The representational geometry of cognitive maps under dynamic cognitive control

Seongmin A. Park [1,5], 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 the brain abstracts non-spatial relationships between entities or task states 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 stable map-like representations in a medial temporal lobe and orbitofrontal cortical network that reflected both task-relevant and irrelevant dimensions and dynamic, orthogonal representations of only relevant task dimensions in a frontoparietal network. These representational motifs also emerged with distinct temporal profiles over the course of training in the RNN, with map-like representations appearing first. We further show that increasing control demands due to incongruence (conflicting responses) between current task-relevant and irrelevant dimensions impact the geometry of subjective representations, and the degree of this effect further accounts for individual differences in cognitive control. Taken together, our findings show how complementary representational geometries balance stability and behavioral flexibility, and reveal an intricate bidirectional relationship between cognitive control and cognitive map geometry.

[1] Center for Mind and Brain, UC Davis, USA

[2] Department of Computer Science, UC Davis, USA

[3] Center for Neuroscience, UC Davis, USA

[4] Department of Psychology, UC Davis, USA

[5] Institute of Cognitive Science Marc Jeannerod, UMR5229, CNRS, Lyon, France

* 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 adaptive behavior. This behavioral flexibility depends on the development of representations that can be shared across multiple tasks, as well as the capacity for flexible top-down selection of information according to changing task goals (Bernardi et al., 2020; Flesch, Juechems, 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 into a unitary combined representation, even when multiple task dimensions are sampled in different task contexts (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 the benefits of 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, 2023; Musslick & Cohen, 2021; Rigotti et al., 2013). An intriguing but unexplored possibility is that an analogous tradeoff 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 attention toward 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, Juechems, 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 may apply to dimensions from mnemonic representations, particularly cognitive maps. When information is retrieved from memory, the brain may favor high-dimensional population coding enabling any specific dimensions to be flexibly selected based on the current goal. Moreover, the representational geometry defined by the 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, as is commonplace in the real world.

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, and, in turn, may produce 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 currently task-relevant and irrelevant dimensions. For example, while José is more senior than Gabriela and has more influence in the board room, Gabriela is more popular and, hence, has more influence at a social gathering. Recent findings suggest that the representational geometry of task representations might be determined by task demands such that they facilitate generalization to similar contexts that share a decision policy (i.e., choice rule) (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 congruence versus incongruence play a role in shaping the resulting subjective representations, we would expect the brain to adopt a versatile representational strategy whereby congruence between task dimensions promotes sharing, and therefore generalization across those dimensions, while incongruence promotes segregation, thereby allowing greater flexibility across contexts or tasks with divergent responses. Consequently, congruence may incentivize generalization across the two dimensions which in geometrical terms would produce warping in the cognitive map 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 inferences.

To facilitate effective value computations at the time of decision making, control- and decision-related brain regions may dynamically collapse or compress the multidimensional map-like representations into one-dimensional (1D) representations according to the current task goal (**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 in proficient participants who underwent intensive training (Flesch, Juechems, 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.

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. We found 2D representations that were stable 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-dependent 1D representations with dynamic compression of task-irrelevant dimensions relative to task-relevant dimensions in the dorsomedial frontal cortex (dmFC), lateral prefrontal cortex (lPFC), posterior cingulate cortex (PCC), and inferior parietal lobule (IPL). Consistent with human fMRI, we found both context-invariant 2D 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) learned the relative ranks of 16 individual faces on two social hierarchy dimensions (**Fig.1a**) - competence and popularity - using a similar previously detailed training protocol (Park et al., 2021b). On each trial, participants selected which of two sequentially presented faces ranked higher along one of the two dimensions, which was indicated by a colored context cue. During training both human participants and the RNN learned only about relationships between neighboring entities – i.e., individuals whose rank differed only by one level along the given dimension - through feedback on their decisions. Human participants learned the relationships in each of the two dimensions on different days, but the model learned them in a single block with interleaved 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 that were not seen during training. During this test, contexts were randomly interleaved for the human participants as well as the model, thus requiring subjects to flexibly switch between contextually defined task dimensions. It is important to note that neither participants nor the model were provided with any information about the true 4x4 grid structure of the social hierarchy, nor were they asked to construct it spatially.

