by our analyses. We conducted extensive analyses to explore these possibilities, by building plausible control RDMs for "difficulty" and "end-anchoring" and using model comparison to arbitrate between them (Figure S2). We propose an account of our neural data based on rank (or magnitude), i.e., that BOLD signals code for the abstract property for "more" or "less" on a one-dimensional continuum. In Figure S2, we consider two competing accounts: that our BOLD data can be explained by "difficulty" (that is, the relative distance between each rank and the previous rank to which it was compared in the one-back task), and "end anchoring" (that is, that the BOLD signal codes for whether each item was unambiguously less (e.g., rank 1 in *test long*), intermediate (e.g., ranks 2-11 in *test long*) or unambiguously more (e.g., rank 12 in *test long*)).

To compare these accounts, we adopt a stylised "population coding" approach which assumes that the relevant input quantities are processed by an encoding model consisting of neurons with (potentially noisy) Gaussian tuning curves that tile the space of possible ranks (rank model) or difficulties (difficulty model) or end anchor status (end anchor model). Thus, in the rank (or magnitude) model there are neurons turned to various positions in rank space, and in the difficulty model these neurons are assumed to be tuned to positions in difficulty space, and in the end anchor model neurons are tuned to whether the item is at end, intermediate, or the other end of the ordinal scale. In Figure S2A (middle row) we show the relative RDMs for rank, difficulty, and end anchoring in *test short* and *test long*. Some dots in the multidimensional scaling plots (Figure S2 bottom row) fall in the same location, obscuring the pattern, so we describe it verbally here for *test short*, and note the same patterns hold for *test long*. In the rank plot (Figure S2A far left column), the following ranks overlap: [1,7; 2,8; 3,9; 4,10; 5,11; 6,12]. In other words, the stimuli form two parallel (overlapping) lines, organised by rank, very similar to the PPC and somewhat resembling the dmPFC in Figure 3C. By contrast, in the difficulty plot (Figure S2A 2nd column), there are only 3 levels of difficulty. This is because the difficulty is identical for items [1,6,7,12 (easiest); 2, 5, 8 11 (middle); 3 4 9 10 (hardest)]. The dots are not arranged in a line, or a curved line; rather, they are folded back on themselves, so that the most extreme (e.g., both 1 and 6) occupy exactly the same location. For the end anchor RDM (Figure S2A 3rd column), there are just three locations, with the majority of points occupying an intermediate position between the two extremes.

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

We found that the effect of magnitude (or rank) remained strongly statistically significant in both PPC (*test short*: t_{33} = 3.77, p < 0.001; *test long*: t_{33} = 4.21, p < 0.001) and dmPFC (*test short*: t_{33} = 3.34, p < 0.001; *test long* t_{33} = 6.17, p < 0.001) even when both control RDMs were included. Overall, difficulty was a weak explanation for the neural data RDMs, except in PPC for *test short* where it explained residual variance in neural signals (t_{33} = 2.09, p = 0.02); by contrast, a reliable effect of end-anchoring was observed in both PPC (*test short*: t_{33} = 3.69, p < 0.001; *test long*: t_{33} = 4.54, p < 0.001) and dmPFC (*test short*: t_{33} = 4.60, p < 0.001; *test long*: t_{33} = 4.73, p < 0.001, Figure S2B). Thus, it appears that the effect of magnitude is expressed alongside an end-anchoring effect, such that unambiguous items are represented as more distal in neural space than would otherwise be expected. This is also visible in the MDS plots in Figures 3C and 5D.

Replay-based methods

An alternative hypothesis to our model, which employs mutual updates to neural representations of objects to preserve knowledge during learning, is that previously knowledge is preserved by replaying previously learned input-output pairs as for example proposed in variants of Complementary Learning Systems (CLS) theory.[53,54] A replay-based account would require that participants store instance-based memories for comparisons between adjacent items (e.g., item i_3 and item i_4) and their respective target values (e.g. ''less than'') in a memory bank. Then, during training on the boundary condition, the network is also trained on the previously learned relationships in the memory bank. We considered the possibility that human participants store and mentally replay pairwise associations learned during previous training, intermingling these with instances of boundary training to avoid catastrophic interference.[54,53] However, our boundary training consisted of just 20 trials comparing i_6 and i_7 with no subsequent rest period; providing very little time for replay to occur. To test this contention more formally, we used neural network simulations to calculate whether this was sufficient time for replay to permit updating to occur in a way that would recreate human results. We considered full, random and ordered replay as candidate mechanisms. The results, which we detail in Figures S3A and S3B, imply that participants would have had to replay the full set (including the boundary condition) of 11 training pairs a total of 4 times per trial (44 training pairs/trial) to obtain full accuracy within 20 *train long* trials. Whilst quantification is difficult here, this seems to speak against strategies that reorganize the embedding space exclusively during the *train long* period. This analysis also seems to mitigate against potentially competing accounts of transitive inference such as REMERGE[36] which rely on offline replay to organize the embedding space even within a single context. Thus, we provide some computational evidence that the short (20 trials) list linking boundary training block is unlikely to leave sufficient time for replay.

