The representational geometry of cognitive maps under dynamic cognitive control

Seongmin A. Park [1], Maryam Zolfaghar* [2,3], Jacob Russin* [3,4], Douglas S. Miller [1], Randall C O'Reilly [2,3,4], Erie D. Boorman [1,3,4]

Recent work has shown that abstract, non-spatial relationships between entities or task states are organized into representations called cognitive maps. Here we investigated how cognitive control enables flexible top-down selection of goal-relevant information from multidimensional cognitive maps retrieved from memory. We examined the relationship between cognitive control and representational geometry by conducting parallel analyses of fMRI data and recurrent neural network (RNN) models trained to perform the same task. We found both 2D map-like representations in a medial temporal lobe and orbitofrontal cortical network and simultaneous 1D orthogonal representations of relevant task dimensions in a frontoparietal network, supporting representational stability and flexibility, respectively. These representational motifs also emerged with distinct temporal profiles over the course of training in the RNN. We further show that increasing control demands due to incongruence (conflicting responses) between current task-relevant and irrelevant dimensions produce warping along the context-invariant axis in subjective representations, and the degree of warping further accounts for individual differences in cognitive control. Together, our findings show how complementary representational geometries balance generalization and behavioral flexibility, and reveal an intricate bidirectional relationship between cognitive control and cognitive map geometry.

[1] Center for Mind and Brain, UC Davis

[2] Department of Computer Science, UC Davis

[3] Center for Neuroscience, UC Davis

[4] Department of Psychology, UC Davis

* Contributed equally

Corresponding to S. A. Park (seongmin.a.park@gmail.com) and E. D. Boorman (edboorman@ucdavis.edu)

Generalizing previous experiences to make flexible decisions according to current goals is a hallmark of behavioral flexibility. This behavioral flexibility depends on the ability both to build representations shared across multiple tasks and also flexible, top-down selection of information according to changing task goals (Bernardi et al., 2020; Flesch et al., 2022; Freund et al., 2021; Higo et al., 2011; Koechlin, 2016; Miller & Cohen, 2001; Musslick & Cohen, 2021; Nee, 2021; Takagi et al., 2021; Vaidya & Badre, 2022). Recent findings suggest that the hippocampus (HC), entorhinal cortex (EC), and orbitofrontal cortex (OFC) organize relationships between items, individuals, events, or abstract task states, even when multiple task dimensions are sampled in different task contexts, into a unitary combined representation (Behrens et al., 2018; Bellmund et al., 2018; Eichenbaum & Cohen, 2014; Garvert et al., 2017; Knudsen & Wallis, 2021; Park et al., 2020, 2021a; Schuck et al., 2016). Recent computational models of cognitive control suggest that the brain trades off the benefits of low-dimensional representations for fast learning and generalization across multiple tasks against high-dimensional representations for flexible decision making (**Fig. 1a**) (Badre et al., 2020; Bernardi et al., 2020; Buschman, 2021; Ehrlich & Murray, 2022; Fusi et al., 2016; Garner & Dux, 2022; Musslick & Cohen, 2021; Rigotti et al., 2013). An intriguing but unexplored possibility is that an analogous tradeoff between generalized and flexible representations may determine the geometry of cognitive maps defined by multiple attributes.

Previous studies of cognitive control have revealed the mechanisms by which the brain biases explicitly presented sensory features, such as color, motion, shape, or orthographic features, while participants integrate sensory inputs providing evidence favoring different, sometimes conflicting, responses (Aoi et al., 2020; Bichot et al., 2019; Flesch et al., 2022; Mante et al., 2013; Takagi et al., 2021). While these studies have made dramatic strides towards understanding the top-down biasing of sensory information based on current task sets, relatively little is known about how cognitive control is implemented through the lens of selecting dimensions from mnemonic representations, particularly cognitive maps. In particular, the representational geometries defined by relative location between entities' or task states' representations on a neuronal population's manifold may play a central role in shaping control demands when they are not determined by incoming sensory feature values, but instead by learned relationships shaped by experience and retrieved from relational memory.

To examine how changing task goals interact with the representational geometry of cognitive maps to guide flexible decision making, we assessed how neural representations of two-dimensional (2D) social hierarchies are shaped by different cognitive control demands (**Fig.1a**). When making relational inferences between entities – e.g. who is more senior, or, on the other hand, more popular, amongst colleagues – control demands increase due to interference caused by incongruence between current task-relevant and irrelevant dimensions – e.g. José is more senior than Gabriela and so has more influence in the board room, but Gabriela is more popular and, hence, more influence at a social gathering. Recent findings suggest that the representational geometry of task representations might be sculpted by task demands to facilitate generalization to similar contexts that share a decision policy (i.e. choice rule or mapping from entities to decision values) (Barsalou, 1983; Bernardi et al., 2020; Ehrlich & Murray, 2022; Ho et al., 2022; Mack et al., 2016; Nieh et al., 2021; Tang et al., 2019). If control demands resulting from different levels of congruence play a role in shaping the resulting subjective representations, we would expect the brain to adopt a versatile representational strategy whereby congruence promotes sharing, and therefore generalization, across dimensions of the task, while incongruence promotes segregation, thereby allowing greater flexibility across contexts or tasks with divergent responses. Consequently, the effects of congruence might generate warping in the map-like representations such that distances along the congruent axis are elongated compared to distances along the incongruent axis (**Fig.1b**). In the current study, we test for such systematic warping in the neural representations of cognitive maps and its influences on individual differences in cognitive control during model-based inferences.

To facilitate effective value coding at the time of decision making, control- and decision-related brain regions may dynamically collapse or compress the multidimensional map-like representations into one-dimension (1D) representations according to the current task goals (**Fig.1c**). When features of sensory stimuli guiding responses are presented explicitly, previous studies have shown top-down control from the prefrontal cortex gating the inputs of task-relevant information on the early sensory cortex (Bichot et al., 2019; Higo et al., 2011; Rougier et al., 2005) or attenuating task-irrelevant information relative to task-relevant information, even of proficient participants who underwent intensive training (Flesch et al., 2022; Takagi et al., 2021). Other studies have instead proposed that currently task-irrelevant evidence is simultaneously and orthogonally encoded faithfully in prefrontal cortex without any top-down gating (Aoi et al., 2020; Mante et al., 2013; Panichello & Buschman, 2021). Whether or how these competing mechanisms apply to inferences over internal representations of cognitive maps, where the task-relevant dimensions are not determined by features of sensory inputs, but by the learned relationships between entities retrieved from memory, is unclear. Specifically, the representational geometries in the same frontal-parietal network might be dynamically changed across trials according to the current task goals by compressing the distances along the task-irrelevant dimension compared to those along the task-relevant dimension.

To address these questions, we combined recurrent neural network (RNN) models with human fMRI data on a task that required subjects to make context-dependent inferences about relationships between entities stored in cognitive maps. We found 2D shared representations across behaviorally-relevant contexts in the hippocampus (HC), entorhinal cortex (EC), and orbitofrontal cortex (OFC), even though only 1D relationships were behaviorally relevant for current decisions. In parallel, we found context-specific 1D representations with dynamic compression of task-irrelevant dimensions relative to task-relevant dimensions in the dorsomedial frontal cortex and lateral prefrontal cortex (dmFC and lPFC), and posterior cingulate cortex and inferior parietal lobule (PCC and IPL). Consistent with human fMRI, we found both context-invariant 2D shared representations and context-dependent 1D representations in the RNN trained on the same task. Moreover, we found evidence of warping in both HC representations and the RNN such that representational geometries were skewed along the congruent versus incongruent axis. Notably, our results show parallels between representations of cognitive maps in human brains and those that emerge in the RNN, and thus provide insight into the neural mechanisms that support the construction and geometry of cognitive maps, the flexible prioritization of goal-relevant information, and their interplay.

Results

To examine how representational geometry and cognitive control may interact to form subjective cognitive maps, we performed new analyses on previously collected fMRI data (Park et al., 2020) and compared neural representations in humans to those in hidden layers of RNN models trained on the same task (not published previously). Both human participants (n=27) and RNN models (classic long short-term memory [LSTM] networks) (Hochreiter & Schmidhuber, 1997) were trained to learn the relationships between 16 individual faces on two social hierarchy dimensions (**Fig.1a**), called competence and popularity, using highly similar previously detailed training protocols (Park et al., 2021b). During training participants and the RNN learned relationships between neighboring entities – i.e., individuals whose rank differed only by one level, in one dimension at a time, through feedback on their decisions. Human participants learned the relationships in each of two dimensions on different days while the model learned them in the same block with intermixed trials (See methods) to avoid catastrophic forgetting (McClelland et al., 1995; McCloskey & Cohen, 1989). After training, both human participants (**Fig.1d**) and the RNN (**Fig.1e** and **Eq.12-14**) were asked to infer relationships between two sequentially presented individuals who had not been learned or compared during training in the given social hierarchy dimension. The social hierarchy dimension relevant to the current goal was signaled by a colored context cue at the beginning of every trial and contexts were randomly interleaved, thus requiring subjects to flexibly switch between contextually-defined task dimensions. It is important to note that neither participants nor the model were provided any information about the true social hierarchy structure, nor were they asked to construct it spatially.

Using both human fMRI data and the RNN, first, we examined simultaneous unitary map-like representations of the social hierarchies in both dimensions (context-invariant representations) and tested for the effects of cognitive control modulating representational geometries to facilitate execution of the current task while resolving interference from task-irrelevant information (context-dependent representations). Second, we tested for hypothesized warping in the subjective 2D map-like representations. Third, we explored the relationship between cognitive control and different levels of warping in the representations across individual participants.

Simultaneous shared context-invariant and orthogonal context-dependent representations

In our previous work (Park et al., 2020), we provided evidence suggesting that the brain builds a shared representation across different social hierarchy dimensions and uses it to make inferences, even when only one dimension is relevant for current decisions. Specifically, we found that the Euclidean distances between individuals in the true 2D social hierarchy structure explained the decision-related BOLD activity in the EC and ventromedial prefrontal cortex (vmPFC), as well as reaction times (RTs), better than 1D task-relevant rank differences alone. Moreover, we found 2D map-like representations of the social hierarchy in the HC, EC, and OFC, such that closer individuals in the true hierarchy (measured by pairwise Euclidian distance) were represented progressively and linearly more similarly (measured by pairwise pattern similarity).

While benefiting from shared representations across multiple task contexts for generalization and fast learning, the brain may also benefit from flexibly transforming this higher-dimensional representation into 1D representations according to the current task goals to effectively resolve interference between information in the task-relevant and task-irrelevant dimensions. Motivated by recent evidence in the perceptual selective attention literature (Aoi et al., 2020; Flesch et al., 2022; Mante et al., 2013; Panichello & Buschman, 2021; Takagi et al., 2021), we tested for the orthogonal representation of the axes of the two different relevant

dimensions and those of the irrelevant dimensions, with compression of evidence along the irrelevant axis. To test for simultaneous representations, and control for alternative representational geometries in the brain, we examine to what extent each of several candidate model representational dissimilarity matrices (RDMs) explains the pattern dissimilarity associated with individuals in the social hierarchy using a general linear model (GLM). Specifically, in the same GLM, we inputted the pairwise Euclidean distances in the 2D social hierarchy and those computed from orthogonal 1D representations with dynamic compression according to the task context, as well as two other control RDMs (task context and learning group; see Methods), to pit them against one another to explain variance in pattern dissimilarity (**Supplementary Fig.1**; **Eq.1**).

