Neural state space alignment for magnitude generalisation in humans and recurrent networks

Hannah Sheahan†*, Fabrice Luyckx*, Stephanie Nelli, Clemens Teupe and Christopher Summerfield†

Department of Experimental Psychology
University of Oxford
Oxford, UK

* equal author contribution

† to whom correspondence should be addressed:
hannah.sheahan@psy.ox.ac.uk
christopher.summerfield@psy.ox.ac.uk

Abstract

A prerequisite for intelligent behaviour is to understand how stimuli are related and to generalise this knowledge across contexts. Generalisation can be challenging when relational patterns are shared across contexts but exist on different physical scales. Here, we studied neural representations in humans and recurrent neural networks performing a magnitude comparison task, for which it was advantageous to generalise concepts of “more” or “less” between contexts. Using multivariate analysis of human brain signals and of neural network hidden unit activity, we observed that both systems developed parallel neural “number lines” for each context. In both model systems, these number state spaces were aligned in a way that explicitly facilitated generalisation of relational concepts (more and less). These findings suggest a previously overlooked role for neural normalisation in supporting transfer of a simple form of abstract relational knowledge (magnitude) in humans and machine learning systems.

Introduction

Humans can think and reason in ways that abstract over the physical properties of the world[1,2]. For example, we understand that cheetahs and space rockets can both move "fast" even though animals and vehicles belong in different semantic categories and do not look alike. Cognitive scientists have long built theories about how humans learn concepts and reason abstractly but much less is known about their neural representation[3,4]. One view is that conceptual knowledge relies on neural ensembles that code for relations among stimuli but are invariant to their physical properties[5-13]. Recent evidence hints that when sets of stimuli share relational structure across contexts, they are embedded on parallel low-dimensional neural manifolds, so that a linear decoder learned in one context can be readily repurposed for another[14-18]. By aligning neural state spaces between contexts in this way, one can generalise relational knowledge, for example applying a criterion that distinguishes fast and slow animals to discriminate fast and slow vehicles, such as space rockets and bicycles (**Fig. 1a**). The neural geometry implied by this coding scheme thus offers a theory for how humans engage in abstract forms of reasoning that involve the use of analogy and metaphor[19].

However, there is significant challenge for relational generalisation that we call the *mapping problem*. The mapping problem occurs when stimuli are analogously related across contexts, but in one context the structure is rotated, rescaled or otherwise misaligned with respect to the other. To illustrate, we might consider both a record-breaking sprinter and a marathon champion to be "fast" runners, but in one case this might entail running one hundred metres in less than ten seconds, and in the other a marathon in under two hours (**Fig. 1b**). For generalisation to be effective, we need to form a concept of "fast" that is not tied to a specific physical value (e.g. in m/s), but that encodes relative speed in each respective context. Understanding how these abstract sorts of invariances are acquired in either biological or artificial neural networks, however, remains a challenge to researchers in neuroscience and machine learning alike.

Here, we use simulations with recurrent neural networks and neural recordings from humans to ask how stimuli with a common relational structure across contexts were represented in neural state spaces, with a focus on the mapping problem. The task involved comparing the magnitude of arbitrary, physically dissimilar stimuli that were sampled from one of three overlapping ranges (contexts). For humans, we used symbolic numbers (Arabic digits) as stimuli because adults have already learned that they denote positions on a one-dimensional manifold (number line) with visual symbols that are physically dissimilar in arbitrary ways. For neural networks, we use arbitrary non-overlapping (one-hot) inputs whose assigned magnitude could be inferred via supervision signals during training. When we conducted multivariate analyses on neural signals or hidden unit activations observed during the task, using dimensionality reduction to visualise the structure of the neural state spaces, we observed neural state space alignment across contexts in both humans and neural networks. In both model systems, numbers are organised according to their magnitude onto three parallel, equidistant neural manifolds (number lines), one for each context. Moreover, in both systems these manifolds were compressed (divisively normalised) and centred (subtractively normalised) so that numbers that denoted "more" or "less" could be linearly discriminated along a single dimension irrespective of their context. In other words, without being explicitly regularised to do so, neural networks autonomously learned to align representations in a way that supports

generalisation between contexts. In doing so, they learned to represent numbers with a neural geometry that matched that in the brains of human participants. This suggests a hitherto overlooked role for neural normalisation[20] in supporting learning and transfer of relational knowledge.

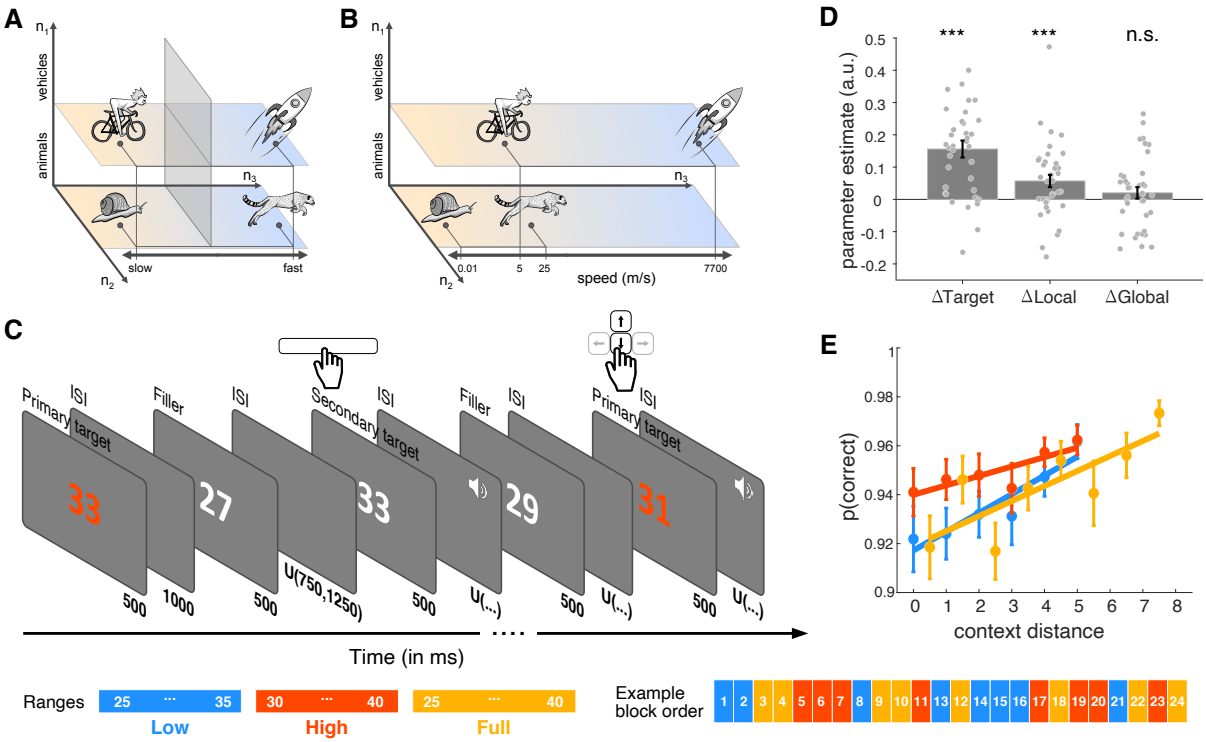

Fig 1. The mapping problem, the task and human behavioural results. (A) A sketch of aligned neural representations of stimuli in different contexts (e.g. animals, vehicles), in a 3-dimensional neural space spanned by the firing rates of 3 hypothetical neurons [n_1, n_2, n_3]. One of these neurons encodes magnitude (n_3) while another encodes context (n_1). In this example, a decoder (grey vertical plane) trained to discriminate the relative speed of bicycles and space rockets could be repurposed to distinguish snails from cheetahs. Such decoder generalisation permits context-general inferences, such as saying that the space rocket and the cheetah are both "fast" even though space rockets and cheetahs belong in different categories and do not look alike. (B) In contrast, a coding scheme based on an absolute metric, such as numerical speed in m/s, positions stimuli that are in different contexts but with the same within-context relations, very differently. (C) Schematic of the task performed by humans. The primary task was to compare the relative magnitude of primary target numbers (coloured) and indicate a binary "more" or "less" response using the up/down keys. Numbers were drawn from three temporally blocked ranges (contexts): a low range (blue), a high range (red) and a full range (golden). Blocks of trials were presented eight times per context, and the order of blocks was pseudorandom. (D) Parameter estimates from logistic regression of human accuracy against the target distance (Δ Target), local context distance (Δ Local) and global context distance (Δ Global). Dots show individual subject fits. (E) Human accuracy increases as a function of context distance for each context. Error bars show s.e.m in panels D and E.