Using both human fMRI data and hidden states of the RNN, first, we examined 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.

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 univariate 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, multivariate analyses of BOLD patterns elicited by face presentations identified 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 increasingly and linearly more similarly (measured by pairwise pattern similarity).

While sharing representations across multiple task contexts may benefit generalization and allow faster learning across contexts, 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, Juechems, et al., 2022; Mante et al., 2013; Panichello & Buschman, 2021; Takagi et al., 2021), we tested for evidence that the two dimensions were represented orthogonally, with information along the currently irrelevant dimension being compressed. To test

these hypotheses while controlling for alternative representational geometries in the brain, we examine the extent to which 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 1D orthogonal representations of the relevant dimensions with the irrelevant dimension dynamically compressed 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), precuneus/posterior cingulate cortex (PCC, [-4,-64,28], t=13.60), right inferior parietal lobule (IPL, [46,-64,22], t=5.20) extending into temporoparietal junction area (TPJ), and left IPL extending into 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** in supplementary information for a full list of brain areas surviving correction. Taken together, these representational similarity analyses (RSA) show that the HC, EC and OFC exhibit stable, task-invariant 2D map-like representations that are shared across contexts, in parallel to dynamic, task-dependent 1D orthogonal representations with compression along the irrelevant dimension, in a frontoparietal network ideal for computing context-dependent decision values. The extent to which neural activity patterns reflected 2D and 1D representations was neither positively nor inversely correlated with each other across participants, supporting the idea of simultaneous representations of both types of representational geometry (**Table S2**).

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 components specialized to produce these representations. Just as in the fMRI data, the pairwise distances between representations in the neural network model were explained by the 2D Euclidean distances between individuals in the context-invariant 2D social hierarchy structure (**Fig.2a**), as well as those in the context-dependent 1D orthogonal representations with dynamic compression (**Fig.2b**). The 2D map-like representations emerged at a relatively early phase of training (the representation was significant [p<0.05] at 150 steps of training [the median run: t=4.35, p=1.65e-5]; the peak was at training step 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 model also utilized task-dependent orthogonal representations, and this tendency for dynamic compression increased continuously over training (significant at training step 250 [the median run: t=4.53, p=1.86e-5]; peak at training step 1000 [the median run: t =11.75, p<1e-4]). Thus, both shared 2D and orthogonal 1D representations emerged as complementary, efficient representations to solve our relational inference task, but with different time courses.

Effects of cognitive control on representational geometry

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 the extent to which 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 after controlling for the distances, subjects were slower on incongruent than congruent trials. Subsequently, by computing effects of the degree of congruence (defined as the vector angles of inferred trajectories, θ) on the relationship between Euclidian distance and RTs (ε), we approximated the hypothesized subjective representational geometry and the implied level of warping per participant (**Supplementary Fig.2; Eq.3-9** for the procedure and confirmatory simulations using synthetic data). The levels of warping were estimated as the ratios of the incongruent diagonal distances to the congruent diagonal distances. The mean ratio between diagonals was 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 a 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 the diagonal ratio (relative values) 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* regions of interest (ROIs) anatomically based on our previous findings(Park et al., 2020) and found that the Euclidean distances in the *warped* representations inferred from RTs could explain additional variance in 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 true Euclidean distances between individuals in the 2D social hierarchy (**Fig.3c**). This finding indicates that the hypothesized warped representations explain additional variance in the effect of Euclidian distance on decision-related univariate 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), or on 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 (not 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 in the true 2D social hierarchy. Specifically, whole-brain searchlight-based RSA (**Supplementary Fig.3** and **Eq.10**) of fMRI data revealed this warped geometry 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 S3** for a full list of brain areas). The HC effect was also significant after excluding individuals who either 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 S3** for a full list of brain areas). This finding suggests those individuals with greater representational warping in these regions experienced greater demands on cognitive control during decisions. This relationship between individual differences in representational geometry and 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 S4** for a full list of brain areas surviving at permutation-based pFDR<0.05 based on 1000 permutations). Control leave-one-out 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 performed identical RSA analyses on the hidden layers of the RNN (h_2 and h_3 in **Fig1e**) throughout learning. Motivated by the human fMRI data, we specifically tested for warping in the 2D map-like representation in the hidden layers by comparing distances between representations of congruent and incongruent pairs. 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**). This finding was remarkable, as the warping phenomenon emerged in the RNN simply as a consequence of training it on the same task and without building any additional constraints into the architecture. Notably, warping in the representations persisted even after inference performance had reached an asymptote (**Fig.4b**). Control analyses indicated that warping in the model's representations was still present when excluding boundary faces who had the highest or lowest rank in both dimensions (**Supplementary Fig.5**). Thus, the warping effect could not be due to the representations of any specific face, nor these outlier face positions. Together, these