Consistent with our previous study (Park et al., 2020), this new analysis found the 2D map-like representations in the HC (peak=[28,-2,-18], t=4.67), EC ([20,0,-36], t=4.02) and vmPFC encompassing the medial OFC ([14,42,-22], t=5.83), and lateral OFC ([28, 22, -22], t= 3.96 in right and [-24, 26, -16], t= 4.55 in left) (pTFCE<0.05; **Fig.2a**). In parallel, we found the 1D orthogonal representation of the relevant dimensions with the irrelevant dimension dynamically compressed relative to the current task-relevant dimension in the dorsomedial PFC (dmPFC, [-4,54,24], t=5.99), the dorsomedial frontal cortex (dmFC) encompassing the pre-supplementary motor area (pre-SMA) ([4,14,48], t=5.59), lateral PFC (lPFC, [42,26,18], t=9.03 in right; [-42,24,20], t=13.48 in left) extending to lateral frontal lobe, precuneus/posterior cingulate corex (PCC, [-4,-64,28], t=13.60), right inferior parietal lobule (IPL, [46,-64,22], t=5.20) extending to temporoparietal junction (TPJ) , and left IPL extending to TPJ and posterior superior temporal sulcus (STSp) ([-48,-72,18], t=6.72) and fusiform gyrus ([30,-84,-10], t=5.54 in right; [-40,-60,-20], t=6.44 in left) (pTFCE<0.05; **Fig.2b**). See **Table S1** for a full list of brain areas surviving correction. Taken together, these representational similarity analyses (RSA) show that the HC, EC and OFC exhibit task-invariant stable 2D map-like representations that are shared across contexts in parallel to task-dependent 1D orthogonal representations with dynamical compression in a frontoparietal network ideal for computing the goal and context-dependent decision values.

Shared 2D and orthogonal 1D representations emerge in the RNN

To test if, under what conditions, and when during training this representational strategy might emerge as an efficient solution to our task, we performed the same GLM-based RSA analyses of our RNN model (**Supplementary Fig.1**). We also found the emergence of both representations in the neural network model, even though we did not design the models with any specialized components to produce these representations. The pairwise distances between representations in the neural network models were explained by the 2D Euclidean distances between individuals in the task-invariant 2D social hierarchy structure (**Fig.2a**), as well as those in the task-specific 1D orthogonal representations with dynamic compressions (**Fig.2b**). The 2D map-like representations emerged at a relatively early phase of training (the representation was significant [p<0.05] from the training step, 150 [the median run: t=4.35, p=1.65e-5]; the peak training step was 300 [the median run: t=8.31, p<1e-4]) and were maintained throughout training (t=5.63, p<1e-4 at the last training step). Simultaneously, the models also utilized task-dependent orthogonal representations, and this tendency for dynamic compression increased continuously over training (significant from the training step, 250 [the median run: t=4.53, p=1.86e-5]; the peak training step was 1000 [the median run: t =11.75, p<1e-4]). Thus, both shared 2D and orthogonal 1D representations emerge as complementary, efficient representations to solve our relational inference task, but with different time courses.

Effects of cognitive control on the representational geometries

Based on empirical and theoretical evidence that task demands during learning might shape a task space's subjective representation (Barsalou, 1983; Bernardi et al., 2020; Ehrlich & Murray, 2022; Ho et al., 2022; Mack et al., 2016; Nieh et al., 2021; Tang et al., 2019), we considered to what extent cognitive control demands might influence the relational configuration between entities and hence the representational geometry of the 2D cognitive map. Specifically, we test the hypothesis that increasing control demands for incongruent compared to congruent pairs produces warping along the congruent axis in the subjective representation of the 2D social hierarchy. We tested this hypothesis in four independent ways.

First, using multiple regression analysis (**Eq.2**), we found that RTs for inferences depended on the congruence of the relationships between individuals in the two task dimensions, even after controlling for the effects of other distance measures of inferred trajectories, including 2D Euclidean distances and 1D rank differences in the task-relevant dimension (**Fig.3a**; -0.016±0.004 (SE), t=-4.09, p=3.67e-04). This analysis indicates that the degree to which Euclidean distances accounted for RTs (ε) depended on whether pairs were congruent (faster RTs) or incongruent (slower RTs) — providing an individual signature of cognitive control. By computing effects of the degree of congruence (defined as the vector angles of inferred trajectories, θ) on RTs, while controlling for the effects of the distances of inferred trajectories across trials, we approximated the hypothesized subjective representational geometry and the implied level of warping across participants (**Supplementary Fig.2; Eq.3-9**) for the procedure and confirmatory simulations using synthetic data). The levels of warping were estimated as the ratio of the incongruent diagonal distances to the congruent diagonal distances. The mean ratio between diagonals were 0.961±0.008 (SE). We found that the hypothetical subjective representations estimated from the RTs were significantly warped along the context invariant axis compared to the shuffled baseline (t=2.93, p=0.007, 1000 permutations; **Supplementary Fig.2**). The distribution of the estimated level of warping across participants is shown in **Fig.3b**. Importantly, we used relative values of the diagonal ratio to index individual differences in cognitive control in further analyses presented below.

Second, we tested if the inferred trajectories over the hypothesized warped representations better explain the neural encoding of decision values for inferences in the EC and vmPFC. We defined *a priori* region of interests (ROIs) anatomically based on our previous findings (Park et al., 2020) and found that the Euclidean distances in the *warped* representations could explain additional activity in the EC (peak [x,y,z]=[18,-10,-26], t=2.91, pTFCE=0.04 in the EC ROI) and the vmPFC ([-2,28,-12], t=3.75, pTFCE=0.04 in the vmPFC ROI), after controlling for the variance accounted for by Euclidean distances between individuals in the 2D social hierarchy (**Fig.3c**). This finding indicates that the hypothesized warped representations explain additional variance in decision value-related activity in these brain regions. Additionally, we tested an alternative, but non-mutually exclusive hypothesis that changes in cognitive control might stem from the costs of task switching, rather than warped neural representations. Notably, we did not find significant effects of switching task contexts on RTs (0.002±0.007 (SE), t=0.22, p=0.83), nor neural activity across the whole brain even at a lenient threshold (p<0.005, uncorrected) in our task.

Third, for the critical test of warped map-like representations in human participants, we directly compare distances between representations of congruent and incongruent pairs sampled from different trials across blocks. The results of our RSA of fMRI data consistently showed warping along the congruent compared to incongruent axis — lower pattern similarity between representations of congruent pairs of faces than equidistant incongruent pairs of faces. Specifically, whole-brain searchlight-based RSA (**Supplementary Fig.3** and **Eq.10**) of fMRI data revealed the warped geometry along the congruent over incongruent axes at the group level representation in the right HC ([28,-12,-22], t=4.50), dorsolateral prefrontal cortex ([42, 34, 24], t=4.56), fusiform gyrus ([44,-38,-14], t=4.60), inferior temporal gyrus ([64,-12,-30], t=4.79), and

supramarginal gyrus ([52,-30,42] t=5.51) (pTFCE<0.05 whole-brain corrected, **Fig.3d**; See **Table S2** for a full list of brain areas). The HC effect was also significant after excluding individuals who had the highest or lowest rank in both social hierarchy dimensions (t=4.06, pTFCE=0.05 corrected in the anatomical HC ROI), indicating that the observed warping was not only caused by specific faces, such as those at the boundaries (highest or lowest ranks in both dimensions).

Fourth, we harnessed our inter-individual behavioral measure of warping to test for brain regions showing greater warping in those participants who exhibited greater behavioral evidence for RT costs on incongruent compared to congruent pairs. This analysis revealed that individual differences in our behavioral measure of warping correlate with the levels of warping observed in the amygdala (right [28,0,-12], t=6.67; left [-20,-8,-16], t=3.30), ventromedial/medial prefrontal cortex (vmPFC/mPFC) ([-5,34,2], t=4.52), dorsomedial PFC (dmPFC, [22, 48, 16], t=9.90), lateral OFC (lOFC, [34, 32, -18], t=4.68), and fusiform gyrus (right [28, -54, -6] t=5.88; left [-36,-32,-18], t=4.24) (pTFCE<0.05 whole-brain corrected; **Fig.3e**; See **Table S2** for a full list of brain areas). The correlation between the activity estimated from anatomically defined ROIs were r=0.60 (p=0.001) and r=0.49 (p=0.009) in the right and left amygdala, respectively, and r=0.61 (p=0.001) in the vmPFC, suggesting those individuals with greater representational warping in these regions experienced greater demands on cognitive control during decisions. The relationship between the representational geometry and individual differences in cognitive control during inferences was also supported by the results of another independent analysis, termed intersubject representational similarity analysis (IS-RSA) (Chen et al., 2020; Cohen et al., 2017; Hasson et al., 2004). The IS-RSA allowed identification of brain areas sharing similar representational geometries of the social hierarchy for participants who showed similar behavioral patterns in cognitive control during inferences, without assuming any specific underlying structure of the representations (**Supplementary Fig.4**). Consistent with the previous analysis, we found that individual differences in cognitive control were captured by differences between pattern similarity in the left amygdala ([-16,-2,-12], t=5.73), mPFC/dmPFC ([-8,56,38], t=6.28), fusiform gyrus (right [24,-50,-6], t=7.05; left [-20.-54.-10] t=4.30) and posterior cingulate cortex (PCC) ([-8,-48,28], t=7.28) across participants (**Fig.3f**; See **Table S3** for a full list of brain areas surviving at permutation-based pFDR<0.05 based on 1000 permutations). Control leave-one-out (loo) analyses indicated that the results were not driven by the level of warping of any single subject (**Supplementary Fig.4**). Collectively, the above lines of evidence converge to suggest people formed warped representations of the 2D social hierarchy, reflecting the asymmetric demands of cognitive control likely experienced during learning, and the resulting geometry of subjective representations accounts for individual differences in cognitive control.

Consistently warped representations emerge in the RNN

To explore under what conditions such representations might emerge during learning, we then perform identical RSA analyses to the hidden layers of the RNN (h_2 and h_3 in **Fig1e**) along the training trajectory to interrogate the embedding of the social hierarchy formed at different stages of training. Motivated by the human fMRI data, we specifically test for warping in the 2D map-like representation in the hidden layers by comparing distances between representations of congruent and incongruent pairs. Remarkably, consistent with the fMRI results, we found evidence for warping in the hidden layers (ω_{RNN}; **Eq.15**) along the congruent compared to incongruent axis that emerged during early to intermediate training timesteps (the peak: training step=300 [the median run: t=6.86, p<1-e-4]; the last training step=1000 [the median run: t=5.29, p<1e-4]) (**Fig.3g**), simply as a result of training on the same task and without any constraints hard-coded into the RNN. Notably, warping in the representations persisted even after prediction performance had reached an asymptote and following intensive training. Control analyses indicated that warping in the model's embedding was still present when excluding boundary faces who had the highest or

lowest rank in both dimensions (**Supplementary Fig.5**). This analysis specifically allowed us to test for warping by comparing the same representations of entities between congruent and incongruent pairs. Thus, the result indicates that the warping effect could not be due to the representations of any specific face, nor these outlier face positions. Together, these findings suggest that such warped 2D map-like representations are a sufficiently efficient representation for solving the context-dependent inferences in the current task that emerges during learning through back-propagation in a standard LSTM network.

Inefficient use of context information for learning incongruent relationships induces warping

Having found evidence that the brain's 2D map-like representation is warped according to subjective control demands, we next examined the hypothesis that cognitive control may determine the level of warping in the subjective representations during learning. Specifically, we addressed how incongruence between dimensions may cause the observed warping during formation of cognitive maps and result in different levels of warping in the representations across individuals. We focused on the effects of (in)congruence for the following reasons: On congruent trials, context information that indicates the task-relevant dimension does not strictly need to be maintained for decision making because the correct answer does not depend on the current context; On incongruent trials, on the other hand, context information is required at the time of decisions in order to overcome interference between conflicting responses and select the appropriate action (**Fig.4a**). To investigate why warping emerges, we interrogated the trajectory of the warping in emergent representations in the RNN during training and performed causal experiments to assess the impact on these representations.