Results

The task required agents to compare the relative magnitude of a current and previous target (number) item, indicating "higher" or "lower" with a binary response. Magnitudes were drawn from different ranges in three temporally blocked contexts, which were repeated in pseudorandom order. For humans, we used Arabic digits, a class of visual stimulus that our adult participants had already learned to associate with magnitude. Target stimuli occurred in each of three contexts, which spanned digit ranges 25-35 (low), 30-40 (high) and 25-40 (full context) in distinct blocks. Targets were signalled by a distinctive cue that was also indicative of the context (font colour of the number), and responses were made with two fingers of the right hand. Interspersed between targets were 2-4 "filler" numbers (white for humans) that were drawn from across the full range. A secondary task that did not require higher vs. lower magnitude comparison was performed on these stimuli by the humans, namely pressing the space bar (with the left hand) when a filler exactly matched the previous target (∼12% of fillers). A visualisation of the task that humans performed is shown in **Fig. 1c**. For neural networks, inputs were *a priori* arbitrary and their magnitude was learned from supervision signals.

Human behaviour. We first focus on the primary (magnitude comparison) task which healthy human adults (n = 38) performed with a mean accuracy of ∼94±4% and response times that averaged 675ms. We begin by offering a normative account of this task. An observer with perfect memory can ignore the context provided by recent numbers and simply maintain the previous item for comparison with the current item (e.g. compare 31 with 33 in **Fig. 1c**). However, for an agent with imperfect memory, the context in which numbers occur becomes relevant. For example, a participant who forgets the 33 but responds "less" to 31 will most likely be correct in the high context[21,22]. Critically however this latter strategy will be more effective when participants have some notion of whether each target is "more" or "less" relative to the local context within the current block, because a number may be low in one context but high in another (for example 31 is "more" than the local average in the low context, and so would prompt the incorrect answer). This effect is well known to lead to estimation biases that depend on the local history and have been associated with neural signals in the parietal cortex[23].

Here, behavioural analyses revealed that participants used memory and contextual information when making decisions (**Fig. 1d**). Logistic regression indicated that accuracy was predicted by both the disparity between current and past target number (*target distance*; $t_{37} = 5.9$, $p < 0.001$) and the distance between the target number and the median number in the current block (*local context distance* $t_{37} = 3.0$, $p < 0.004$). After accounting for these sources of variance, however, distance to the overall median number (across all blocks; *global context distance*) had no impact on performance ($p > 0.1$). We plot accuracy as a function of local context distance for the low, high and full contexts in **Fig. 1e**.

Neural network behaviour. For comparison, we trained recurrent neural network models (RNNs; n = 10) to perform an equivalent task on symbolic (one-hot) inputs (see **Fig. 2a**). As for human experiments, inputs were sampled from different ranges in blocks of 120 trials. Each network had a single recurrent layer and a single feedforward hidden layer and was trained to minimise errors on the task. We encouraged the network to find context relevant in the same

way as humans using a virtual inactivation (VI) approach. This involved setting the input to zero for a fraction ε of targets, as if stimuli were missed or forgotten, as they might be by a human with imperfect memory. Applying VI during training obliged the network to learn to use the context (i.e. local average of numbers) as a cue for magnitude comparison, because when numbers were lost to memory, knowledge of the context improved accuracy. Applying VI during test allowed us to measure this impact of context on responding independent of the target distance. After $\sim 10^6$ training steps, networks converged to near perfect accuracy on held out stimulus sequences irrespective of whether virtual inactivation was applied during training at $\varepsilon_{train} = 0.1$ (99.80 ± 0.05%) or not (99.96 ± 0.03%).

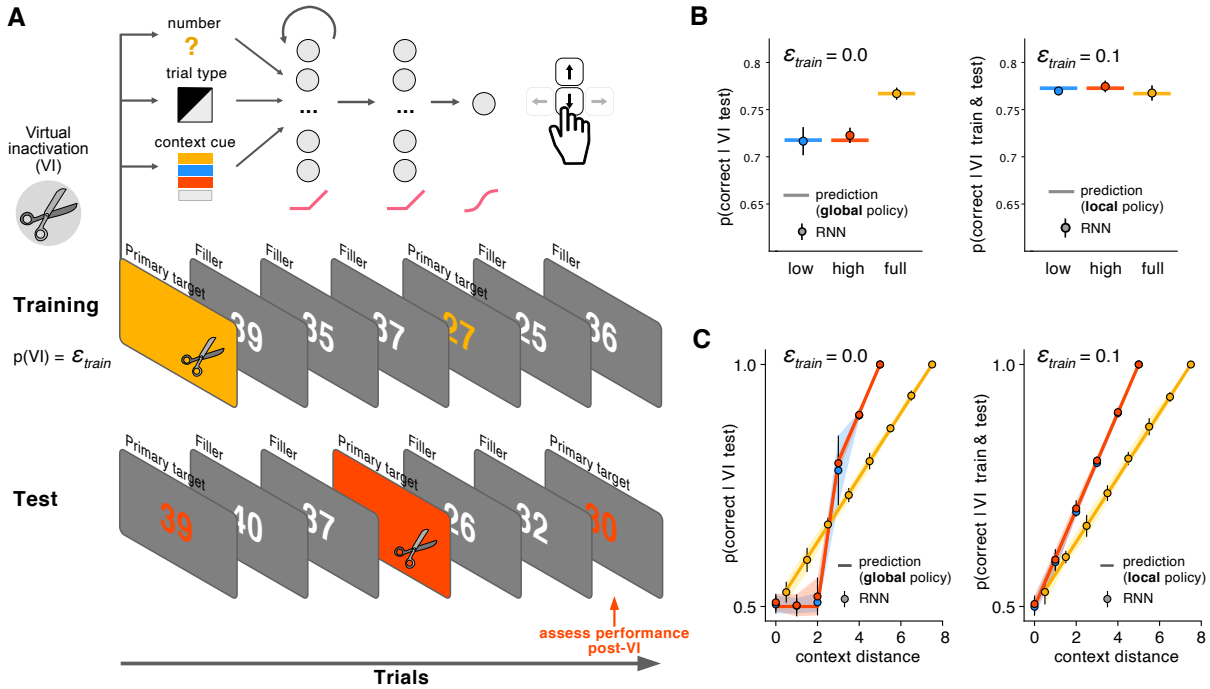

Fig 2. Recurrent neural network architecture, virtual inactivation and network behaviour. (A) Architecture of a recurrent neural network model trained to perform the same magnitude comparison task as humans. At each step the network received symbolic (one-hot) concatenated inputs indicating the number, trial type (filler or primary target), and context (zero for filler trials) of the current trial and indicated a binary "more" or "less" response with a single output node. The activation functions for each layer are shown in pink. Virtual inactivation (shown here with scissors) involved setting the number input on a given trial to zero, as if the number had been hidden or missed by the network. On each training trial, the network inputs corresponding to the primary target were virtual inactivated with probability ε_{train}. At test, each primary target was virtually inactivated in turn and performance assessed on the primary target that followed. (B) Performance in each context predicted by an agent who optimally uses either the global (left, horizontal bars) or local (right, horizontal bars) context as a cue to magnitude comparison following a hidden (VI) target. RNNs (filled circles) without VI during training (left) and with VI during training (right) perform at exactly the levels predicted by the global and local agents respectively. (C) RNN accuracy (filled circles) following a VI shown as a function of local context distance for each context. RNN accuracy matches the predictions (coloured lines) of an agent using the global context (left) when the RNN was without VI in training (left). In contrast, the performance of RNNs with VI during training match predictions of an agent using the local context (right). Context colouring: low (blue), high (red), full (golden). Error bars in B and C show standard deviation across different random model initialisations and datasets.