As our neural network model fits the human data after being exposed to exactly as many *train long* trials as our participants, we simulated how many ''virtual replay trials'' after each training step on the boundary condition would be required to fit the data using four different replay curricula: (1) Ordered Replay: Replay samples from the memory buffer in an ordered fashion (e.g. 1<2, 2<3, 3<4...) (2) Random Full Replay: Randomly sample replay items from the memory buffer such that each memory is sampled exactly once (e.g. 2<3, 5<6, 9<10...) (3) Random Replay: Randomly sample replay items from the memory buffer, allowing for multiple draws of the same memory (e.g. 2<3, 5<6, 2<3...) (4) Highest Loss Replay: Sample the memory that currently leads to highest loss. We further distinguished between a memory buffer that does not include the boundary condition and one that does include it (IB = Include Boundary). Thus, the question becomes how many cycles (''epochs'') of replay from the memory buffer per training step on the boundary condition are required to disentangle the overlapping representations. As for the original model reported in the manuscript, each of these curricula were repeated for n=20 random seeds of the network to obtain measures of consistency. The simulations revealed that replaying the full set at least 4 times per each boundary training step (44 training pairs/trial) is required to obtain full accuracy within 20 boundary training trials (Figures S3A and S3B). This seems to be an implausible level of online replay required to account for our effects. We also confirmed that convergence time when replaying the full set 4 times (44 training pairs/trial) could not be reduced by adjusting the learning rate during replay (with respect to the learning rate η=0.05 during the boundary training trials, Figures S3C and S3D).

REMERGE model

We also implemented REMERGE, a well-known computational account of transitive inference (Figure S3E).[36] We provide predictions for REMERGE showing that they are quite discrepant with observed human data, including both behaviour and neural geometry (Figures S3F–S3I). We also considered the possibility that any replay is not confined to the *train long* period but continues into the subsequent *test long* epoch. Whilst this is possible, it would predict that over the course of *test long*, greater performance improvements would be observed for those item pairs that benefit from replay, i.e., the item pairs whose rank is reversed during the *train long* period. We saw no evidence for this (Figures S3J–S3M).

REMERGE is a two-layer neural network with recurrent connections from the hidden (i.e., conjunctive) to the feature layer and a linear readout from the hidden to the output layer (Figure S3E). Network inputs u encode the two input items as a two-hot vector, which is 0 everywhere except at the two positions of the inputs where it is set to 1. The neural dynamics for an input u evolve according to the following set of coupled differential equations:

$$\frac{dx}{dt} = -x + W_1 h + u \text{ and } \frac{dh}{dt} = -h + W_1 x,$$

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

Where x and h denote the dynamic feature and conjunctive layer of the network and W_1 is the recurrent weight matrix which connects the feature and the conjunctive layer with excitatory weights of size 0.5. The network's output at time t is then calculated from the hidden layer by a linear readout $y_t = W_2 h_t$. Excitatory and inhibitory readout weights are set to 1 and $-an$, respectively. For the exact wiring diagram of the network please refer to Figure S3E. Please note that all synaptic weights are fixed and not optimised during simulations.

Reaction times of the REMERGE model are calculated by means of a race condition to a fixed threshold between the two output nodes that refer to the two currently active input nodes (Figure S3E). We note that the REMERGE model makes incorrect predictions about the pattern of RTs that we see in human behaviour (Figure S3F cf. Figures 2B and 4A).

The choice matrix of the REMERGE model accurately predicts human behaviour during *test short* (Figure S3G). It is calculated by applying a softmax to the two output nodes that refer to the two input items after the temporal dynamics have converged. This active selection of output nodes is crucial for the REMERGE model to make accurate predictions, as other output nodes may dominate the output otherwise. Further, during *test short*, the network's temporal dynamics evolve in two disconnected subnetworks, independently coding for the two contexts. Both the active selection and the independent network dynamics become apparent when calculating the MDS and RDM of the representations in the conjunctive layer: the two contexts are encoded in two orthogonal subspaces and are not ordered according to their value (Figure S3H). Therefore, the REMERGE model's coding scheme is substantially different from the overlapping and elongated neural codes observed in BOLD during *test short* (cf. Figures 3B and 3C).