findings suggest that warping emerges naturally in a standard recurrent neural network trained via backpropagation on our task.

Inefficient use of context 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 examined how incongruence between dimensions experienced throughout cognitive map formation may cause the observed warping and result in different levels of warping 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 for cognitive control processes 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 over the course of 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 in the early stages of training (**Fig.4b**; See **Table S5** for the results of logistic regression). Moreover, this bias in performance toward congruent trials (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 early in 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_c in **Fig.1e** and **Eq.16**) and examined changes in the geometry of 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**), consistent with the notion that the inefficient use of context produced the observed warping.

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 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 tested for stronger reinstatement of these context-specific activity patterns at the time of *decision making* (F2 presentation) on incongruent trials than congruent trials, which identified significant clusters 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) (**Fig.4f**; See **Table S6** for additional regions outside of these ROIs). Importantly, this GLM controlled for the potential effects of the context, incongruence, and motor responses by including these as control RDMs. These findings support the hypothesis that there is 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 first, regions showing greater activity on incongruent than congruent decisions and second, 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 compared to congruent inferences (**Fig.4g** and **Table S7**).

The above behavioral and modeling analyses suggested that participants who ineffectively use context information to resolve incongruency should show greater warping in their representations. To test this idea, we first captured the fidelity with which the context was reinstated by calculating 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 behavior was less influenced by incongruence more faithfully reinstated the appropriate context representation in HC (**Fig.4a**) (**Supplementary Fig.7** and **Eq.11**). This effect was specific to incongruent inferences, because 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 findings suggest a critical link between the utilization of contextual information for effective cognitive control and the observed representational warping.

Discussion

Our findings show how latent non-spatial, abstract cognitive maps are represented when learning 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. First, both a HC-EC-OFC/vmPFC network and the 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 effective selection of the task-relevant dimension, this multidimensional representation was modified into one of two orthogonal 1D number lines suppressing the task-irrelevant dimension in a frontoparietal cortical network and in the RNN. Finally, increased control demands for inferences made in the presence of incongruence produced individual differences in behavioral indices measuring effective control, which predicted the extent to which subjective 2D representations were warped along the context-invariant (congruent) axis. 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 human participants were never explicitly instructed about the underlying latent structure, representations in HC, EC, and OFC/vmPFC 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 task dimensions were represented orthogonally when they were behaviorally-relevant, with the irrelevant dimension compressed, in a separate network of brain regions including the dmFC, lPFC, PCC, and IPL. Importantly, these orthogonal 1D representations afford behavioral flexibility by dynamically compressing task-irrelevant dimensions relative to task-relevant dimensions. 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, Juechems, 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, 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 even after extensive training. 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 (a stable multidimensional representation) (**Fig.1a**) allows for fast learning in a new task context via generalization (which works particularly well for learning congruent relationships in our task) and a single computation across dimensions, thereby mitigating costs of task/context switching (Badre et al., 2020; Garner & Dux, 2023; 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 multidimensional representation that can serve generalization across dimensions, and flexible readout of information depending on the current goal. In the context of our task, computing shared decision 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 wherein the relationships in different dimensions formed a low-dimensional number line started to transition toward orthogonal high-dimensional representations wherein the relationships in the two dimensions became represented on separate axes of a single 2D 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 context-free values based only on win/loss ratio regardless 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 increased training (**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 accord 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 select 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 resolves 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 for future studies concerns how the HC contextual representations and 2D map-like representations are related during goal-directed selection of specific information from cognitive maps.