Assessing RNN performance on a subset of test trials for which the previous target input was virtually inactivated ($\varepsilon_{test} > 0$) offered information about how they were performing the task, both with and without VI during training. On these test trials, memory for the previous target is erased so that a network that ignores context will perform at chance. As a point of comparison, we computed the performance ceiling displayed by an agent who optimally uses either the local context (i.e. numbers within the current block) or global context (i.e. all numbers) as a lone cue to magnitude comparison. Networks that had undergone VI during training achieved near equivalent accuracy in low, high and full conditions, performing at precisely the level of an optimal agent using the local context (**Fig. 2b**). By contrast, networks that enjoyed perfect memory during training performed better on the full than high or low conditions, matching predicted performance levels for an ideal agent that used only the global context. In other words, whereas all networks learned to use context as a cue to respond, only when capacity was limited during training did the networks learn to exploit the range of numbers in the current context to maximise their accuracy; we confirmed this observation statistically by comparing the residuals of the fit to either model (both t-values > 16; $p < 0.001$ in both cases). In this and all subsequent statistical analyses, we treat the individual network (n = 10; each with their unique initialisation and stimulation sequence) as the unit of replication. In **Fig. 2c** we plot the network performance as a function of local context distance in each condition (low, high, full) for both networks trained with $\varepsilon_{train} \in [0, 0.1]$ (see Fig. S1 for performance of networks trained with $\varepsilon_{train} \in [0.2, 0.3, 0.4]$). When VI is applied, performance is a linear function of context distance in each condition, whereas without VI, performance remains at chance for context distances of ≤ 2 in the low and high conditions, because for these instances a comparison to the global average fails to offer the correct answer.

Neural network state space. With these results in hand, we used representational similarity analysis (RSA) combined with multidimensional scaling (MDS) to visualise the embedding of numbers and contexts in the network neural state space. Focussing on fully trained networks, we computed correlation distance among hidden unit activations evoked by each number in each context and plotted the resulting neural states in three dimensions. In **Fig. 3a-b** we show the projection of each stimulus in each context (coloured dots) into a space spanned by the first three dimensions of the activations from the RNN hidden layer. As can be seen, the network learned to organise numbers according to their magnitude onto three parallel lines, one for each context. This occurs both with and without VI during training. However, consistent with the finding that local context is a more salient cue for responding when $\varepsilon_{train} = 0.1$, these context-specific number lines were more widely separated when VI was applied at training, as revealed by a statistical comparison of the Euclidean distance among their centroids ($t_9 > 25$, $p < 0.001$).

Critically, however, it can also be seen that the neural representation of number in each context was compressed and centred so that the three lines span a common distance in one of the three dimensions of neural state space. This means that for the network, common positions on each line do not denote specific numbers, but rather encode abstract quantities corresponding to “more” or “less” within each context. Forming an abstract concept of “more” or “less” presumably facilitated the readout of signed context distance, permitting the network to solve the mapping problem, which may be particularly useful under the VI manipulation where capacity was limited. We note however, that this normalisation only occurred when the

network received different ranges of numbers in three temporally distinct contexts (blocks), matching the task performed by humans. When inputs from different contexts were interleaved (but still signalled with a unique cue) the network failed to normalise, representing the signals in their native (absolute) frame of reference (**Fig. 3c-d**).

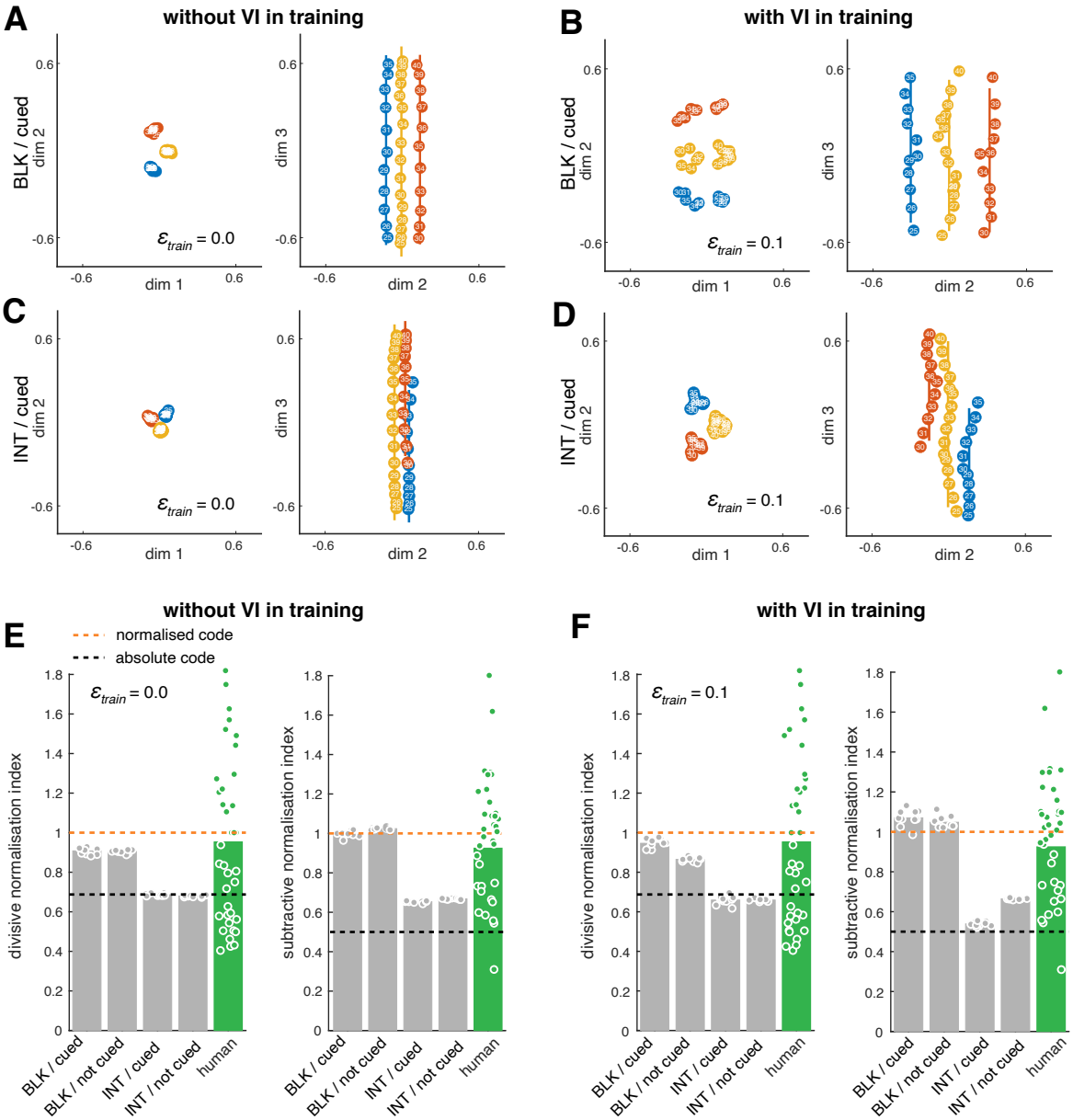

Fig 3. Neural network activations and normalisation metrics. Neural network models were trained either with or without virtual inactivation (VI), with either temporally blocked (BLK) or interleaved (INT) contexts, and with contexts either cued explicitly in the input (cued) or not (not cued). (A-B) Multidimensional scaling (MDS) visualisation of hidden unit activations in networks that were blocked and for which context was explicitly cued. (C-D) MDS activations from networks for which context was explicitly cued but for which contexts were temporally interleaved. Filled circles show the state space representation of stimuli in the low (blue), high (red) and full (golden) contexts; inset white numbers indicate the corresponding symbolic number input to the network. Coloured lines show the best fit lines model for quantifying the representational geometry. (E-F) Divisive and subtractive normalisation

indices from fits to the representations which resulted from each network training condition (grey), and from human neural recordings (green). Grey dots are fits to individual random model initialisations and datasets. Green dots are fits to individual human subjects (see below). Dashed horizontal lines are provided as references for the divisive and subtractive normalisation indices expected under fully normalised (orange) and fully absolute (black) coding schemes.

To quantify these observations, we fit a state space model to the RNN data RDM (n = 10). The model was fit by varying the angle and length of three parallel number lines within a three-dimensional state space. Gradient descent was used to minimise their discrepancy with the neural state space data. This model, whose identifiability we verified using a parameter recovery approach (**Fig. S1**), allowed us to compute and compare the best-fitting line lengths in each context. Under a full (divisive) normalisation scheme, all three lines should have the same length (divisive normalisation [DN] index = 1), whereas without normalisation, the ratio of numbers in the low/high to full conditions should be ~ 0.69, reflecting their relative range (ranges of 16 vs. 11).