In order to switch the networks behaviour from *test short* to *test long*, an additional set of weights is introduced to the weight matrices W_1 and W_2, which is connecting the two subnetworks (Figure S3E). Again, the model accurately predicts the human choice matrix (Figure S3G). However, analysing the MDS and RDM of the conjunctive layer reveals that the REMERGE model employs a hierarchical code in contrast to the elongation code observed in BOLD (Figure S3H, cf. Figures 5A and 5B).

If embeddings are adjusted incrementally across the *test long* period by mentally replaying items not experienced during train long (e.g. 8 and 6) and using the evoked error to learn offline. This model implies that performance will generically improve across the course of *test long*, but it also makes a more specific prediction: that this improvement should be most pronounced for those item pairs whose relative rank is changed by the information provided during *train long* (i.e., item pairs that cross contexts). We would expect this benefit to be greater than (say) for those item pairs that have already been extensively trained during train short (those within contexts), which will already be correct from the start of *test long* because inferences can be based on learning that occurred during *train short*.

Fortunately, our design cycled through all 144 (12 x 12) item pairs twice during *test long*, allowing us to compare average performance on the first and second presentation. We plot mean accuracy independently for each presentation of each item (Figures S3J and S3K). The purple, blue and cyan boxes divide the item pairs into the within-context (cyan) and cross-context (purple) item pairs, coding separately for the item pairs 6 and 7 that were trained during *train long*. We also show the mean accuracy for first and second item. As can be seen, whilst there are increases from first to second presentation, these are of similar magnitude for the within-context and between-context item pairs. Accordingly, there was a main effect of presentation ($F_{1,33} = 5.04$, $p = 0.032$) but no interaction between presentation and item pair type ($p = 0.233$). The main effect of item pair type that is visible in the plot (different intercept for each line) is driven by difference in the fraction of equal rank items (e.g. item 4 with item 4) for which there is no correct answer.

Another way to conduct this analysis is to directly compare those cross-context item pairs for which the rank changes after *train long* relative to those for which it does not (Figures S3L and S3M, compare the purple and cyan triangles). As would be expected, items that change rank with *train long* elicit more errors overall. However, as can be seen from panel D, these item pairs improve from first to second presentation at approximately the same rate, and once again there was a main effect of presentation type ($F_{1,33} = 15.9$, $p < 0.001$) but no interaction with item pair type ($p = 0.965$).

One might question why there is improvement from first to second item in the first place, but it seems plausible that generic improvements in performance occur across the block as participants become more familiar with the task, as similar increases were seen during *test short*.

Cross task normalization

Since an elongated brispiness axis was observed during both *test short* and *test long* in ROIs from our number localiser, we asked about the relationship between these representations *across tasks*. To do this, we fit a GLM to data jointly across all three experiment types. As with those reported in the main text, this GLM used delta functions convolved with the canonical haemodynamic response function (HRF) and time-locked to trial events. The design matrix for this model included 30 regressors, one for each of the Arabic digits or objects locked to stimulus onset, as well as 6 additional nuisance regressors for head motion. We then constructed RDMs as in the main text to assess the similarity of evoked neural patterns, this time focusing on how patterns were related *across* the experiments.

We hypothesized that neural populations could code for either ground truth magnitude, in which case we would expect the *test long* axis to be twice as long as in *test short* and the number localiser (RDM_{none}, Figure S6B), or for relative magnitude, meaning the axes in all three experiments should be compressed to the same length ($RDM_{sub,div}$, Figure S6B). These were computed using ground truth object ranks for *test short* and *test long*, call them $rank_{test\ short}$ ([1:6]) and $rank_{test\ long}$ ([1:12]). "None" was constructed by computing the distances between raw, ground truth ranks, i.e. under the assumption that those ranks were not normalized. "Sub"

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

was constructed with the assumption that the vectors were normalized by subtracting by the mean element (e.g. $rank_{test\ short}$ – mean($rank_{test\ short}$)) effectively centring them with respect to each other before computing the distances. Finally, "sub, div" was constructed by normalizing the vectors both subtractively and divisively (e.g. $rank_{test\ short}$ – mean($rank_{test\ short}$)/max($rank_{test\ short}$)), which centers the vectors and equalizes their length. We found evidence for this normalised coding scheme in both PPC and dmPFC (Figure S6B; one-sided t-tests on Pearson correlations: $RDM_{sub,div}$ t_{33} = 3.86, 6.29, p's<0.001; RDM_{none} t_{33} = -3.07, -0.5, p's>0.95; RDM_{sub} t_{33} = -1.8, 0.15, p's > 0.6; PPC and dmPFC respectively).