Examining performance on congruent and incongruent trials separately revealed that the RNN started performing relatively better on inferences of congruent compared to incongruent trials at the early training steps (**Fig.4b**; See **Table S4** for the results of logistic regression). Moreover, this bias in performance (black line in **Fig.4b**) correlated with changes in the levels of warping (ω_{RNN}) (**Eq.18**; β=0.78±0.10 (SE), t=7.85, p<1e-4). While these effects were relatively large during early training, the representations remained warped even after the RNN's performance equated for congruent and incongruent trials (**Fig.4c**). This observation suggests that the warping in the RNN might be linked to an initial tendency to ineffectively utilize context information that would mitigate the interference on incongruent trials. To test the relationship between the use of context information and warping in the RNN, we performed an ablation experiment – we degraded the embedding vector of the context cue (by setting α to 1 to 10; e_a in **Fig.1e** and **Eq.16**) and examined changes in the geometries in the hidden layer's representations of the faces (h_2 and h_3 in **Fig1e** and **Eq.12**). The results of the ablation experiment showed increases in warping in the representations as a function of the degree of degraded context cue information (β=-0.97± 0.20 (SE), t=-4.84, p<1e-4) (**Fig.4d** and **Eq.17**).

Inspired by this observation in the RNN model, we performed additional behavioral and fMRI analyses. First, in a larger cohort of participants trained behaviorally, we found that participants whose accuracy was more asymmetric (greater in congruent compared to incongruent pair trials) likewise showed RTs that were more strongly explained by the effects of congruence, potentially consistent with more warped representations (**Fig.4e** and **Supplementary Fig.6**; r=0.34, p<0.001, n=218). This result and our RNN analyses point to a key role for context information in guiding decisions on incongruent trials. Second, we therefore tested for greater reinstatement of the current context information, specifically at the time of decisions, on incongruent than congruent trials. We first defined ROIs that showed significant context representations at the time of the context cue, which included the bilateral HC, mPFC, anterior insula, dorsal anterior cingulate cortex (dACC)/pre-SMA, retrosplenial cortex (RSC)/PCC and premotor cortex (**Supplementary**

Fig.7, pTFCE<0.05 whole-brain corrected). Next, we found stronger reinstatement of these context-specific activity patterns relevant to the current goal at the time of *decision making* (F2 presentation) on incongruent trials than congruent trials in the left HC ([-26,-8,-22], t=3.25), mPFC([-12,48,-4], t=3.19), dACC/pre-SMA([-6,12,44], t=6.59), anterior insula ([-30, 18, -8], t=5.48 in right and [36, 22, -2], t=6.44 in left), and RSC/PCC ([-4,-28,28], t=6.35), among other areas (pTFCE<0.05 corrected in the combined ROI), after controlling for the potential effects of the context, incongruence, and motor responses (**Fig.4f**; See **Table S5** for additional regions outside of these ROIs). These findings support the hypothesis of greater reinstatement of context-specific codes during inferences in the presence of incongruence.

This finding suggested that decision-related areas communicate with context-coding areas in the brain particularly on incongruent trials. We thus tested for increasing functional connectivity between regions showing greater activity on incongruent than congruent decisions and regions encoding context information at the time of inferences on incongruent compared to congruent trials. Specifically, a univariate analysis comparing incongruent and congruent inferences revealed a significant effect in the mid-anterior cingulate cortex (ACC), after controlling for subjective decision difficulty (Euclidian distance of inferred trajectories over the warped space) ([2,12,36], t=3.57). Next, we performed a psychophysiological interaction (PPI) analysis by inputting the activation in ACC (including voxels within a 10mm radius sphere around the peak) as a seed and incongruent versus congruent decisions as the psychological variable. This analysis revealed increased functional connectivity of the ACC to the bilateral HC ([30, -28, -14], t=2.93 in right; [-28, -24, 12], t=2.78 in left, pTFCE<0.05 in *a priori* anatomically-defined HC ROIs) and the PCC ([14,-32,28], t=4.47, pTFCE=0.04 corrected in the whole brain) on incongruent inferences (**Fig.4g** and **Table S6**).

Taken together, our behavioral and modeling analyses suggested that participants who use context information ineffectively to resolve incongruency should show greater warping in their representations. To capture the fidelity of the context reinstatement, we first calculated the similarity between the context cue activity pattern and the activity pattern on incongruent decisions in the corresponding context, with respect to the opposing cue's activity pattern. Next, we leveraged our behavioral RT marker of warping and found that the participants who showed less behavioral warping showed greater fidelity of context reinstatement during inferences on incongruent trials in the bilateral HC (**Fig.4h**, r=0.429, p=0.026 in right HC; r=0.369, p=0.059 in left HC). This finding indicates that participants whose inference processes were less influenced by incongruence had HC activity patterns whose two vector angles associated with retrieval of different contexts (competence and popularity) during inferences of incongruent trials (**Fig.4a**) were more similar to the respective vector angles associated with presentations of different contexts (**Supplementary Fig.7** and **Eq.11**). This effect was specific to incongruent inferences, since the degree of behaviorally estimated warping was not significantly explained by the degree of context reinstatement for congruent inferences (**Supplementary Fig.7**). Together with the RNN and human fMRI analyses, these analyses show a critical role for utilizing contextual information for effective cognitive control on representational warping.

Discussion

Our findings show how latent non-spatial, abstract cognitive maps are represented when learning the relationships piecemeal from two different task contexts, and how they are used to make flexible decisions in only one task dimension at a time. To address these questions, we used parallel analyses of an RNN model and human fMRI data and have shown three key phenomena about the relationships between the representational geometry of cognitive maps interacting with changing task goals and ensuing cognitive control demands resulting from interference between task relevant and irrelevant responses. First, both a hippocampal-entorhinal-vmPFC network and RNN learn representations of the latent task structure in multidimensional (2D) manifolds such that the relational information learned from different task contexts were represented as different axes of the same 2D representation that was shared across contexts. Second, to facilitate effectively selecting between the task-relevant and task-irrelevant dimensions, task-relevance modified the multidimensional representational geometry into orthogonal 1D number lines of the two task-relevant dimensions depending on the current task goal, and a suppression of the task-irrelevant dimension in a frontoparietal cortical network and the RNN. Finally, increasing control demands specifically for inferences on incongruent relationships produced individual differences in behavioral indices of control demands on incongruent compared to congruent relationships and 2D representations to be warped along the context-invariant (congruent) axis. This tendency created differences in subjective representations such that different levels of warping in neural representations account for individual differences in behavioral indices of cognitive control. The current findings cast cognitive control in a new light, whereby an individual's representational geometry is both sculpted by and shapes cognitive control when retrieving specific dimensions of endogenous representations from memory.

Although neither the model nor human participants were explicitly instructed about the underlying latent structure, representations in HC, EC, and OFC and in hidden layers of the RNN model captured this structure in that the similarity between representations of each entity (individual face) was correlated with their true pairwise Euclidean distances in the 2D social hierarchy (Park et al., 2020, 2021a). In addition to these stable representations that facilitate efficient generalization across task contexts, we also found simultaneous 1D representations in which the axes of the two different relevant dimensions and irrelevant axes were represented orthogonally in both separate brain areas including the dmFC, lPFC, PCC, and IPL and the RNN model. Importantly, these orthogonal 1D representations afford behavioral flexibility by dynamically compressing task-irrelevant dimensions relative to task-relevant dimensions as a function of which dimension is currently relevant. This finding suggests that the human brain may use similar principles for top-down biasing or selection of task-relevant from irrelevant dimensions as previously proposed for explicitly presented perceptually signaled features (Flesch et al. 2022; Takagi et al. 2021) to perform inferences over particular task-relevant dimensions of cognitive maps retrieved from memory. Notably, we also found both 2D and orthogonal 1D representations emerge in the representations of faces in the hidden layers of our RNN, albeit with distinct time courses. Consistent with the representations found in the human brain, task-general 2D representations formed in the RNN at earlier phases of training and remained after extensive training even after reaching asymptotic high-level performance. The RNN more gradually developed the 1D orthogonal representations as training progressed, and this tendency to utilize task-specific representations increased throughout training.

While other studies have observed multidimensional task structure representations (Bernardi et al., 2020; Nieh et al., 2021; Park et al., 2020, 2021a; Theves et al., 2020), little is known about the factors determining the representational geometry of such cognitive maps. If neural systems construct separate representations for separately learned task dimensions, these

could afford behavioral flexibility but in theory should incur costs for task switching between different cognitive maps. On the other hand, having a task-independent representation shared across multiple dimensions allows for fast learning via generalization and a single computation across dimensions, thereby mitigating costs of task/context switching (Badre et al., 2020; Garner & Dux, 2022; Musslick & Cohen, 2021; Vaidya & Badre, 2022). Our results indeed showed that both medial temporal and interconnected orbital and ventromedial prefrontal brain regions (Barbas & Blatt, 1995) and RNNs can construct a shared representation that can serve generalization across dimensions. In the context of our task, computing shared vectors across dimensions is beneficial for inferences on congruent relationships but not for inferences on incongruent relationships. Convergent pieces of evidence suggest that an initial shared representation whereby the relationships in different dimensions were collapsed into a 1D number line started to transition to 2D representations whereby the relationships in the two dimensions became represented on separate axes of a single representation (**Fig.1a**). Specifically, we found that those participants who showed greater accuracy asymmetries on incongruent compared to congruent trials also exhibited RTs better explained by the context-independent rank differences (or "model-free" values based only on win/loss ratio independently of the task-relevant dimension), whereas RTs of those who showed lower asymmetries were better explained only by the context-dependent task-relevant values (**Fig.4e** and **Supplementary Fig.6**). Complementing these behavioral data, the simulation results from the RNN also support this interpretation by showing that the shared representations that initially enabled accurate inferences on the congruent relationships better than incongruent relationships evolved into a 2D structure enabling accurate inferences on both with increasing in training steps (**Fig.4c**). This finding suggests that learning how many dimensions are relevant to potential behavior and how they are related to each other gradually transforms the representational geometry.

We also tested for the effects of congruence on the representational geometry of cognitive maps in the human brain by comparing distances between representations of congruent and incongruent pairs sampled from different trials across blocks. We found warping in the HC representations along the congruent compared to incongruent axis – greater differences in pattern similarity between representations of equidistant congruent pairs of faces than incongruent pairs of faces. Additionally, the levels of warping estimated in the amygdala, mPFC, and PCC were correlated with individual differences in our behavioral measure of cognitive control, suggesting those individuals with greater representational warping in these regions experienced greater demands on cognitive control during decisions. These results were also supported by an independent IS-RSA analysis. These findings jive with previous studies which have reported sensitivity to transitively inferred social hierarchy rank in activity in the amygdala, HC, and mPFC, and with evidence amygdala gray matter volume correlated with individual differences in transitive inferences in the social domain (Kumaran et al., 2012, 2016) and social group size (Bickart et al., 2010; Lewis et al., 2011; Noonan et al., 2014; Sallet et al., 2011; Testard et al., 2022).

Our findings further suggest that the warped representations along the congruent versus incongruent axis further influence the control demands which arise from interference between conflicting responses. To resolve this conflict, the brain needs to deploy control resources to the single task dimension relevant to the current goal signaled by one of two discrete color cues. The results of the ablation simulation support the importance of contextual representations by showing that increasingly degrading the contextual input impaired learning of incongruent relationships in particular and subsequently caused greater warping in 2D social hierarchy representations. We found consistent findings in the human fMRI data: the HC, mPFC, anterior insula, and RSC/PCC representations specific for a single contextual cue were reactivated to a greater extent during inferences on incongruent compared to congruent pairs. Importantly, among these ROIs, the level of reactivation in the HC specifically correlated with the level of warping estimated from behavioral RTs across participants. Taken together, our data suggest that efficient use of contextual

information resolve conflict by orthogonalizing dimensions in the shared representations, and the subjective representational geometries likely affect the control demands required for flexible decision making. Consistent with this interpretation, we additionally found increasing functional connectivity between the ACC, which showed heightened activity on incongruent compared to congruent inferences (controlling for Euclidian distances of subjective inferred trajectories), and bilateral HC, which showed both the 2D context-invariant representation and context reinstatement during incongruent inferences. An intriguing open question is how the context-specific HC representations and 2D map-like representations are related during goal-directed selection.