The relative line lengths are plotted on the leftmost bars **Fig. 3e** (without VI during training) and **Fig. 3f** (with VI). As can be seen, when magnitude ranges were separated into temporally distinct contexts (denoted BLK), the DN index approaches 1, irrespective of whether contextual cues are offered, and independent of whether VI was applied at training or not. When ranges were interleaved (INT), however, the DN index is ~ 0.69 or lower, indicative of an absolute code. In other words, without the benefit of blocked temporal context, the network does not use the context cue to distinguish among the different ranges. We also computed a subtractive normalisation (SN) index by comparing the offset in centroids among the high, low and full conditions along the principal magnitude coding axis. Again, blocking (but not interleaving) of contexts led to a full subtractive normalisation (SN ~ 1) whereby the centres of each line were brought into the same register, indicative of neural state space alignment. For completeness, we also fit models with the restriction that line lengths in each context should be equal (relative model) or that they should reflect the range of numbers in each context (absolute model). For (cued) BLK conditions, 10/10 RNNs were better fit by the relative model, whereas for (cued) INT conditions, 10/10 RNNs were better fit by the absolute model.

These investigations of the RNN neural state space offered several insights about how transitive orderings (e.g. numerical magnitudes) may be represented in different contexts. First, networks embedded arbitrary inputs onto parallel number lines in a way that reflected their transitive ordering. When different ranges of numbers occurred in different temporal contexts, these number lines were normalised so that each embedding space stretched from "less" to "more" within that context. When VI was applied at training, the number lines spread out in dimensions orthogonal to the magnitude axis, and it was presumably this which facilitated the use of context-specific rather than context-general information to help solve the task.

Human neural data. Finally, we turn to analysis of neural data recorded with scalp electroencephalography (EEG) whilst humans performed the numerical comparison task. Our main focus is the results of a multivariate analysis approach designed to interrogate the nature of the neural state space and its alignment across contexts. However, we also observed univariate signals with a right occipitoparietal focus that covaried positively with magnitude and with target distance in an early window (from 200-500 ms post-stimulus) and signals that

covaried with local context distance over left posterior electrodes in a later window (500-800 ms post-stimulus). These effects, which survived correction for multiple comparisons using a familywise error test, are shown in **Fig. 4c.**

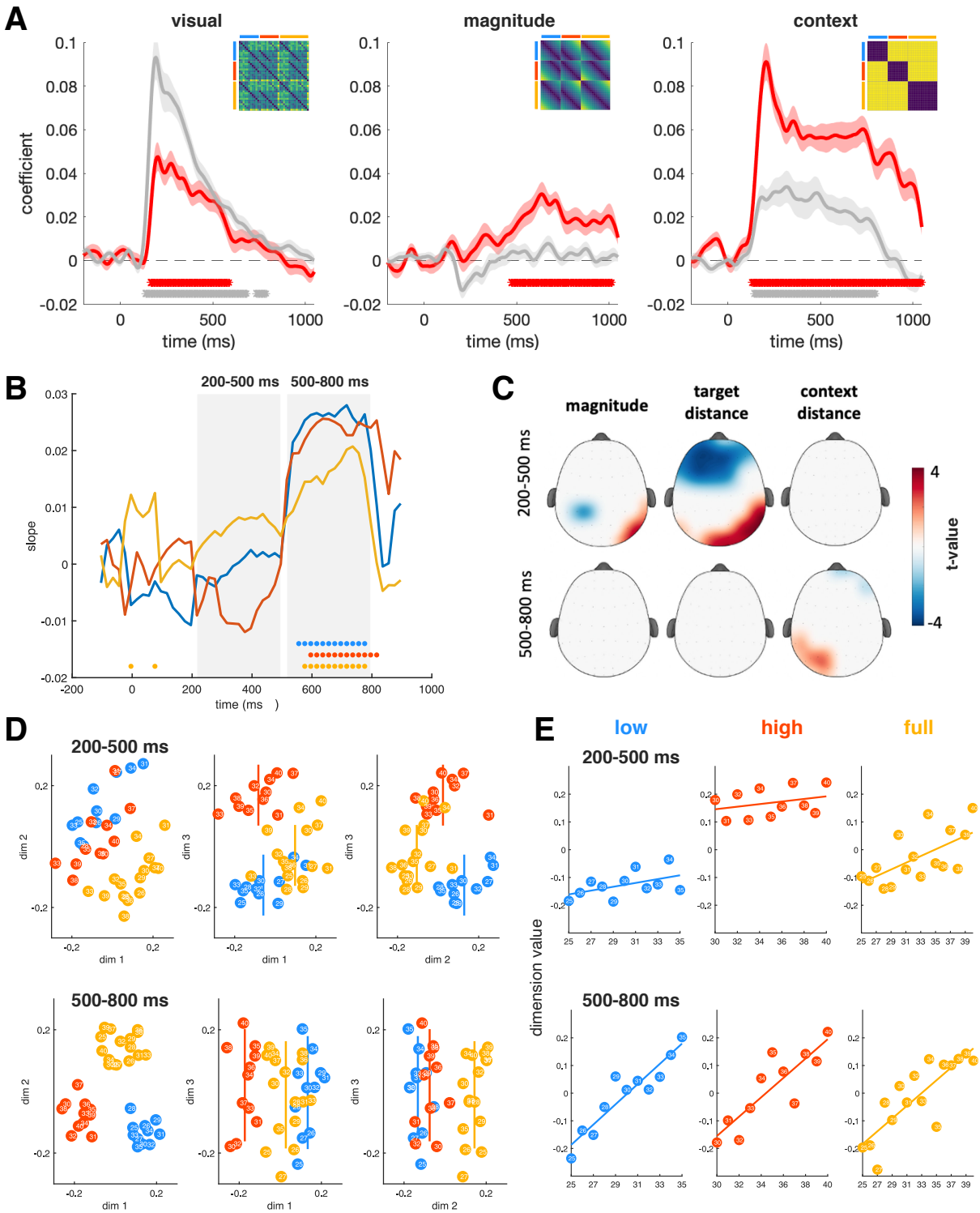

Fig 4. Human neural activity. (A) Model representational dissimilarity matrices (shown inset) labelling the symbolic number within each context were regressed against the time course of neural activity within each trial (x-axis). The time course of regression coefficients is shown for both the filler trials

(grey) and primary targets (red), and their cluster-corrected significance at each time point is indicated along the bottom of each figure with overlapping red/grey asterisks. Three coefficient time courses are shown resulting from regression against: an RDM indicating the pixel-wise visual similarity between inputs (left panel), an RDM indicating the numerical magnitude of each input (middle panel), and an RDM indicating the context (right panel). (B) Slopes of the relationship between number and dimension value calculated within a sliding temporal window for each of the three contexts (blue, low; red, high; gold, full). Slopes were normalised by a shuffled control. Asterisks show reliable deviation from zero. (C) Univariate effects of magnitude, target distance, and local context distance rendered onto the scalp. Warm colours indicate positive relationship; cold colours a negative relationship. Only effects that survived correction for multiple comparisons are shown. (D) Multi-dimensional scaling visualisation of human neural activity corresponding to the primary targets shown in each context, mapped into a 3-dimensional space. Top panels, early window; bottom panel, late window. The neural codes associated with each number can be seen to form equidistant clusters according to the context in which the number was presented (left panel). Plotted lines show the best fit models for quantifying the representational geometry in each time window. (E) A complementary visualisation plotting the position of each target along the third MDS dimension against the numerical value of the target in each context, such that each x-axis is scaled to indicate "small" and "large" numbers in each context (for both time windows). Plotted lines were fit to capture this slope of this relationship for each context and time-window. The numerical value of each target can be seen to increase monotonically as a function of its position along this dimension, and the slope of the best-fitting line is approximately equal in each context for the late time window only. Filled circles show the state space representation of stimuli in the low (blue), high (red) and full (golden) contexts; inset white numbers indicate the corresponding Arabic number shown on the trial. Coloured lines show the best fit lines model for quantifying the representational geometry.

We constructed model RDMs on the basis of three variables: numerical magnitude, context, and the pixelwise similarity among inputs (Arabic digits). We then regressed these three model RDMs against a data RDM obtained from scalp EEG patterns at each timepoint post-target; for a point of comparison, we also conducted the same analysis on the filler stimuli. The resulting coefficients are shown in **Fig. 4a**. As expected, early timepoints are dominated by visual similarity, for both targets and fillers, to an approximately equal extent (left panel). However, from about 600ms the EEG signals evoked by targets carried information about both context and magnitude; the context information was weaker, and the magnitude information was absent for the fillers. We also tried to decode responses to target stimuli (which were made with two fingers of the right hand) but were unable to do so at any timepoint across the epoch (all p-values > 0.2), making it unlikely that any partial correlation with response drives these or any subsequent results.