Neural network simulations

We implemented a two-layer feedforward neural network in Python with coupled input weights to study the computational underpinnings of the task, its solutions and possible failure modes. All simulations were run for $n = 20$ random seeds and plots show averages across seeds. The network received two one-hot vectors, coding respectively for the object on the left (*xa*) and right side (*xb*) of the screen (a one-hot vector for object *i* has zeros everywhere except at the *i*-th position which is equal to one). The one-hot vectors were then propagated forward by the coupled weights *W1* and *-W1* respectively, followed by a rectified linear unit (*ReLU*) to create the hidden layer representation of 20 neurons. Finally, to mimic the binary output produced by humans responding more or less, the hidden layer representation was projected onto a single output value using the readout weights w_2:

$$\hat{y} = w_2 \mathrm{ReLU}\,(W_1 x_a - W_1 x_b)$$

We designed the model to produce a binary output to mirror how humans responded ("more" or "less"). The simplest way to do this is to use a single output unit for which values $\hat{y} > 0$ signal "more" and values $\hat{y} < 0$ signal "less." To assuage any concerns surrounding this choice in architecture, we additionally ran all of our simulations using a model with 2 output nodes and found that this choice had no impact on our simulation results (Figures S4D–S4G; See STAR Methods on Neural Network Representations). This implies that the one-dimensional solution observed in the hidden layer is not due to our using a single readout neuron. Nor is it the case that successful transitive-inference task performance necessitates the observed one-dimensional solutions. For example, when neural networks were initialized with large starting weights (in the "lazy regime"; see Figure S5) they solved the task using a high-dimensional solution with quite different geometry. Indeed, behavioural changes do not need to be expressed in adjustments to the geometry of neural representations,[55] as an ordered transitive relation can be enacted by changes in the decoder alone (Figures S5E and S5F).

By coupling the input weights, we ensured that the hidden layer representation of each object was independent from the position on the screen. Note that due to objects being represented as one-hot vectors, hidden representations of objects are independent from each other, i.e. the *i*-th column of the weight matrix W_1. Further, due to the symmetry in the first-layer weights, if the two one-hot vectors encode the same object the network's output is zero.

The network was optimised using stochastic online gradient descent

$$\Delta W = -\eta \frac{\partial \mathcal{L}}{\partial W}$$

on single pairs of objects, i.e. a batch size of 1, with learning rate $\eta = 0.05$ on the mean squared error (MSE; we used MSE loss rather than cross-entropy to avoid problems associated with saturating outputs) between the network's output $\hat{y}$ and the target values $y = 1$ for $i_a > i_b$ and otherwise $y = -1$:

$$\mathcal{L} = \frac{1}{2}(\hat{y} - y)^2$$

Since inputs to the network were two one-hot vectors, in each training step only two columns of the first layer weight matrix W_1 were updated, we denote these two column vectors by Δw_{1a} and Δw_{1b}.

Synaptic weights were initialised from a zero-centered Gaussian distribution with standard deviation $\sigma = g * \sqrt{1/fan\_in}$ where $g = 0.025$ and $g = 1$ in hidden and readout layers respectively. The hidden layer weights were initalised to small values to encourage a low- dimensional ("rich") solution.[31] We employed a training procedure very similar to that used for human subjects. Networks were first trained for 8 cycles, where each cycle was comprised of 120 trials (or 8 blocks of 60 trials per each per context) - leading to a total of 960 steps of gradient descent training. Subsequently, we performed 20 training steps on the two objects of the boundary condition. Note that like humans, despite being trained exclusively on adjacent items, neural networks learned faster and performed better on non-adjacent items.

Learning relational certainty

In order to recover the rapid *knowledge assembly* observed in humans, we adapted vanilla SGD, by applying mutual updates on synaptic weights W_1 based on the pairwise certainty that two object representations bear an accurate relation to one another in

embedding space, and in addition, correcting for potential drift in the readout weights w_2. For a given trial t with inputs $x_{a,t}$ and $x_{b,t}$ we compute the certainty value as a sigmoidal function of the loss incurred $\mathcal{L}_t$, with the slope α and the bias β of the sigmoid as potentially free parameters (here we set them to $\alpha = 1000$ and $\beta = 0.01$)

$$\Phi(\mathcal{L}_t) = \frac{1}{1+\exp(\alpha(\mathcal{L}_t - \beta))}$$

Pairwise certainty values were then stored in the certainty matrix A as an exponential moving average

$$A_{ba}^{t+1} = A_{ab}^{t+1} = (1-\gamma)A_{ab}^{t} + \gamma\Phi(\mathcal{L}_t)$$

where the free parameter γ determined how quickly old values were discounted.