We found evidence of complementary representational geometries for efficiently encoding abstract relational information and flexibly selecting behaviorally relevant dimensions from those representations in the human brain as well as the neural networks model. Our findings suggest that cognitive control shapes the representational geometry of cognitive maps which can create warping in the subjective representations. We further showed that biases in model-based inferences could arise from the subjective representational geometry of cognitive maps. This insight opens a new research direction to better understand individual differences in learning and decision making, which could provide a framework to understand and potentially overcome irrational biases in decision making, social prejudices, and psychiatric conditions (Gillan et al., 2016; Schafer & Schiller, 2018; Shields & Gremel, 2020; Whittington et al., 2022). Moreover, the current study demonstrates the value of integrating a neural network modeling approach with neuroimaging through the interrogation of representations that emerge through learning, and may help to address current limitations of modern neural networks used for artificial intelligence (Hassabis et al., 2017; Saxe et al., 2020). Taken together, our results reveal an intricate relationship between cognitive control during cognitive map formation, the resulting representational geometry, and its role on subsequent control during decisions.

References

Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, *23*(11), 1410–1420. https://doi.org/10.1038/s41593-020-0696-5

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2020). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.31234/osf.io/asdq6

Barbas, H., & Blatt, G. J. (1995). Topographically specific hippocampal projections target functionally distinct prefrontal areas in the rhesus monkey. *Hippocampus*, *5*(6), 511–533. https://doi.org/10.1002/HIPO.450050604

Barsalou, L. W. (1983). Ad hoe categories. *Memory & Cognition*, *11*(3), 211–227. https://doi.org/10.3758/BF03196968

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Bichot, N. P., Xu, R., Ghadooshahy, A., Williams, M. L., & Desimone, R. (2019). The role of prefrontal cortex in the control of feature attention in area V4. *Nature Communications 2019 10:1*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-13761-7

Bickart, K. C., Wright, C. I., Dautoff, R. J., Dickerson, B. C., & Barrett, L. F. (2010). Amygdala volume and social network size in humans. *Nature Neuroscience 2010 14:2*, *14*(2), 163–164. https://doi.org/10.1038/nn.2724

Buschman, T. J. (2021). Balancing Flexibility and Interference in Working Memory. *Annual Review of Vision Science*, *7*(1). https://doi.org/10.1146/annurev-vision-100419-104831

Chen, P. H. A., Jolly, E., Cheong, J. H., & Chang, L. J. (2020). Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, *216*, 116851. https://doi.org/10.1016/j.neuroimage.2020.116851

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience 2017 20:3*, *20*(3), 304–313. https://doi.org/10.1038/nn.4499

Ehrlich, D. B., & Murray, J. D. (2022). Geometry of neural computation unifies working memory and planning. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(37), e2115610119. https://doi.org/10.1073/PNAS.2115610119/SUPPL_FILE/PNAS.2115610119.SAPP.PDF

Eichenbaum, H., & Cohen, N. J. (2014). Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? *Neuron*, *83*(4), 764–770. https://doi.org/10.1016/j.neuron.2014.07.032

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

Freund, M. C., Etzel, J. A., & Braver, T. S. (2021). Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach. *Trends in Cognitive Sciences*, *25*(7), 622–638. https://doi.org/10.1016/J.TICS.2021.03.011

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Garner, K. G., & Dux, P. E. (2022). Knowledge generalization and the costs of multitasking. *Nature Reviews Neuroscience 2022*, 1–15. https://doi.org/10.1038/s41583-022-00653-x

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

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. *ELife*, *5*(MARCH2016). https://doi.org/10.7554/ELIFE.11305

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*, 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity during Natural Vision. In *Science* (Vol. 303, Issue 5664). https://doi.org/10.1126/science.1089506

Higo, T., Mars, R. B., Boorman, E. D., Buch, E. R., & Rushworth, M. F. S. (2011). Distributed and causal influence of frontal operculum in task control. *Proceedings of the National Academy of Sciences*, *108*(10), 4230–4235. https://doi.org/10.1073/pnas.1013361108

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136. https://doi.org/10.1038/s41586-022-04743-9

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Knudsen, E. B., & Wallis, J. D. (2021). Hippocampal neurons construct a map of an abstract value space. *Cell*, *184*(18), 4640-4650.e10. https://doi.org/10.1016/J.CELL.2021.07.010

Koechlin, E. (2016). Prefrontal executive function and adaptive behavior in complex environments. In *Current Opinion in Neurobiology* (Vol. 37, pp. 1–6). https://doi.org/10.1016/j.conb.2015.11.004

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations Underlying Social Hierarchy Learning: Distinct Neural Mechanisms for Updating and Representing Self-Relevant Information. *Neuron*, *92*(5), 1135–1147. https://doi.org/10.1016/j.neuron.2016.10.052

Kumaran, D., Melo, H. L., & Duzel, E. (2012). The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. *Neuron*, *76*(3), 653–666. https://doi.org/10.1016/j.neuron.2012.09.035

Lewis, P. A., Rezaie, R., Brown, R., Roberts, N., & Dunbar, R. I. M. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. *NeuroImage*, *57*(4), 1624–1629. https://doi.org/10.1016/J.NEUROIMAGE.2011.05.030

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

Mante, V., Sussillo, D., Shenoy, K. v., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

McClelland, J. L., McNaughton, B. L., & 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. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, *24*(C), 109–165. https://doi.org/10.1016/S0079-7421(08)60536-8

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

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/J.TICS.2021.06.001

Nee, D. E. (2021). Integrative frontal-parietal dynamics supporting cognitive control. *ELife*, *10*. https://doi.org/10.7554/ELIFE.57244

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Ann Koay, S., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *80 | Nature |*, *595*. https://doi.org/10.1038/s41586-021-03652-7

Noonan, M. A. P., Sallet, J., Mars, R. B., Neubert, F. X., O'Reilly, J. X., Andersson, J. L., Mitchell, A. S., Bell, A. H., Miller, K. L., & Rushworth, M. F. S. (2014). A Neural Circuit Covarying with Social Hierarchy in Macaques. *PLOS Biology*, *12*(9), e1001940. https://doi.org/10.1371/JOURNAL.PBIO.1001940

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

Park, S. A., Miller, D. S., & Boorman, E. D. (2021a). Inferences on a Multidimensional Social Hierarchy Use a Grid-like Code. *Nature Neuroscience*, *24*(9).

Park, S. A., Miller, D. S., & Boorman, E. D. (2021b). Protocol for building a cognitive map of structural knowledge in humans by integrating abstract relationships from separate experiences. *STAR Protocols*, *2*(2), 100423. https://doi.org/10.1016/j.xpro.2021.100423

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. *Neuron*, *107*(6), 1226-1238.e8. https://doi.org/10.1016/j.neuron.2020.06.030

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7338–7343. https://doi.org/10.1073/PNAS.0502455102/SUPPL_FILE/02455SUPPTEXT.PDF

Sallet, J., Mars, R. B., Noonan, M. P., Andersson, J. L., O'Reilly, J. X., Jbabdi, S., Croxson, P. L., Jenkinson, M., Miller, K. L., & Rushworth, M. F. S. (2011). Social network size affects neural circuits in macaques. *Science*, *334*(6056), 697–700. https://doi.org/10.1126/science.1210027

Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience 2020 22:1*, *22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

Schafer, M., & Schiller, D. (2018). The Hippocampus and Social Impairment in Psychiatric Disorders. *Cold Spring Harbor Symposia on Quantitative Biology*, *83*, 105–118. https://doi.org/10.1101/SQB.2018.83.037614

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Shields, C. N., & Gremel, C. M. (2020). Review of Orbitofrontal Cortex in Alcohol Dependence: A Disrupted Cognitive Map? *Alcoholism: Clinical and Experimental Research*, *44*(10), 1952–1964. https://doi.org/10.1111/ACER.14441

Takagi, Y., Hunt, L. T., Woolrich, M. W., Behrens, T. E., & Klein-Flügge, M. C. (2021). Adapting non-invasive human recordings along multiple task-axes shows unfolding of spontaneous and over-trained choice. *ELife*, *10*. https://doi.org/10.7554/eLife.60988

Tang, E., Mattar, M. G., Giusti, C., Lydon-Staley, D. M., Thompson-Schill, S. L., & Bassett, D. S. (2019). Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, *22*(6), 1000–1009. https://doi.org/10.1038/s41593-019-0400-9

Testard, C., Brent, L. J. N., Andersson, J., Chiou, K. L., Negron-Del Valle, J. E., DeCasien, A. R., Acevedo-Ithier, A., Stock, M. K., Antón, S. C., Gonzalez, O., Walker, C. S., Foxley, S., Compo, N. R., Bauman, S., Ruiz-Lambides, A. v., Martinez, M. I., Skene, J. H. P., Horvath, J. E., Higham, J. P., … Sallet, J. (2022). Social connections predict brain structure in a multidimensional free-ranging primate society. *Science Advances*, *8*(15), 5794. https://doi.org/10.1126/SCIADV.ABL5794/SUPPL_FILE/SCIADV.ABL5794_SM.PDF

Theves, S., Fernández, G., & Doeller, C. F. (2020). *The Hippocampus Maps Concept Space, Not Feature Space*. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

Vaidya, A. R., & Badre, D. (2022). Abstract task representations for inference and control. *Trends in Cognitive Sciences*, *26*(6), 484–498. https://doi.org/10.1016/J.TICS.2022.03.009

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). How to build a cognitive map. *Nature Neuroscience 2022 25:10*, *25*(10), 1257–1272. https://doi.org/10.1038/s41593-022-01153-y