Next, we used MDS to visualise relations among the neural codes associated with each number in each context (**Fig. 4b, d, e**). Within each time bin, we fit of a model that assumed that number relations were described by three parallel lines in neural state space. In **Fig. 4b**, we show how the slopes of those number lines vary over time using a sliding window approach and in **Fig. 4d**, we plot the first three dimensions of the neural state space in the early window (200-500 ms; upper panels) and late window (500-800 ms; lower panels). In the early period it can be seen that there is a reliable segregation of multivariate neural signals associated with each of the three contexts, with the neural distance between numbers within a context observed to be reliably smaller than those between contexts, compared to an appropriate shuffled null distribution ($p < 0.001$). Moreover, whilst there appears to be an emergence of three number

lines (rightmost panel), the slopes of these lines did not fall within the top 5% of a null distribution constructed by shuffling the neural RDMs, and it can be seen that there is no alignment among the centroids of these lines.

However, as we move into the later time bin (lower panels) we can see that just as for the RNN, the numbers spread out along parallel lines in the third dimension, whereas the three contexts themselves were separated in the first two dimensions, lying approximately at the apices of an equilateral triangle (left panel). We also see a highly reliable effect of context as quantified by a permutation test ($p < 0.001$). The third dimension seemed to code for magnitude, with the lower numbers in each context exhibiting negative scores along this dimension and the higher numbers positive scores. To test this latter contention more formally, we plotted the magnitude of each number in each context against its value on the third dimension (**Fig. 4e**). Positive slopes were recovered in each case, and the slopes of the relationship between dimension scores and numerical magnitudes were more positive than expected by chance for all three contexts, as evaluated relative to a shuffled control (all p-values < 0.001). The abscissa in **Fig. 4e** is scaled such that relatively "small" and "large" numbers are spaced equally in the three contexts, and under this scaling, it can be seen that the slope of the best-fitting line is approximately equal in each context. This suggests that as for RNNs, the neural number lines are normalised in a way that facilitates neural state space alignment. Indeed, the average DN index for humans was 0.95 ± 0.41, reliably below 1 but significantly closer to the value expected under full normalisation than no normalisation; similarly, the value for SN was 0.93 ± 0.38 (**Fig. 3e-f**).

Discussion

We studied the neural representation of number and context in humans and neural networks performing a sequential magnitude comparison task. Consistent with previous reports, we found that human decisions were guided by contextual signals as well as memory for the previous item. This implies that humans maintain an estimate of the local average number within a block to help guide responding when memory has failed. The context-related decision information ("context distance") is encoded in univariate neural signals over occipito-parietal electrodes, albeit in a rather late window beginning at approximately 500 ms post-stimulus. These findings are broadly consistent with previous reports that where imperative information is weak or absent, judgments are biased towards the central tendency of the stimulation history[21,22,24], and that (in the rodent) this contextual information is coded in parietal circuits[23]. Recurrent neural networks trained to perform the task do not naturally encode local contextual information when their memory is fully intact. However, when their memory is artificially rendered fallible during training, the networks also learn to capitalise on local contextual information to make judgments of relative magnitude in a way that closely resembles the human participants.

Our major question was the nature of the neural representation of magnitude and context in humans and neural networks. We studied this by reducing the dimensionality of neural signals recorded at multiple scalp electrodes in humans, and hidden unit activations read out from neural networks trained to perform the task, visualising and quantifying the neural state spaces in which each number and context was embedded. The most striking finding is the highly conserved way that humans and neural networks represent magnitude and context in this task.

In both model systems, magnitudes are projected onto parallel neural "number lines" whereby stimuli with greater magnitude difference elicited more dissimilar neural signals in each context. Moreover, in both systems these number lines are normalised so that they span a common distance within neural state space, meaning that the ends of each line correspond to "more" and "less" – a relational quantity, rather than a specific numerical value.

A low-dimensional code for symbolic number has been reported before in M/EEG signals[15,25,26]. This neural code must be abstracted away from the physical properties of the inputs, because physical similarity among Arabic digits is not determined by their relative cardinality. Here, we used two-digit numbers, and so there was some unavoidable correlation between magnitude and physical similarity due to decade boundaries (e.g. 29 and 30), but we still observed a multivariate code for number even after regressing out pixelwise similarity among digits (**Fig. 4b**), and no such code was observed for comparable filler objects, so we think it is unlikely that putative magnitude effects are driven by visual appearance. Similarly, because we were unable to decode the response made by participants to target stimuli (with two fingers of the right hand), we think it is unlikely that putative magnitude effects are related to motor signals.

However, rather than a generic code for number, we observed that both humans and neural networks learned to additionally segregate information by context, so that rather than a single number line, we observed three parallel neural number lines. In humans, this pattern did not occur instantly but emerged gradually over the course of the epoch. Early in the epoch (e.g. from 200 ms) neural signals distinguished context itself, an effect that may have been driven in part by the different font colour in which targets occurred across blocks. During this early window we observed nascent number lines in distinct parts of state space, but without parallel arrangement. From 500ms, however, numbers were arranged into context-specific parallel number lines.

We argue that this neural geometry has at least two desirable properties. Firstly, it allows number information in each context to be kept separate. This is useful because (as discussed above) when memory is imperfect the local context provides helpful information about how to respond; if all numbers were projected onto a single line, this contextual information would be unavailable. Indeed, when we examined the state spaces of neural networks that had not experienced virtual inactivation during training, and that did not use local context distance as a cue, the three parallel number lines were much closer together. In fact, when training was interleaved so that time could not be used as an additional cue for context, they lay virtually on top of one another (**Fig. 3**). Secondly, the fact that number lines are parallel in neural state space facilitates generalisation between physically similar stimuli that share a common magnitude. This neural geometry ensures that a linear decoder trained in one context could be successfully applied to read out magnitudes in another[14]. Indeed, a long tradition emphasises that humans generalise naturally between space, time and number, by using a magnitude representation that is shared across different input modalities[27,28]. This theory also successfully predicts the existence of neurons that code for magnitude with a shared code across input domains; such cells have been identified in the parietal cortex of macaque monkeys experiencing trains of auditory or visual pulses[29]. In humans, the parietal cortex is a hub for numerical cognition[30] and a site where overlapping representations of "distance" signalled by spatial, temporal or social comparisons are observed in fMRI signals[31]. Similarly, when humans learned to rank images of animals according to their probability of paying out a

reward, shared multivariate patterns in EEG signals come to code for the number and value, as if there were a corresponding neural signal for higher numbers and higher event probabilities[15]. Together, these findings imply a general principle whereby neural state space alignment permits generalisation across contexts[14].

However, generalisation of relational information between contexts can be hampered by what we call the "mapping problem": the need to represent stimulus geometry on a common scale. For example, our task requires participants to estimate whether a number is reflective of "more" or "less" in each context in case memory is imperfect. In recurrent neural networks, we can see that this is achieved as the neural number lines are both subtractively and divisively normalised so that they span a common distance along one axis of neural state space, meaning that this axis runs from "more" to "less" rather than simply indexing numerical value. Of note, no normalisation is explicitly engineered into our networks; rather, they autonomously learn to encode numbers in this context-normalised fashion because doing so enhances performance.

In humans, we saw evidence for the same normalisation process: the neural number line in the "full" condition was compressed to the same length in neural state space as that for the "low" and "high" conditions (divisive normalisation), and the centroids of the numbers were aligned along one dimension (subtractive normalisation). It has long been known that biological brains are prone to normalise incoming sensory signals, with neural activity typically expressed relative to the local average response[20]. Explanations for this ubiquitous motif often focus on the need to make efficient use of computational resources[32]. For example, divisive normalisation helps make efficient use of the dynamic range of a neuron or population, and subtractive normalisation can "explain away" redundant information in an input signal[33]. Here, we suggest a complementary role for normalisation in aligning neural codes to facilitate generalisation across contexts. Other studies offer hints of a comparable process. For example, when monkeys reproduced either long or short temporal intervals (contexts), the underlying neural dynamics in the medial prefrontal cortex were found to 'stretch' or scale in time according to context[34].