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 network 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 context-dependent inferences could arise from the subjective representational geometry of cognitive maps. This insight opens a new research direction to better discern individual differences in learning and decision making, and 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), which often exhibit excessive reliance on contextual information (Lake et al., 2017), and lack human-like control processes (Russin 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

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N. J., Habel, U., Schneider, F., & Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: Intersubject variability and probability maps. *Anatomy and Embryology*, *210*(5–6), 343–352. https://doi.org/10.1007/s00429-005-0025-5

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

Benjamini, Y., & Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60. https://doi.org/10.2307/1165312

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., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), E10313–E10322. https://doi.org/10.1073/PNAS.1800755115/SUPPL_FILE/PNAS.1800755115.SAPP.PDF

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2022). *Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals*. https://doi.org/10.48550/arxiv.2203.11560

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

Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and Modulatory Interactions in Neuroimaging. *NeuroImage*, *6*(3), 218–229. https://doi.org/10.1006/nimg.1997.0291

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. (2023). Knowledge generalization and the costs of multitasking. *Nature Reviews Neuroscience*, *24*(2), 98–112. 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 goal directed 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*(2), 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

Kingma, D. P., & Ba, J. L. (2014, December 22). Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1412.6980

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. *Current Opinion in Neurobiology*, *37*, 1–6. https://doi.org/10.1016/j.conb.2015.11.004

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(NOV), 4. https://doi.org/10.3389/neuro.06.004.2008

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). *Building machines that learn and think like people*. https://doi.org/10.1017/S0140525X16001837

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

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., & Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, *26*(4), 490–503. https://doi.org/10.1016/j.mri.2007.08.006

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

Neubert, F.-X., Mars, R. B., Sallet, J., & Rushworth, M. F. S. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), 1–10. https://doi.org/10.1073/pnas.1410767112

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

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

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

Paszke, A., Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, & Gregory Chanan. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*.

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

Russin, J., O'reilly, R. C., & Bengio, Y. (2020). Deep learning needs a prefrontal cortex. *International Conference on Learning Representations (ICLR)*.

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. (2022). A Neural Network Model of Continual Learning with Cognitive Control. *Cognitive Science Society (CogSci 2022)*. https://doi.org/10.48550/arXiv.2202.04773

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

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

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. *The Journal of Neuroscience*, *40*(38), 7318–7325. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

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

Yushkevich, P. A., Amaral, R. S. C., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A. C., Carr, V. A., Chakravarty, M. M., Chételat, G., Daugherty, A. M., Davachi, L., Ding, S.-L., Ekstrom, A., Geerlings, M. I., Hassan, A., Huang, Y., Iglesias, J. E., … Zeineh, M. M. (2015). Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *NeuroImage*, *111*, 526–541. https://doi.org/10.1016/j.neuroimage.2015.01.004

Zilles, K., & Amunts, K. (2010). Centenary of Brodmann's map conception and fate. *Nature Reviews Neuroscience*, *11*(2), 139–145. https://doi.org/10.1038/nrn2776