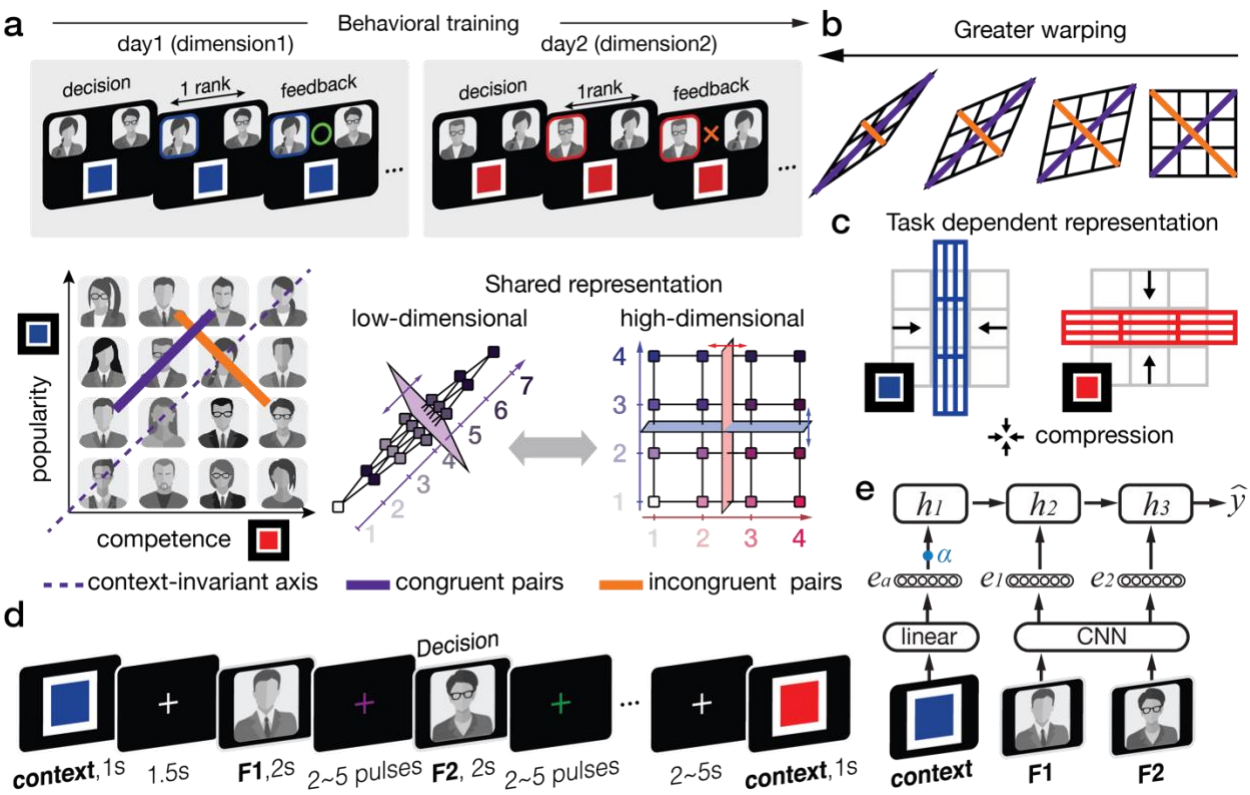


Figure 1. Experimental task and the interaction between cognitive control and representations of relational structure

a. Behavioral training for learning the relationships between 16 faces in two independent social hierarchy dimensions (competence and popularity). To represent the relational task structure, brain areas could build a representation containing relationships in both dimensions that can be shared across two task dimensions (shared representation). Different levels of cognitive control during learning congruent and incongruent relationships could influence the dimensionality of the shared representation in the brain. (1) The brain could construct a low-dimensional shared representation if the brain updates the ranks of faces from the win-loss ratio regardless of the current task-relevant dimension. A low-dimensional representation would facilitate generalization by allowing the same decision policy (purple plane) to be applied to inferences about relationships in both task dimensions. (2) The brain could construct a high-dimensional (2D) representation to effectively resolve conflicting task-irrelevant information on incongruent trials when learning relationships in the current task-relevant dimension. The high-dimensional representation enables accurate, flexible decision making across both inference task contexts.

b. Individual differences in two-dimensional (2D) shared representations. Different levels of cognitive control might influence the distance between congruent (purple) and incongruent relationships (orange) during the formation of the cognitive map. This could further cause individual differences in representational geometry in terms of levels of warping along the context-invariant axis.

c. One-dimensional (1D) task-dependent representations. To facilitate effective value coding for the current task goal, the brain might dynamically expand multidimensional map-like representations along the task-relevant axis and/or compress the task-irrelevant information according to the current task goals.

d. An example trial of the fMRI experiment. While stimuli were sequentially presented in the fMRI, participants were asked to make novel inferences about the relationships between F1 and F2 and select the one who is a higher rank in the given social hierarchy dimension at the time of F2 presentation. The task-relevant dimension was signaled by a colored context cue at the beginning of every trial and randomly intermixed across trials.

e. Recurrent neural network (RNN) model architecture. The RNN was trained to perform the same task with human participants. The model used a convolutional neural network (CNN) to process images and long short-term memory (LSTM) to process the sequence of inputs over time. e: embeddings; h: hidden layers; $\hat{y}$: outcome; α: the level of degradation associated with additional ablation simulation.

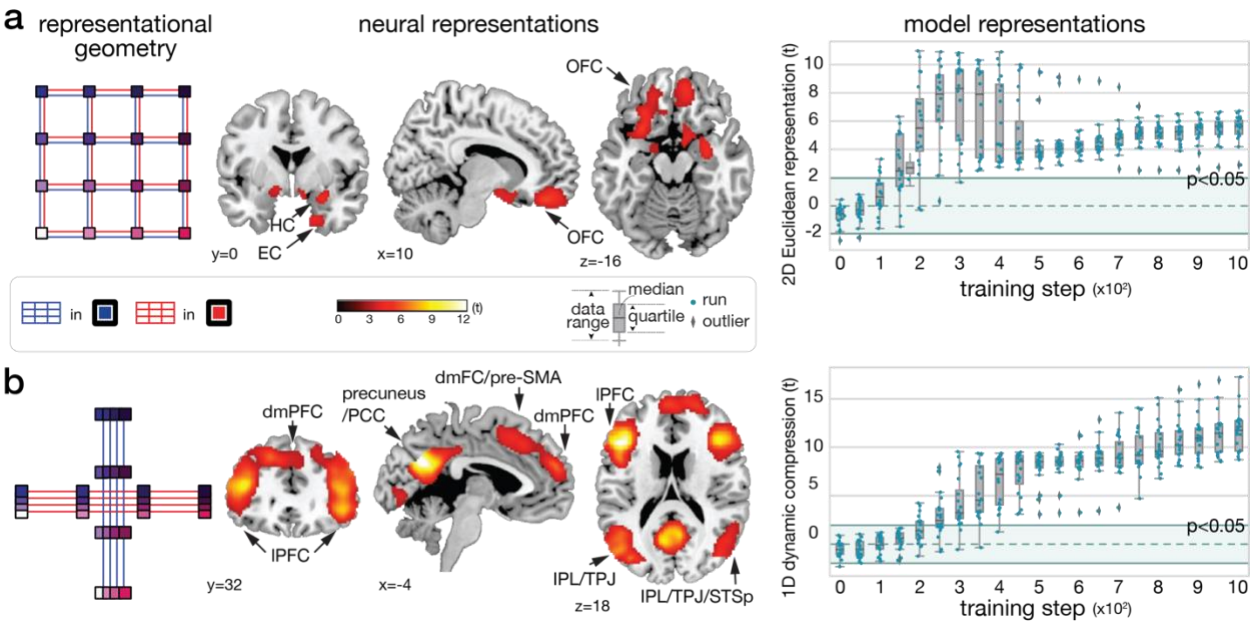


Figure 2. Simultaneous 2-D context-invariant and 1-D orthogonal context-dependent representations in the brain and the RNN

Simultaneous representations of different geometrical structures (left panels) were found in both the brain (middle panels) and the RNN (right panels). Different representational geometry models were fit against each other within a single general linear model (GLM) and tested with representational similarity analysis by estimating pairwise distances between individuals from their activity patterns' dissimilarity in the brain or in the hidden layers of the RNN.

a. RSA results of the 2-D shared representation RDM (left panel) are shown in the hippocampus (HC), entorhinal cortex (EC), and orbitofrontal cortex (OFC) in the brain (middle). The same analysis of the RNN show changes in the level of representation in the hidden layers over the course of training (right panel).

b. RSA results of the 1-D orthogonal representation in the brain and RNN. Significant effects are shown in the lateral prefrontal cortex (lPFC), dorsomedial frontal cortex (dmFC) and prefrontal cortex (dmPFC), posterior cingulate cortex (PCC), and inferior parietal lobule (IPL). RSA shows timecourse of emergent 1-D representation during RNN training (right panel).

a.and b. All the t-statistic brain maps are displayed at a cluster-corrected threshold $p_{TFCE}<0.05$ over the whole brain for all brain regions. Box plots show the levels of representations (t values) over 20 runs (a dot indicates each run). Box, lower and upper quartiles; line, median; whiskers, range of the data excluding outliers; ♦ sign, the whiskers' range of outliers.

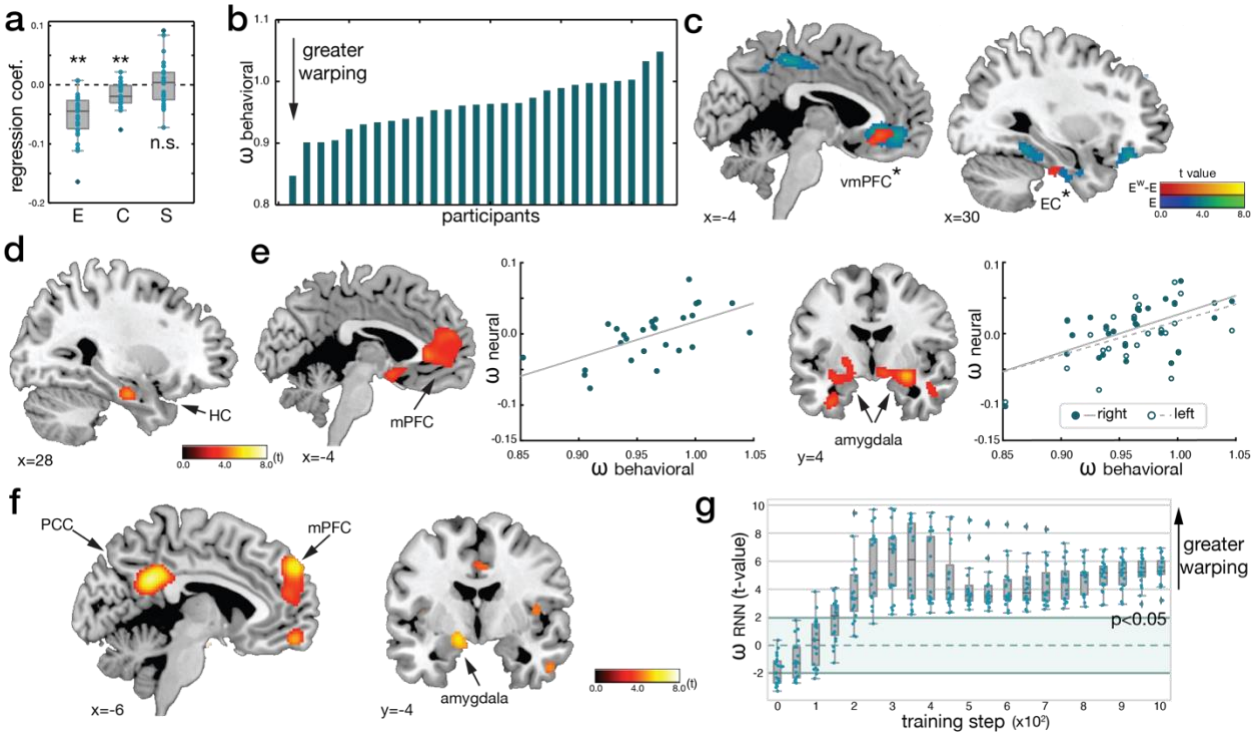


Figure 3. Evidence of warping in the shared representation

a. Multiple linear regression results show that both the Euclidean distance of inferred trajectory for novel inferences of the relationship between faces (E) and the context congruency (C), but not the context switching (S), significantly explain variance in reaction times (RT), while competing with other distance terms (**Eq 2**).

b. individual differences in the level of warping inferred from RT ($\omega_{behavioral}$). Smaller $\omega_{behavioral}$ values correspond to greater warping along the context-invariant axis.

c. The entorhinal cortex (EC) and ventromedial prefrontal cortex (vmPFC) encode the Euclidean distance of the inferred trajectory in subjective warped representations. Additional variance in BOLD activity at the time of decisions explained by Euclidian distances in subjective representations (E^W) over and above the Euclidian distance in the true (non-warped) 4x4 social hierarchy is shown in red-yellow ($p_{TFCE}<0.05$ small volume corrected within anatomically defined independent ROIs; denoted by ✻ next to EC and vmPFC). Euclidean distances defined on the basis of true (non-warped) representations (E) are shown in blue-green. For visualization purposes, the whole-brain map is thresholded at $p<0.005$ uncorrected.

d. Warping was observed at the group level in HC activity patterns (ω_{neuro}), among other regions, along the context-invariant axis. Maps are displayed at a cluster-corrected threshold of pTFCE<0.05 over the whole brain for all brain regions.

e. The level of warping (ω_{neuro}) in the medial prefrontal cortex (mPFC) and amygdala correlated with individual differences in behaviorally estimated warping ($\omega_{behavioral}$). Each dot indicates a participant in the scatter plots. Maps are displayed using conventions in (d).

f. Intersubject RSA (IS-RSA) analysis showed that the representational geometry in the mPFC, posterior cingulate cortex (PCC), and amygdala reflects individual differences in behavioral warping ($\omega_{behavioral}$) across

participants. Displayed clusters of activation correspond to t-values compared to baseline computed from 1000 permutations with random shuffling at the threshold $p_{FDR}<0.05$.

g. Timecourse of changes of the level of warping (ω_{neuro}) estimated from the hidden layer representations in the RNN across training. Notably, the RNN representation remained warped even when the model performed inferences with a high-level accuracy.