Our study leaves a number of questions unanswered. Firstly, we focus here on a very simple form of relational abstraction – the transitive ordering implied by numerical magnitudes. It remains to be seen whether the results described here generalise to more complex relational structures. Secondly, because of the limited spatial resolution of EEG, and the use of multivariate methods that relied upon electrodes from across the scalp, we are unable to say much about the neural locus of our effects. On the basis of the univariate findings (which highlight occipito-parietal electrodes) and past work[23], we think that the parietal cortex is a likely candidate for representing magnitude in parallel, context-specific lines. However, we cannot say this with confidence on the basis of the current data. Finally, we note that normalisation was not ubiquitous across the human cohort. Within the late time window, a relative model fit better in a majority of participants but was not ubiquitous. It would be interesting to conduct more targeted tests to ask whether participants that normalise more effectively also generalise more readily. These caveats aside, however, we believe that these findings offer insights into the neural coding and generalisation of the concept of magnitude, a basic form of abstraction for humans[27].

References

1 Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav Brain Sci* **40**, e253, doi:10.1017/S0140525X16001837 (2017).

2 Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279-1285, doi:10.1126/science.1192788 (2011).

3 Murphy, G. L. *The Big Book of Concepts*. (MIT Press, 2002).

4 Markman, A. B. & Gentner, D. Thinking. *Annu Rev Psychol* **52**, 223-247, doi:10.1146/annurev.psych.52.1.223 (2001).

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

6 Baram, A. B., Muller, T. H., Nili, H., Garvert, M. & Behrens, T. E. J. Entorhinal and ventromedial prefrontal cortices abstract and generalise the structure of reinforcement learning problems. *BiorXiv preprint* (2019).

7 Bellmund, J. L. S., Gardenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: Spatial codes for human thinking. *Science* **362**, doi:10.1126/science.aat6766 (2018).

8 Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490-509, doi:10.1016/j.neuron.2018.10.002 (2018).

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

10 Collins, A. G. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol Rev* **120**, 190-229, doi:10.1037/a0030852 (2013).

11 Summerfield, C., Luyckx, F. & Sheahan, H. Structure learning and the posterior parietal cortex. *Prog Neurobiol*, 101717, doi:10.1016/j.pneurobio.2019.101717 (2019).

12 Doumas, L. A., Hummel, J. E. & Sandhofer, C. M. A theory of the discovery and predication of relational concepts. *Psychol Rev* **115**, 1-43, doi:10.1037/0033-295X.115.1.1 (2008).

13 Alfred, K. L., Connolly, A. C., Cetron, J. S. & Kraemer, D. J. M. Mental models use common neural spatial structure for spatial and abstract content. *Commun Biol* **3**, 17, doi:10.1038/s42003-019-0740-8 (2020).

14 Bernardi, S. *et al.* The geometry of abstraction in hippocampus and pre-frontal cortex. *BiorXiv preprint* (2019).

15 Luyckx, F., Nili, H., Spitzer, B. & Summerfield, C. Neural structure mapping in human probabilistic reward learning. *Elife* **8**, doi:10.7554/eLife.42816 (2019).

16 Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron* **98**, 1005-1019 (2018).

17 Ganguli, S. *et al.* One-dimensional dynamics of attention and decision making in LIP. *Neuron* **58**, 15-25, doi:S0896-6273(08)00168-2 [pii] 10.1016/j.neuron.2008.01.038 (2008).

18 Fitzgerald, J. K. *et al.* Biased associative representations in parietal cortex. *Neuron* **77**, 180-191, doi:10.1016/j.neuron.2012.11.014 (2013).
19 Gentner, D. Bootstrapping the mind: analogical processes and symbol systems. *Cogn Sci* **34**, 752-775, doi:10.1111/j.1551-6709.2010.01114.x (2010).
20 Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat Rev Neurosci* **13**, 51-62, doi:nrn3136 [pii] 10.1038/nrn3136 (2012).
21 Jazayeri, M. & Shadlen, M. N. Temporal context calibrates interval timing. *Nature neuroscience* **13**, 1020-1026, doi:10.1038/nn.2590 (2010).
22 Holllingworth, H. L. The central tendency of judgement. *The Journal of Philosophy, Psychology and Scientific Methods* **7**, 461-469 (1910).
23 Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368-372, doi:10.1038/nature25510 (2018).
24 de Gardelle, V. & Summerfield, C. Robust averaging during perceptual judgment. *Proc Natl Acad Sci U S A* **108**, 13341-13346, doi:1104517108 [pii] 10.1073/pnas.1104517108 (2011).
25 Spitzer, B., Waschke, L. & Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat Hum Behav* **1**, doi:10.1038/s41562-017-0145 (2017).
26 Teichmann, L., Grootswagers, T., Carlson, T. & Rich, A. N. Decoding Digits and Dice with Magnetoencephalography: Evidence for a Shared Representation of Magnitude. *J Cogn Neurosci* **30**, 999-1010, doi:10.1162/jocn_a_01257 (2018).
27 Walsh, V. A theory of magnitude: common cortical metrics of time, space and quantity. *Trends Cogn Sci* **7**, 483-488, doi:10.1016/j.tics.2003.09.002 (2003).
28 Hubbard, E. M., Piazza, M., Pinel, P. & Dehaene, S. Interactions between number and space in parietal cortex. *Nat Rev Neurosci* **6**, 435-448, doi:10.1038/nrn1684 (2005).
29 Nieder, A. Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proc Natl Acad Sci U S A* **109**, 11860-11865, doi:10.1073/pnas.1204580109 (2012).
30 Piazza, M. & Izard, V. How humans count: numerosity and the parietal cortex. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry* **15**, 261-273, doi:10.1177/1073858409333073 (2009).
31 Parkinson, C., Liu, S. & Wheatley, T. A common cortical metric for spatial, temporal, and social distance. *J Neurosci* **34**, 1979-1987, doi:10.1523/JNEUROSCI.2159-13.2014 (2014).
32 Louie, K. & Glimcher, P. W. Efficient coding and the neural representation of value. *Ann N Y Acad Sci* **1251**, 13-32, doi:10.1111/j.1749-6632.2012.06496.x (2012).
33 Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **2**, 79-87, doi:10.1038/4580 (1999).
34 Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nature neuroscience* **21**, 102-110, doi:10.1038/s41593-017-0028-6 (2018).
35 Paszke, A. *et al.* in *NeurIPS.*

Methods

Participants

Thirty-nine participants were recruited for the experiment through the recruitment system at the Department of Experimental Psychology at the University of Oxford. One participant was omitted from the analyses due to technical difficulties with the EEG equipment. All analyses were performed on the remaining 38 participants (15 female, 23 male, age = 27.11 ± SD 6.13). Participants were required to have normal or corrected-to-normal vision, with no history of neurological or psychiatric illness. Participants were compensated for their time at a rate of £10 per hour, with an additional maximum bonus of £5 determined by their performance in the task in which they achieved lower accuracy (see below). The reward contingency was added to ensure participants would pay equal attention to both tasks. Informed consent was given before the start of the experiment. The study was approved by the Medical Science Inter-Divisional Research Ethics Committee (R49432/RE001).

Experimental procedure

Stimuli were created and presented using the Psychophysics Toolbox-3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) for Matlab (MathWorks) and additional custom scripts. The tasks were presented on a 20-inch screen with a resolution of 1600 x 900, at a refresh rate of 60 Hz and on a grey background. Viewing distance was fixed at approximately 62 cm. The 'up' and 'down' arrow keys on a standard QWERTY keyboard were used as response keys for the numerical comparison task and the space bar for the number matching task.

The experiment was a dual-task rapid serial visual presentation paradigm, involving a numerical comparison task and a number matching task (see **Fig. 1**). In each block (n = 24), participants viewed a sequence of 120 two-digit numbers. Numbers relevant for the numerical comparison task ("targets") were presented in coloured font, while those relevant for the number matching task were presented in white ("fillers"). On presentation of a target number, participants were asked to compare its magnitude to that of the previous target number in the stream, responding 'up' when it was greater and 'down' when it was smaller using their right hand. Every target was followed by 2 – 4 fillers. Participants were asked to press space bar with their left hand when a filler number was identical in magnitude to the previous target. This secondary task was imposed to keep participants focussed on all numbers.

The event sequence was as follows. Each block started with the presentation of a central fixation cross (1000 ms) followed by two placeholder hashtag signs (500 ms). Each stimulus (number) was presented for 500 ms with a fixed ISI of 1000 ms after a target and a variable ISI (750-1250 ms) after a filler. Participants were free to respond from stimulus onset until 200 ms before the onset of the next stimulus (this avoided feedback signals overlapping with the presentation of the subsequent stimulus). If a response key was pressed, participants received auditory feedback for 150 ms. Correct responses were followed by a high-pitch tone and all errors resulted in a low-pitch tone. If participants failed to respond within the appropriate time window, a buzz sound was presented for 150 ms. At the end of each block, participants were informed about their percentage accuracy on both tasks.