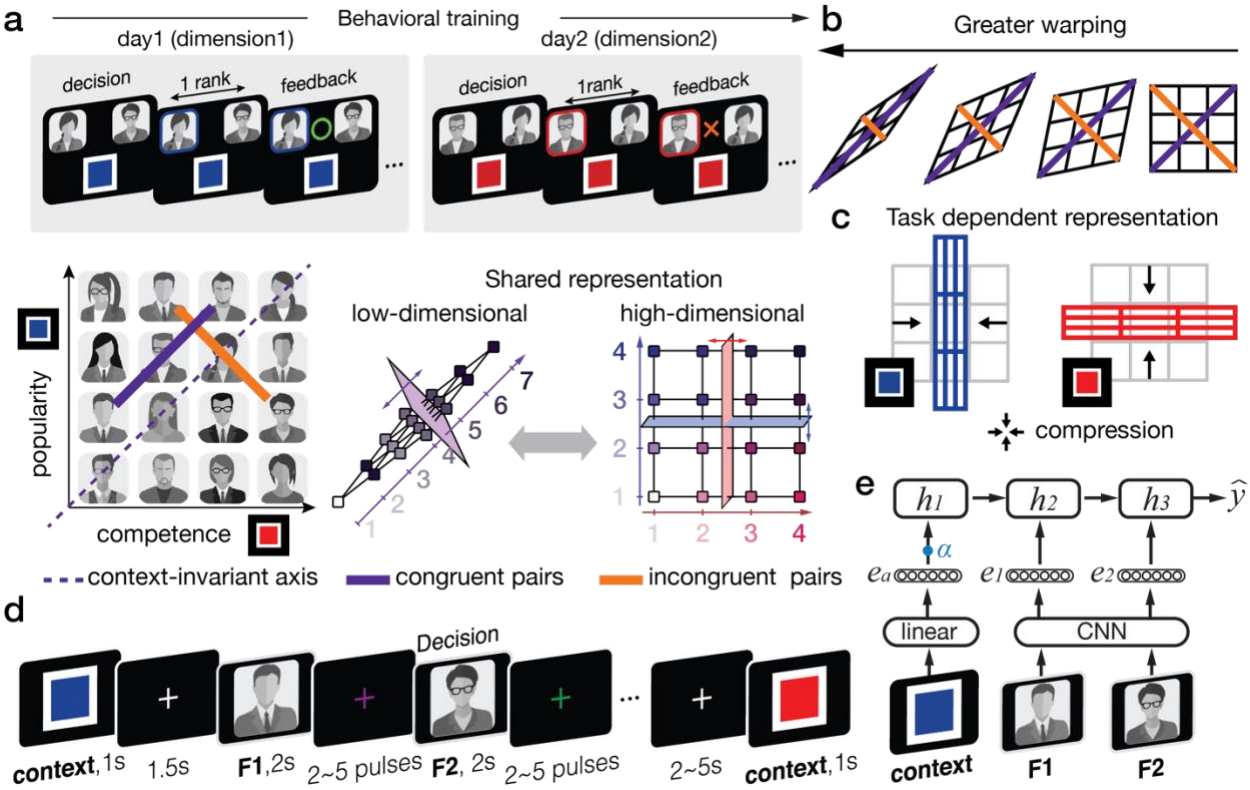


Figure 1. Experimental task and the interaction between cognitive control and representations of relational structure

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 the two task contexts (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 **t**wo-dimensional (2D) shared representations. Different levels of cognitive control might influence the distance between congruent (purple) and incongruent (orange) relationships 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 orthogonal 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 as 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.