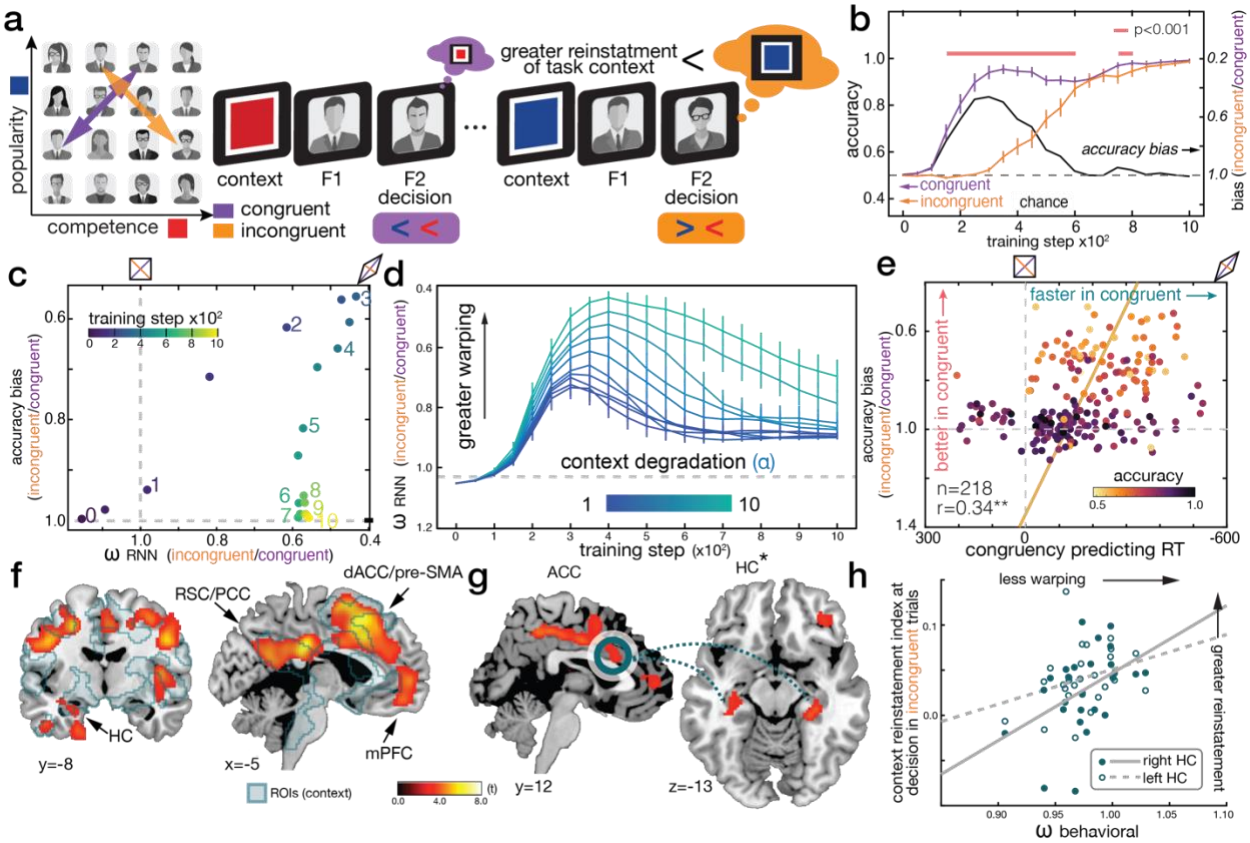


Figure 4. Relationship between context and warping in subjective representations

a. In the current task, participants needed to reinstate task-relevant context information to choose the appropriate decision. This context reinstatement was critically important especially for resolving conflict during inferences on incongruent relationships but not strictly required for making inferences on congruent relationships. This suggests task-relevant context information reinstatement might be observed during inferences on incongruent compared to congruent decisions.

b. Timecourse of the mean inference performance of the RNN across training on the congruent (purple line) and incongruent trials (orange line). The black line indicates the ratio between them, indicating the relative performance accuracy in the incongruent compared to congruent trials. This accuracy bias correlates with timecourse changes in the level of warping in the RNN representation (**Fig.3g**) along the context-invariant axis, suggesting the tendency to learn congruent compared to incongruent relationships earlier could cause warping in the shared representations (ω_{RNN}).

c. The trajectory of the accuracies and accuracy bias between incongruent compared to congruent pairs for the RNN as a function of the level of warping (ω_{RNN}). Each dot indicates the mean accuracy bias across runs at different training steps (0: the beginning to 10: the last training steps) and their mean level of warping at the same training step.

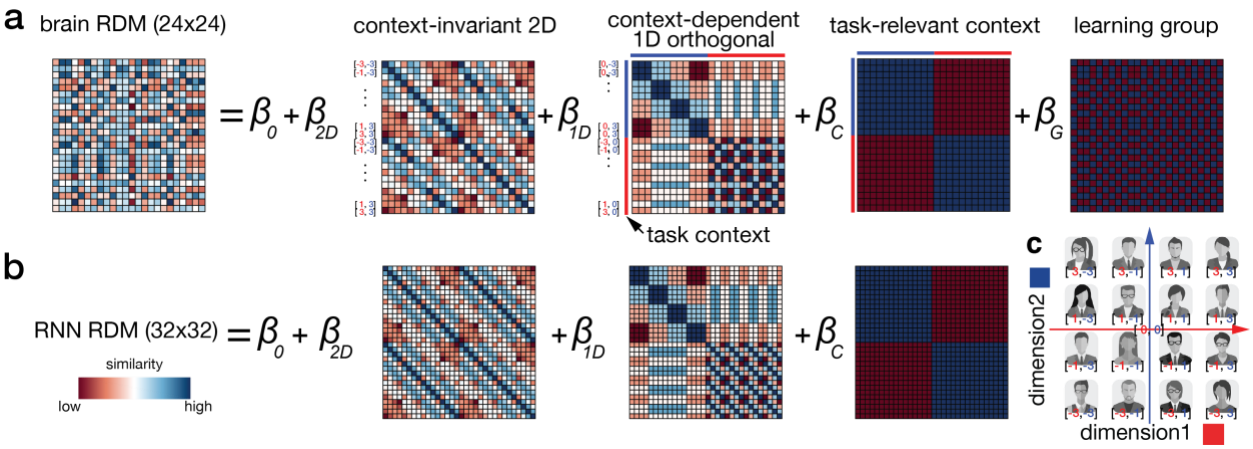
d. Results of ablation simulation. Greater degradation of the task-relevant context information in the RNN (increases in α) incurs greater warping in the shared representation in RNN across training steps, supporting the causal influence of the efficient use of task-relevant context information on constructing high dimensional shared representations (**Fig.1a**).

e. The relationship between individual differences in the accuracy bias estimated from behavioral data acquired during training from the same cohort of the fMRI participants (n=218) and the regression coefficient of context congruence predicting reaction times (RT) in the multiple linear regression. The regression coefficient indicates the degree to which the variance of reaction times (RT) is explained by context congruence while competing with other distance measures (analogous to **Fig.3a**). This shows that the participants who made faster inferences in congruent than incongruent trials, which could be associated with greater warping in the representation, were more likely to make accurate inferences selectively on congruent relationships but less accurate for inferences on incongruent relationships.

f. The activity patterns specific to the current task-relevant context information were reinstated greater at the time of inferences on the incongruent than congruent trials in the hippocampus (HC), medial prefrontal cortex (mPFC), dorsal anterior cingulate cortex / pre-supplementary motor area (dACC/pre-SMA), and retrosplenial cortex/ posterior cingulate cortex (RSC/PCC). The maps are displayed at a cluster-corrected threshold $p_{TFCE}<0.05$ over the whole brain. The independently defined ROIs including all brain areas showing context-specific activity patterns at the time of context cue presentation are shown in green outline.

g. Results of psychophysiological interaction (PPI) analysis showing functional connectivity between the ACC seed ROI that showed greater activation at the time of inferences on incongruent compared to congruent trials and the bilateral HC increased when making inferences on the incongruent compared to inferences on congruent pairs. The maps are displayed at a cluster-corrected threshold $p_{TFCE}<0.05$ over the whole brain (left panel) except for the HC (right panel), where we used a threshold of $p_{TFCE}<0.05$ corrected within anatomically defined ROI (denoted by ⁎), due to our strong *a priori* hypothesis of encoding context information or shared representation.

h. Individual differences between behaviorally estimated warping ($\omega_{behavioral}$) in subjective representations and the level of task context reinstatement at the time of incongruent inferences in left and right HC (**Supplementary Figure 7**). Each circle denotes a subject for left (open circles) and right HC (closed circles).