One each block, targets (for the numerical comparison task) were drawn from a specific range. In the **Low** range context, numbers were uniformly drawn between 25 and 35, in the **High** range context between 30 and 40 and in the **Full** range context between 25 and 40. The white filler numbers always spanned the entire range between 25 and 40 irrespective of block. The experiment consisted of 8 blocks of each context. Each context was associated with a unique colour for the targets (blue: RGB = [86, 180, 233]; orange: RGB = [230, 159, 0]; purple: RGB = [230, 120, 220]) and the colour-to-context mapping was randomised between participants. The block order was randomised in order to reduce temporal similarity between numbers from the same context. Each block contained 29 targets (excluding the first coloured number of the sequence), with a probability of 12% that at least one of the subsequent white numbers matched the previous target. To prevent potential task switching costs, the white filler number immediately after a primary target never required a response, but participants were not made aware of this feature. At the start of the experiments, participants first completed 3 practice blocks of 144 trials each, one for each of the three contexts. These blocks did not count towards their final performance bonus.

EEG acquisition

The EEG signal was recorded using 61 Ag/AgCl sintered surface electrodes (EasyCap, Herrsching, German), a NeuroScan SynAmps RT amplifier, and Curry 7 software (Compumedics NeuroScan, Charlotte, NC). Electrodes were placed according to the extended international 10-20 system, with the right mastoid as recording reference and channel AFz as ground. Additional bipolar electrooculography (EOG) was recorded, with two electrodes placed on either temple for recording horizontal EOG and two electrodes above and below the right eye for vertical EOG. All data was recorded at 1 kHz and low pass filtered online at 200 Hz. All impedances were kept below 10-15 kΩ during the experiment.

EEG pre-processing

The data were pre-processed using functions from the EEGLAB toolbox (Delorme & Makeig, 2004) for Matlab and custom scripts. First the data were down-sampled to 250 Hz, low-pass filtered at 40 Hz and then high- pass filtered at 0.5 Hz. The continuous recording was visually screened for excessively noisy channels and these were interpolated by the weighted average of the surrounding electrodes. The data was then offline re-referenced to average reference. Epochs were extracted from 250 ms before to 1000 ms after stimulus onset. Epochs were baselined relative to the full pre-fixation time window. Epochs containing atypical noise (such as muscle activity) were rejected after visual inspection. We then performed Independent Component Analysis (ICA) and removed components related to eye blink activity and other artefacts (manually selected for each participant).

EEG: Univariate analyses

A regression-based approach allowed us to disentangle the influence of various variables on the univariate signal during the numerical comparison task. Before running the regression, we used Principal Component Analysis (PCA) on the pre-processed data of each participant to reduce noise in the signal. All epochs with primary targets were stacked over all electrodes and trials, creating a feature matrix of (trials*electrodes) by time points. The first 43 principal

components (PC) were retained, which on average explained 90% of the data. The EEG data was then reconstructed by multiplying the PC's with the estimated loadings and the reconstructed data was subsequently used as the dependent variable in our linear regression model.

The regression model contained 4 regressors of interest: (1) numerical magnitude of the current primary target, (2) the absolute difference of the current primary target to the previous target, (3) absolute difference between the *current* target and the mean of the current context, and (4) absolute difference between the *previous* target and the mean of the current context. We included two nuisance regressors to exclude alternative explanations of the univariate results: (5) visual similarity of the current primary target with the previous target and (6) reaction time (RT) on the current trial. Visual similarity was estimated as the correlation distance between black-and-white pixel images of two numbers as they were presented on screen. Trials with no response or RTs beyond 2.5 SD of the inverse RT were excluded from analysis. All regressors were z-scored before estimating the beta coefficients for each time point and electrode. For a control analysis, the same regression model was repeated replacing the regressors coding for distance to the context mean with the distance to the global mean for current and previous target.

EEG: Representational Similarity Analysis (RSA)

We constructed neural Representational Dissimilarity Matrices (RDMs) from the EEG data at each time point. First, data was z-scored over all trials per electrode and time point. Condition-specific neural signals were estimated at each time point and electrode using a regression model with dummy codes for every number in each context (in total 38 predictors: 25-35 for Low, 30-40 for High and 25-40 for Full), where the beta coefficients reflected the average deflection in EEG signal per condition at each electrode. The residuals of the regression were used to estimate the covariance matrix, which was subsequently used to increase the signal-to-noise ratio (SNR) by noise normalizing (whitening) the averaged EEG data through multiplication with the negative half inverse covariance matrix ($\Sigma^{-0.5}$) at each time point (Walther et al., 2016). Finally, we calculated the Pearson correlation distance between each condition, resulting in a 38 x 38 RDM at each time point. To exclude the possibility that more observations in certain contexts were biasing the dissimilarity measures, we constructed RDMs iteratively with subsampled data that equated the frequency of observations in each cell of the RDM. We repeated the RDM construction 100 times, randomly selecting N observations per cell at each iteration, with N determined by the minimum number of observations in any condition for a particular participant. The final neural RDM then consisted of the averaged RDM over all 100 iterations.

We constructed 3 candidate model RDMs that represented different potential structures in the neural signal: context, magnitude and visual similarity. The context model RDM assumed complete similarity (0) between numbers coming from the same range context and complete dissimilarity (1) for numbers coming from different ranges. The magnitude model RDM encoded the absolute difference between each number in each context. The visual similarity RDM was calculated using correlation distance between the vectorized black-and-white pixels of the numbers as they appeared on screen. All model RDMs were z-scored before entering in the regression to assess their relative contribution. Beta coefficients were estimated at each

time point and for each participant separately and the resulting beta series were smoothed over time for visualisation. Statistics are reported at the group level. Significant clusters were identified using cluster-corrected non-parametric permutation tests (Maris & Oostenveld, 2007).

For a supplementary control analysis, two additional subject-specific models were added to the regression: a colour model RDM and a reaction time (RT) model RDM. For the model RDM representing the colour space, RGB values were first transformed into CIE 1976 L*a*b* space to more accurately reflect human perception of colour differences. The dissimilarity between colours was then indexed through $\Delta E*$, a measure of the colour difference between two colours in CIELAB space. The RT RDM was constructed by calculating the average RT to a number independent of the preceding number. RTs were cut-off based on the 2.5 standard deviation from the inverse RT to control for outliers before averaging.

Multidimensional Scaling and Model Fitting

Neural state spaces were visualised by reducing the similarity data (RDMs) to three dimensions using multidimensional scaling with metric stress (equivalent to plotting the first three principal components of the data). Next, we used a model fitting exercise to characterise how the neural representation of number and context is organised in this low-dimensional space. We assumed that each context was characterised by a neural representation of numbers lying on a line, that the centroid of the [x,y,z] coordinates for the numbers in each context was the midpoint of this line, and that the three lines (one for each context) were parallel. We fit a model with 6 degrees of freedom: parameters 1-3 were the lengths of the each of the three lines, and parameters 4-6 were the angles of the (parallel) lines in dimensions [x,y,z]. The model was fit using gradient descent to minimise the discrepancy between simulated number positions and observed number positions in the low-dimensional neural state space. We used parameter recovery to verify that this model could recover ground truth line lengths and orientations in a simulated space. We used this approach both for human data (see **Fig. 4**) and RDMs constructed from RNN hidden unit activations (see below and **Fig. 3**). We used the same approach on the group mean similarity data (average RDM) and individual human/network RDMs. For each subject (n = 38 humans; n = 10 RNNs), we fit three variants of the model: one in which the line lengths were constrained to be the same (relative model); one where the line length in the full condition was constrained to be $16/11$ times longer for the full condition that low or high (reflecting the larger range; absolute model); and one where no constraints were imposed (free model). The residuals (loss) of the absolute and relative models were compared using frequentist statistics.

The line length parameters estimated from the free model were used to compute indices of divisive normalisation as follows:

$$dn = \frac{\lambda_{low} + \lambda_{high}}{2\lambda_{full}}$$

where λ_i is the estimated length parameter for condition i. The index dn will thus be 1 for perfect normalisation, i.e. when the line lengths in full and low/high conditions are equal and

will approach 11/16 if the line length is reflective of the absolute range of numbers in each context. The subtractive normalisation index was calculated as follows:

$$sn = 1 - \left(\frac{[\mu_{full} - \mu_{low}] - [\mu_{full} - \mu_{high}]}{2\lambda_{full}} \right)$$

where μ_i is the centre of dimension d for condition i. This index captures the relative offset of the centroid of the low/high conditions from the full condition, normalised by the line length in the full condition. This index will approach 1 if there is full normalisation (i.e. if there is no offset) and 0.5 for full offset. We ensured that any meaningful offset occurred in dimension $d = 3$ by rotating the lines (using the best fitting parameters 4-6) so that any elongation on a magnitude axis, if present, would occur in the dimension 3 (we also visualised neural state spaces after this rotation had been applied).