In addition, to infer certainty values for pairs of items that were not presented to the network, a fraction of the certainty values for item i_b were added to the certainty values of i_a.

$$A_{a*}^{t+1} = A_{*a}^{t+1} = (1-\gamma)A_{a*}^{t} + \gamma\Phi(\mathcal{L}_t)A_{b*}^{t}$$

and vice versa. Note that the certainty matrix is symmetric and therefore rows Aa^* are identical to columns A^*a. These row-wise updates followed the heuristic: If item a is correctly related in embedding space to item c and item a is correctly related to item b, then infer that item b is also correctly related to item c. While the discount factor gamma for the pairwise certainty values and to spread certainty values across items could be assigned independent values to let them operate on two different time scales, we empirically observed that a shared value is sufficient to model the observed phenomena.

Synaptic weights were then mutually updated by outer products:

$$W_1^{t+1} = W_1^{t} + \Delta w_{1a}\left(A_{a*}^{t+1}c_a\right)^T + \Delta w_{1b}\left(A_{b*}^{t+1}c_b\right)^T$$

where Aa^* denotes the a^{th} column of the certainty matrix. Note that ca and cb are vectors of scaling factors to correct for drift in the readout weights as follows:

$$c_a = \frac{\Delta w_{1a}^T w_2 + \Delta w_2^T(w_{1a}+\Delta w_{1a}) - W_1^T \Delta w_2}{\Delta w_{1a}^T(w_2+\Delta w_2)},$$

In order to perform a SGD step on the items currently presented to the network (i.e. the a-th and b-th column of W_1) the a-th and b-th entry of A_{a*}, c_a and A_{b*}, c_b respectively are set to 1. Importantly, gradient updates to the representation of item b are mutually applied to all other items ia $\neq$ ib, but scaled by certainty Aa,b.

Fitting human choice matrices

To fit human choice matrices we applied a sigmoid function to the linear neural network output

$$\sigma(\hat{y}) = \frac{1}{1+\exp(-s\hat{y})}$$

We fit different parameterisations separately to choice matrices for high and low performers (defined by a median split, as in Figure 4). For each, we performed a grid search on combinations of γ in range 0 to 1 and s in range 0.01 and 100 (both in $\log_{10}$ units) and mapped the resulting deviation between predicted and observed choice matrices for that participant group (Figures S4A and S4B). Because neural network models are stochastic, we repeated the simulations for 20 random initial seeds and averaged the resulting deviance for fitting. Low performers were fit well with $s \approx 1$ and had a U-shaped relationship with accuracy for varying γ, leading to two local minima, such that values that were close to zero (vanilla SGD) and close to one both resulted in a failure to stitch information appropriately (Figure S4B). For low γ, the algorithm fails to acquire certainty and thus does not perform mutual updates, behaving like vanilla SGD. Similarly, for large γ, the certainty matrix is rapidly updated, such that the boundary items form a high certainty cluster separate from the rest of the items (Figure S4C). In this case, after few mutual update steps that partly disentangle the representations, the certainty matrix rapidly approaches zero for non-boundary items, again leading to SGD-like updates on boundary items only. Behaviourally, these two failure modes can be interpreted as either failing to relate items in the two conditions during after boundary training for low γ or by relating the items of the boundary condition as independent group during *train short* for large γ. On the other hand, high performers had a single minimum for $s > 2$ and $\gamma \approx 0.1$ (Figure S4B).

Neural network representations

To confirm that the one-dimensional solution our neural network simulations converged to cannot be attributed to our choice to use one output node, we reran all of our simulations with a model with 2 output nodes. As long as training starts from small initial weights, this architecture is formally equivalent to a network with a single readout, as the parameters of the readout layer converge to identical

Please cite this article in press as: Nelli et al., Neural knowledge assembly in humans and neural networks, Neuron (2023), https://doi.org/10.1016/j.neuron.2023.02.014

values with inverted signs. Thus, using 2 output nodes has otherwise no impact on our simulation results, and results during *test long* using a network trained with 2 output nodes fit to good and poor performers on *test long* (Figures S4D–S4G).

We also showed that the low dimensional representations found in the hidden layer of the neural network are not an inevitable consequence of successful task performance (Figure S5). Using the architecture reported in the main text (Figure 2) when the *test short* and *test long* neural network simulations are started from large initial weights ("lazy" regime), the network converges to a solution in which object representations are distributed over the high-dimensional neural manifold in an unstructured way that does not align with the underlying hierarchy (Figures S5D and S5E).