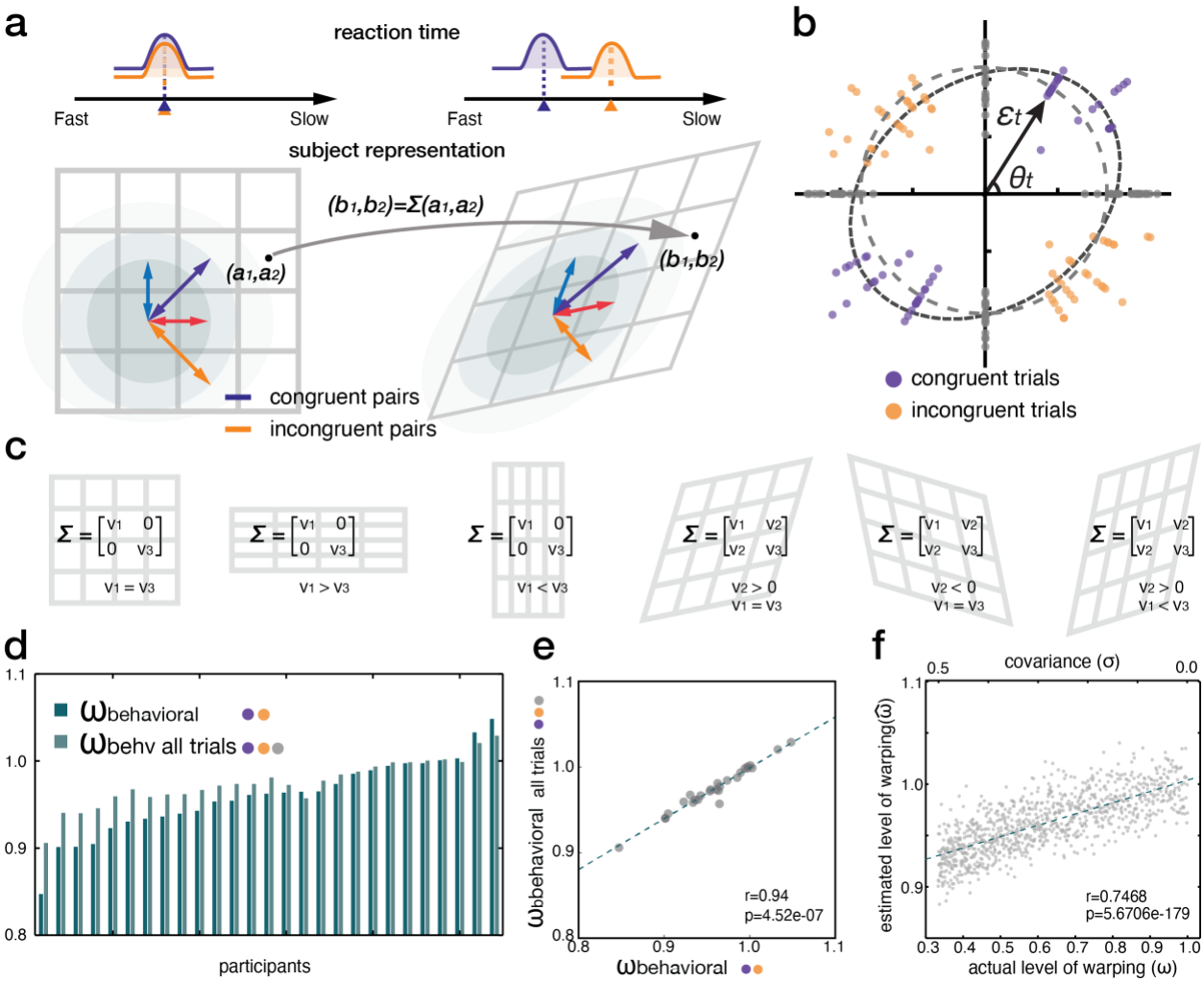


Supplementary Figure 1. The multivariate representational similarity analysis (RSA) using a general linear model (GLM) to examine the representational geometry in the brain and RNN

a. The representational geometry of the 2D shared representation across contexts and 1D task-dependent representation were tested from the representational dissimilarity matrix (RDM) estimated from Euclidean distances between representations of faces in two different task contexts across whole brain areas while controlling potential influences of task context and learning group.

b. The GLM to test the representational geometry of RNN. Considering the training procedure difference, the RDM of the learning group was not inputted as the regressor of non-interest.

c. The coordinates of 16 faces that indicate their position in the 2D shared representation. It was used to compute pairwise Euclidean distances between faces and create the model RDMs. The coordinates of the task-irrelevant axis were set to zero when computing Euclidean distances between faces in the context-dependent 1D orthogonal representation.



Supplementary Figure 2. The procedure of approximation of the level of warping in subjective representations from the reaction times (RT) of inferences

a. A schematic illustration of the rationale underlying the analysis for approximation of the subjective representation from the RT. RT during inferences inversely correlate with the Euclidean distances of inferred trajectories (**Fig.3a**). Therefore, if the representation is not warped (left), RTs of inferences on congruent relationships (purple) should not differ from those on equidistant incongruent relationships (orange). On the other hand, in the warped representation, Euclidean distances of congruent relationships are elongated and those of incongruent relationships are shortened (right). Therefore, if the representation is warped along the context-invariant axis, the RTs would be faster for inferences on congruent than incongruent relationships.

b. To test for the RTs were varied by the angles of inferred trajectories (θ) while controlling for their Euclidean distances, we computed the covariance matrix (Σ) from the logarithm distribution of -RTs (considering their negative correlation) normalized by the distances of inferred trajectories (ε) across trials (**Eq.3-6**). If the representation is not warped the circular distribution of ε should not be biased. If not, the circular distribution should be elongated along the context-invariant axis. From Σ of each participant, we

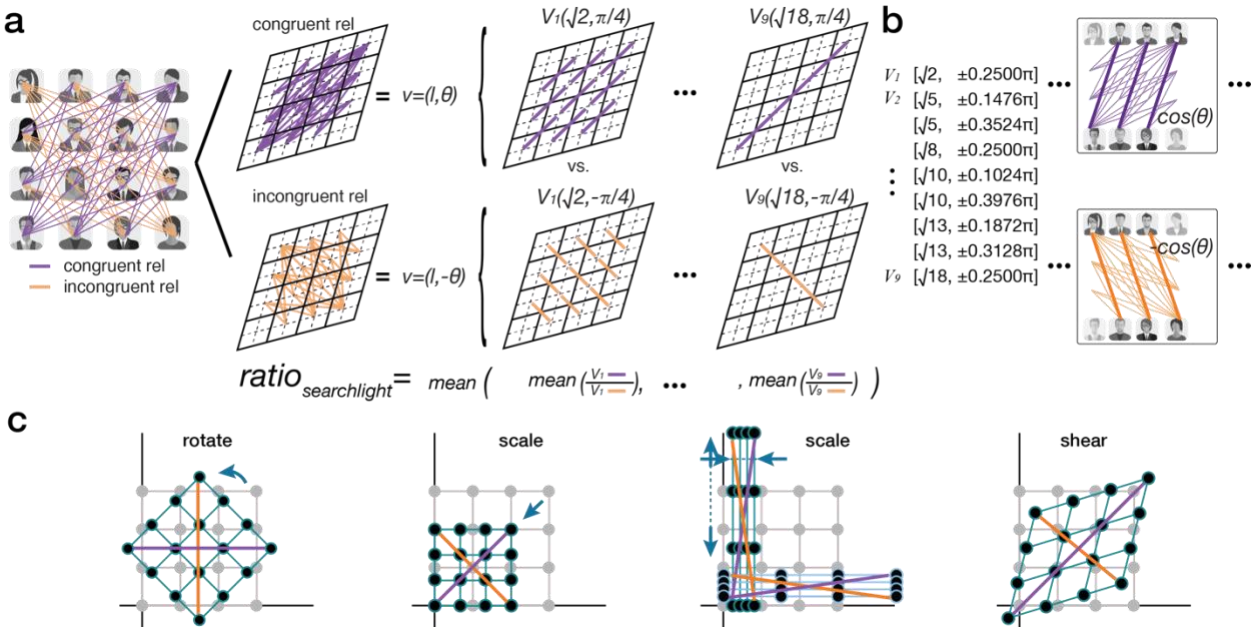
computed the level of warping in the shared representation ($\omega_{behavioral}$) as the ratio of the distances of incongruent pairs to those of congruent pairs.

c. Σ allows capturing not only the potential warping along the context invariant axis but other possible transformations, if there are any, in the shared representation.

d. $\omega_{behavioral}$ can be estimated from the Σ from the distribution of ε including those of $\cos(\theta_t)$ equals to 0 or 1 (grey dots) as well. The level of warping was estimated from RTs of each participant including all trials ($\omega_{behavioral\ all\ trials}$; lighter color) juxtaposed to the sorted $\omega_{behavioral}$ that were estimated only from RTs of congruent and incongruent trials (darker color).

e. The rank of $\omega_{behavioral\ all\ trials}$ was barely different from the rank of $\omega_{behavioral}$, suggesting results of the analysis using $\omega_{behavioral}$ won't be different when using $\omega_{behavioral\ all\ trials}$ instead.

f. Simulation results to validate the current methods. The level of warping estimated from RTs ($\omega_{behavioral}$) linearly correlates with the modulated actual level of warping (ω) in the hidden subjective representations. To do this, we generated synthetic RTs by applying the mean regression coefficients and intercept estimated from participants' data to Euclidean distances of inferred trajectories of the same inference trials with the fMRI task on the representations that we modulated the levels of warping (ω). We modulated ω while applying different levels of Σ while varying σ (v_2 and v_3 components in Σ) from 0.0 (no warping) to 0.5 (large warping). From the synthetically generated RTs, we estimated $\omega_{behavioral}$ using the same method.

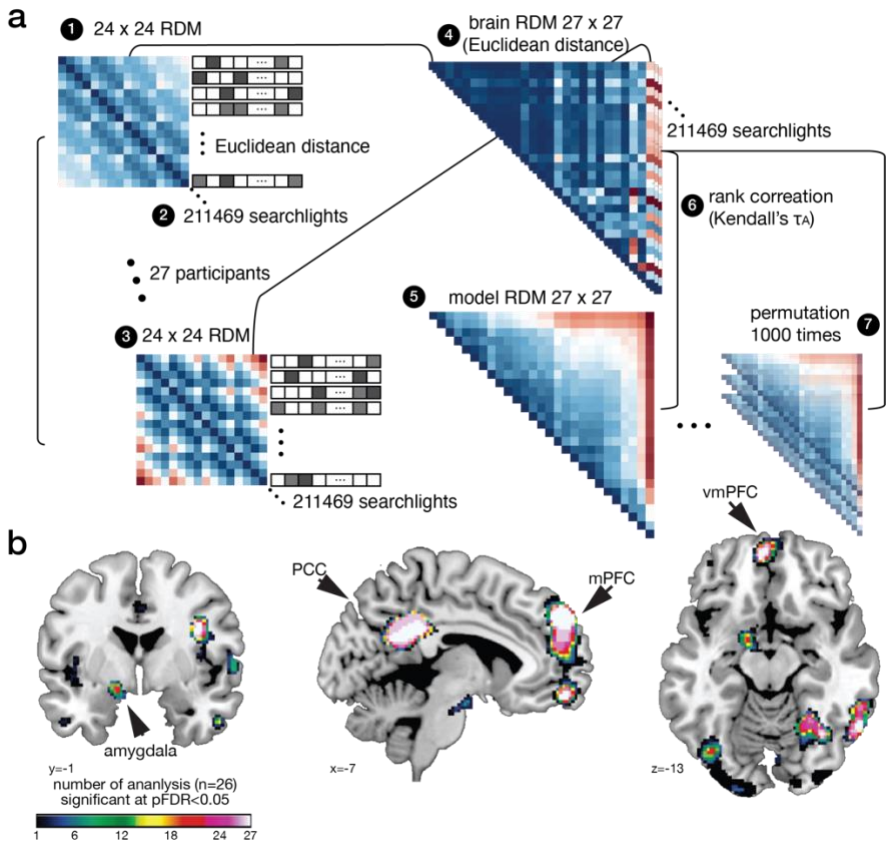


Supplementary Figure 3. The procedure to measure the level of warping (ω_{neural}) in the shared representation from activity patterns in the brain

a. To estimate ω_{neural}, we compared neural activity patterns of all possible pairs of congruent and incongruent relationships and computed the pairwise pattern dissimilarity using Euclidean distances. The sets of pattern dissimilarity estimated from congruent pairs (purple) were separated from those of incongruent pairs (orange). We computed the mean pattern dissimilarity for the pairs sharing the same vector in which both angles (θ) and Euclidean distances (l) between two faces on the 2D social hierarchy are the same. Then, we computed the ratio of the mean pattern dissimilarity of the congruent pairs sharing the same vector (e.g. $v^C = [l, \theta^C]$) to that of the equidistant vectors of incongruent pairs that had symmetric vector angle (i.e. $v^I = [l, \theta^I]$ where $cos(\theta^C) = -cos(\theta^I)$). To test for warping, we computed the mean ratio between pattern dissimilarity of congruent pairs and that of their symmetric incongruent pairs in each searchlight ROI across the whole brain (**Eq.10**).

b. The nine vector sets ([Euclidean distance, vector angle]) are used for computing ωneural using representations of all 16 faces. In the additional analysis, we tested for warping using representations of 12 faces while excluding those whose rank was the highest or lowest in both social hierarchy dimensions. The number of congruent/incongruent pairs sharing the same vector changes when excluding the representations. This additional analysis allowed to test for warping in the representation was not driven by some representations having extreme ranks.