RNN architecture

A simple recurrent neural network model was built to perform the same primary task as human participants. The network was trained to transform symbolic inputs **x** on each trial t using a recurrent layer $\boldsymbol{h}^{(1)}$, followed by a fully connected feed-forward layer $\boldsymbol{h}^{(2)}$, which projected to a single output node y. Rectified linear (ReLU) activation functions *f(u)* were applied to each of the two hidden layers and a sigmoid activation function *σ(u)* was applied to the output layer to constrain the response 0≤y≤1. Thus the network took the form:

$$\boldsymbol{h}^{(1)}(t) = f(\boldsymbol{b}^{(1)} + \boldsymbol{W}^{(1)}(\boldsymbol{U}\boldsymbol{h}^{(1)}(t-1) + \boldsymbol{V}\boldsymbol{x}(t)))$$
$$\boldsymbol{h}^{(2)}(t) = f(\boldsymbol{b}^{(2)} + \boldsymbol{W}^{(2)}\boldsymbol{h}^{(1)}(t))$$
$$y = \sigma(\boldsymbol{b}^{(3)} + \boldsymbol{W}^{(3)}\boldsymbol{h}^{(2)}(t))$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are binary matrices used for concatenating the current input $\boldsymbol{x}(t)$ and the recurrent hidden activations from the previous trial $\boldsymbol{h}^{(1)}(t-1)$. Vectors $\mathbf{b}^{(i)}$ are the biases on layer I, and the two hidden layers $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$ were 220 and 200 units wide respectively. Weights and biases on all layers were initialised with random samples from a uniform distribution spanning ±1/sqrt(n), where n is the number of nodes in the upstream layer. The hidden state on the first trial in the first block of each training and test set was initialized to zero, and subsequent blocks were initialised with the final hidden state of the previous block. Neural networks were built and trained in PyTorch[35].

Training the network

Data was generated to simulate the trial sequence presented to the human participants, and so followed the same generation procedure. As in the human study, each block was 120 trials in length, containing 30 primary targets drawn from numbers corresponding to a single context, while filler trials always spanned the full range. The order of blocks was pseudorandom and the number of blocks in each context was balanced. However, unlike the human participants, the network had no a priori knowledge of numerical magnitude, and so many more blocks of trials (2880 blocks) were generated for training each network than the 24 blocks used per human participant. The RNN was trained only to perform the primary task, and responses on the filler trials were ignored.

On each trial, the input vector **x** consisted of a one-hot representation of the current number, as well as a node indicating whether the current trial was a primary target or a filler trial, and an (optional, see Context Manipulations) one-hot coded vector reflecting the current context. The context cue was included to simulate the colouring of the primary targets by context range in the human experiment. As filler trials were always white in the human experiment, the context cue inputs to the network were always zeroed on filler trials. The primary target numbers input to the RNN for each context were sampled from the same distributions as in the human experiment but were represented by one-hot codes spanning the ranges 1-16 (full range), 1-11 (low range), and 6-16 (high range) to save on input nodes.

The network was trained to perform the primary task (numerical comparison) and during training received feedback on primary target trials according to a binary cross-entropy cost function (but no feedback was given on the first primary target in each block). Errors were backpropagated through time at the end of each block of trials and the weights updated with stochastic gradient descent. Network outputs >0.5 were interpreted as analogous to pressing the 'up' arrow key in the human experiment, meaning that the current number was thought to be greater than the previous primary target while responses <=0.5 were interpreted as presses of the 'down' arrow key. Training hyperparameters were fixed for all networks at 10 epochs, a learning rate 0.0001, and momentum of 0.9. Network parameters were frozen at the end of training, prior to test.

Each network was trained 10 times, with different random initialisations and random datasets. Each dataset consisted of a training set (2880 blocks), and a test set for assessing the network activations (480 blocks).

Context manipulations

To isolate factors that could lead to context-separation and normalisation in both the human and RNN activations, we factorially controlled the provision of explicit context cues to the network (analogous to colour in the human experiment), and the blocking of trials in time by numerical context. We trained four groups of networks to fully cross these two factors. In each group of networks, context cue inputs on primary trials either reflected the numerical range of the block, or these inputs were kept constant across all blocks. Additionally, primary targets were either drawn from a single numerical range within a block, as in the human experiment, or primary targets within each block were drawn from the global distribution of primary target numbers, which spanned all three numerical ranges. Therefore, while the blocking of numerical range changed between groups of networks, the total number of times each number occurred as a primary target across the dataset remained the same for all groups. For each network, the same manipulations were made across training and test sets.

Virtual inactivation

Primary target trials were virtually inactivated (VI; zeroing the inputs that communicated number) randomly with probability ε during training, and errors backpropagated. Groups of

networks were trained with one of four different VI probabilities, $\varepsilon = \{0.0, 0.1, 0.2, 0.3, 0.4\}$. Context was explicitly cued in the input and trials were blocked by number range for all networks with VI during training. To evaluate whether the trained networks learned to use numerical context when solving the primary task, we used VI at presentation of a single primary target in the test set and assessed performance on the subsequent primary target. In effect, this forced the network to 'guess' whether the numerosity of the current input was likely to be greater or less than a 'forgotten' previous primary target. Network weights were frozen prior to any VI assessments on the test set. This process was repeated for all primary targets in the test set and post-VI performance was taken as the mean percentage of trials that the network answered correctly, evaluated on the primary target trial after the VI.

Calculations of normative performance

To evaluate the usage of local context information by the RNN, we compared the post-VI RNN performance to two benchmark levels. These benchmarks were calculated as the best overall performance achievable for an agent following each of two different policies when presented with a primary target (and not the previous primary target). These were π_{local}: a policy that used the local context information when responding. Under policy π_{local} the agent guesses the current primary target x_A to be greater than the previous (inactivated, forgotten) primary target x_B if the current target is greater than the median of the current range of primary targets, assuming the range of primary targets was learned during training. We also evaluated performance under π_{global}: a policy that used the global context when responding. Under policy π_{global} the agent guesses that the current primary target x_A was greater than (inactivated, forgotten) x_B if x_A is greater than the median of all primary targets across all contexts (which was similarly learned during training). That is

$$\pi_{local}: \quad x_A > x_B \;\; if \; x_A > \tilde{x}_{local}$$
$$\pi_{global}: \quad x_A > x_B \;\; if \; x_A > \tilde{x}_{global}$$

The probability that an agent following policy π responds correctly on a random trial is given by

$$P_\pi(c|\pi, \tilde{x}) = \sum_{\mathrm{a}} P_\pi(c|x_A = \mathrm{a}, \pi, \tilde{x})\, P(x_A = \mathrm{a})$$

For either the local or global case, applying one of the above policies gives

$$P_\pi(c|\pi, \tilde{x}) = \sum_{\mathrm{a}=x_{min}}^{\tilde{x}} P(x_A = \mathrm{a})\big(1 - P(x_A > x_B|x_A = \mathrm{a})\big) + \sum_{\mathrm{a}=\tilde{x}+\varepsilon}^{x_{max}} P(x_A = \mathrm{a}) P(x_A > x_B|x_A = \mathrm{a})$$

Under a context-blocked training and test regime, the distribution of x_A is uniform across the local context range of primary targets (full, low or high), and the distribution of x_B is uniform across the same range with the exception that $x_A \neq x_B$. If N_A is the number of possible outcomes of x_A, and N_B the number of possible outcomes of x_B, then

$$P(x_A = \mathrm{a}) = \frac{1}{N_A}$$

$$P(x_A > x_B | x_A = \mathrm{a}) = \frac{\mathrm{a} - x_{min}}{N_B}$$

The performance values presented in **Fig. 2 & S1** were then found by substituting values for x_A for each context and policy.

RNN activations

Network activations were evaluated on the test set. Inputs were passed to the network as in training, and the activation of units in the final hidden layer $\mathbf{h}^{(2)}$ was recorded and these activations averaged across all presentations of the same number, context and trial type (primary, or filler) in the test set.