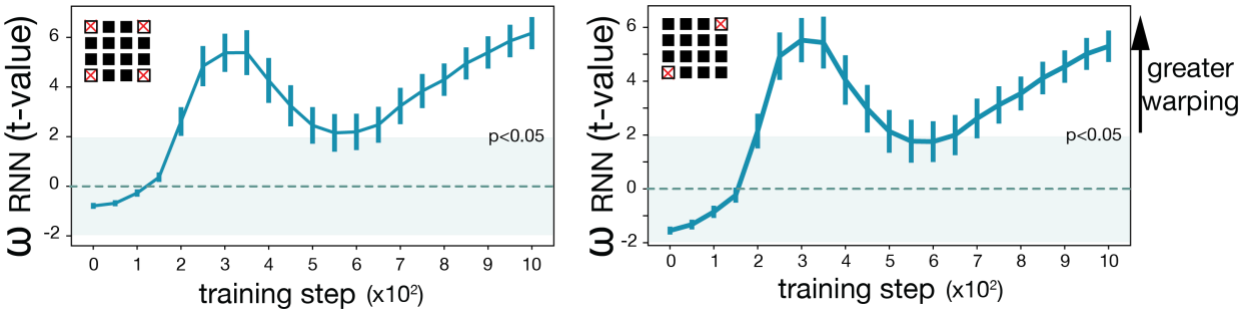
c. This method allows testing the warping only dependent on the shear transformation while independent of other affine transformations (translation, rotation, or scaling in either dimension).



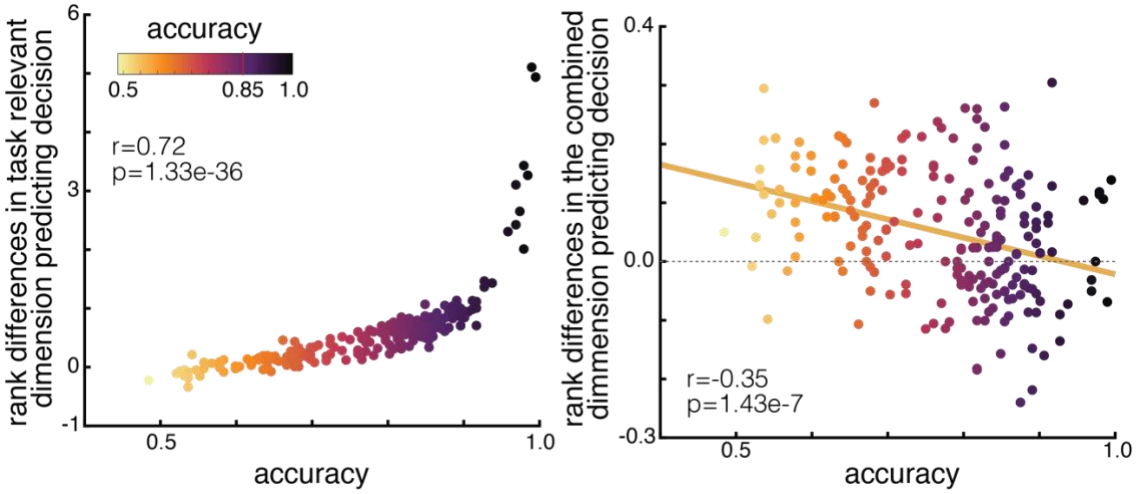
Supplementary Figure 4. intersubject representational similarity analysis (IS-RSA)

a. The procedure of IS-RSA: (1) creating a representational dissimilarity matrix (RDM) by comparing the mean activity patterns associated with each face presented in two task-relevant contexts using Euclidean distances; (2) An RDM was computed from each of 211469 searchlight ROIs across the whole brain; (3) This procedure was repeated across participant (n=27); (4) The brain RDM for IS-RSA was created by comparing RDMs estimated from the same searchlight ROI across participants using Euclidean distances. This created an n×n sized RDM per ROI; (5) The model RDM for IS-RSA was computed from pairwise dissimilarity of the behaviorally estimated level of warping ($\omega_{behavioral}$; **Fig.3b**) across participants. (6) The rank correlation between the brain RDM of each searchlight and the model RDM was computed using Kendall's τ_A across the whole brain; (7) The rank correlation between the brain RDM and the model RDM was computed repeatedly while randomly shuffling the labels of $\omega_{behavioral}$ 1000 times to create the baseline distribution of Kendall's τ_A. The t and p values were computed over the baseline using a normal distribution generated from the mean and standard deviation of the distribution of 1000 Kendall's τA. This procedure was repeated for each searchlight ROI across the whole brain.

b. The results of IS-RSA while including data from 26 participants while excluding one participant (27 times of leave-one-subject-out analysis). The color map indicates the number of analyses showing significant intersubject similarity across 26 participants at the threshold $p_{FDR}<0.05$

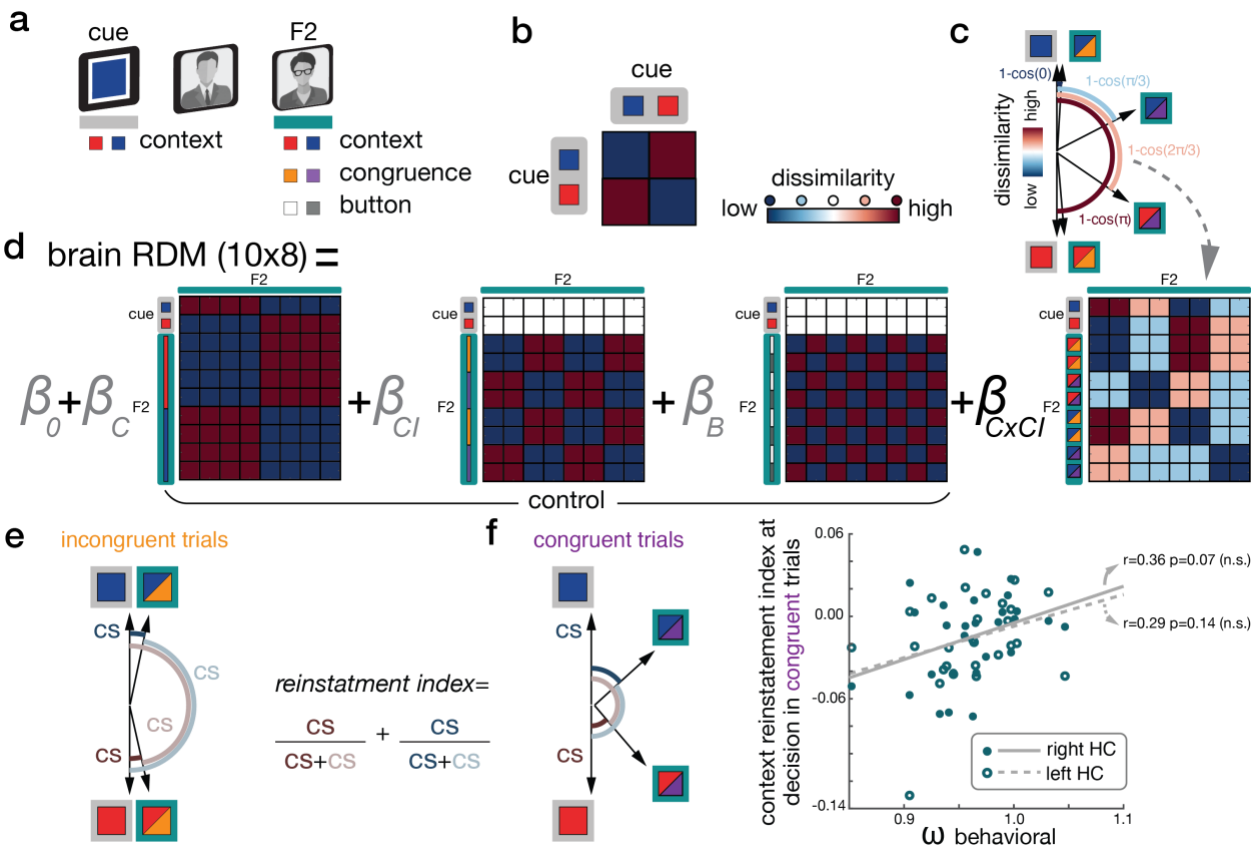


Supplementary Figure 5. Time course changes of level of warping in RNN representations while excluding the representations of faces who had the highest or lowest rank in either social hierarchy dimension (left); those while excluding the representations of faces who had the highest and lowest rank in both social hierarchy dimensions (right). These results show that the warping was not driven by the representations of certain faces but observed systematically across all face representations.



Supplementary Figure 6. Relationships between individual differences in inference accuracy and their tendency of using model-based (left panel) and model-free (right panel) decision values (DVs)

In the current task, participants should use rank differences in the task-relevant dimension (model-based DV equivalent to the distances between individuals in the task-relevant dimension in the high dimensional representation in **Fig.1a**) to make correct inferences, and they should not use rank differences in the combined dimension (model-free DV equivalent to the distances between individuals in the low dimensional representation in **Fig.1a**). The participants who showed low accuracy performance (lighter color) were more likely to use the model-free DV to make inferences while those showed high accuracy (darker color) performance were less influenced by model-free DV. Participants needed to reach above 85% accuracy to continue participating in the fMRI experiment.



Supplementary Figure 7. The multivariate representational similarity analysis (RSA) to test for the greater task context reinstatement at the time of inferences on the incongruent compared to congruent relationships

a. For an additional RSA, we estimated the mean activity patterns of each task context (red: competence; blue: popularity) at the time of cue, context presentation, and mean activity patterns of F2 grouped by task-relevant context (red: competence; blue: popularity), context congruence (orange: incongruent trial; purple: congruent trial), and the button press (white: right choice; grey: left choice) at the time of decision

b. The brain areas containing task context-specific activity patterns at the time of cue were identified by comparing the mean activity patterns at the time of cue associated with two task contexts.

c. A graphical model for creating a model representational dissimilarity matrix (RDM) to test the hypothesis of great context reinstatement at the time of inferences on the incongruent relationship compared to inferences on the congruent relationship. Each arrow indicates the vector of representation of two task contexts and eight decision types. The model RDM reflects $1 - cos(angle)$ between two vectors. i.e. Dissimilarity between two task context-specific representations: $1 - cos(\pi)$; Dissimilarity between representations of the task-relevant context (red or blue) and incongruent decision under the corresponding task context (red+orange or blue+orange): $1 - cos(0)$; Dissimilarity between representations of the task-relevant context (red or blue) and congruent decision under the corresponding task context (red+purple or blue+purple): $1 - cos(\pi/3)$.

d. A brain RDM was created per searchlight ROI across the whole brain. The 8×10 sized RDM was created by comparing the activity patterns estimated at the two context cue presentations and those of eight types

of inferences. To test the hypothesis while controlling for the other potential confounding representations, we used a general linear model (GLM) in which we included other regressors. The three regressors of non-interests included current task-relevant context (C), congruence/incongruence (CI), and button press (B). Using the GLM, we tested the hypothetical representation (the interaction effect between C and CI shown in **c**). This representation was tested within the independently defined ROI including the brain areas having context-specific representation (as the result of **b**).

e. To test the effects of efficient task-context reinstatement on individual differences in warping, we computed the reinstatement index per participant. The reinstatement index indicates to what extent the hippocampus (HC) activity patterns at the time of incongruent inferences reflect the patterns representing the task-relevant context information relative to the patterns representing the task-irrelevant context information. We used cosine similarity (CS) between vectors of those representations to compute the reinstatement index per participant.

f. Left: The same reinstatement index was estimated from HC activity patterns associated with inferences on congruent relationships, instead. Right: The relationships between the reinstatement index of congruent trials and individual differences in levels of warping estimated from reaction times ($\omega_{\text{behavioral}}$) (similar to **Fig.4h**. using activity patterns associated with inferences on congruent relationships)