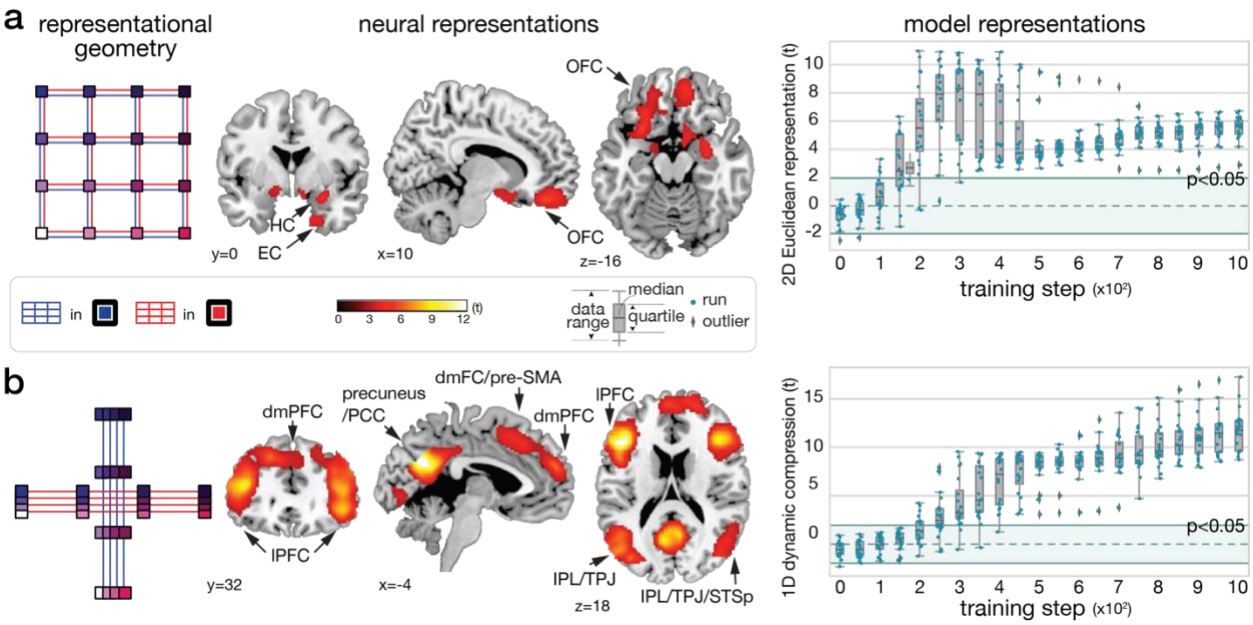


Figure 2. Simultaneous 2-D context-invariant and 1-D orthogonal context-dependent representations in the brain and the RNN

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 (RSA) 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 shows 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.

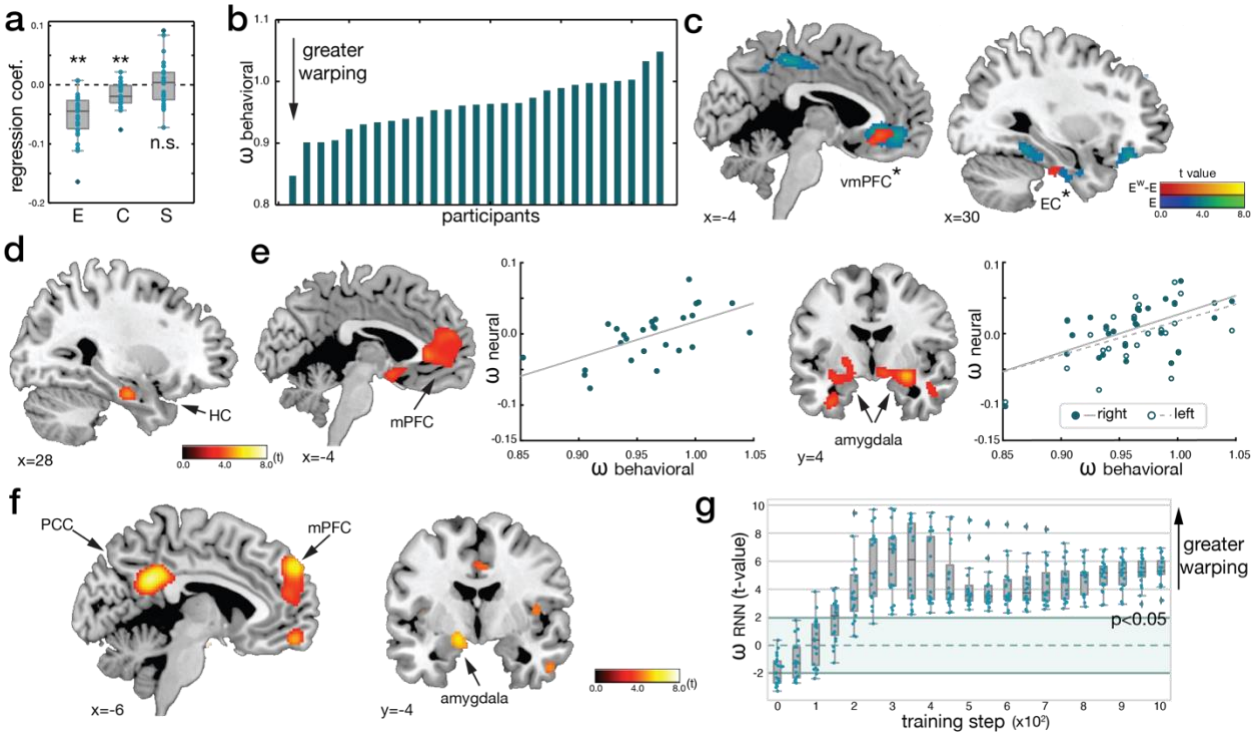


Figure 3. Evidence of warping in the shared representation

a. Multiple linear regression results show that both the Euclidean distance (E) of the inferred trajectory for novel inferences of the relationship between faces 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 (E) 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 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.

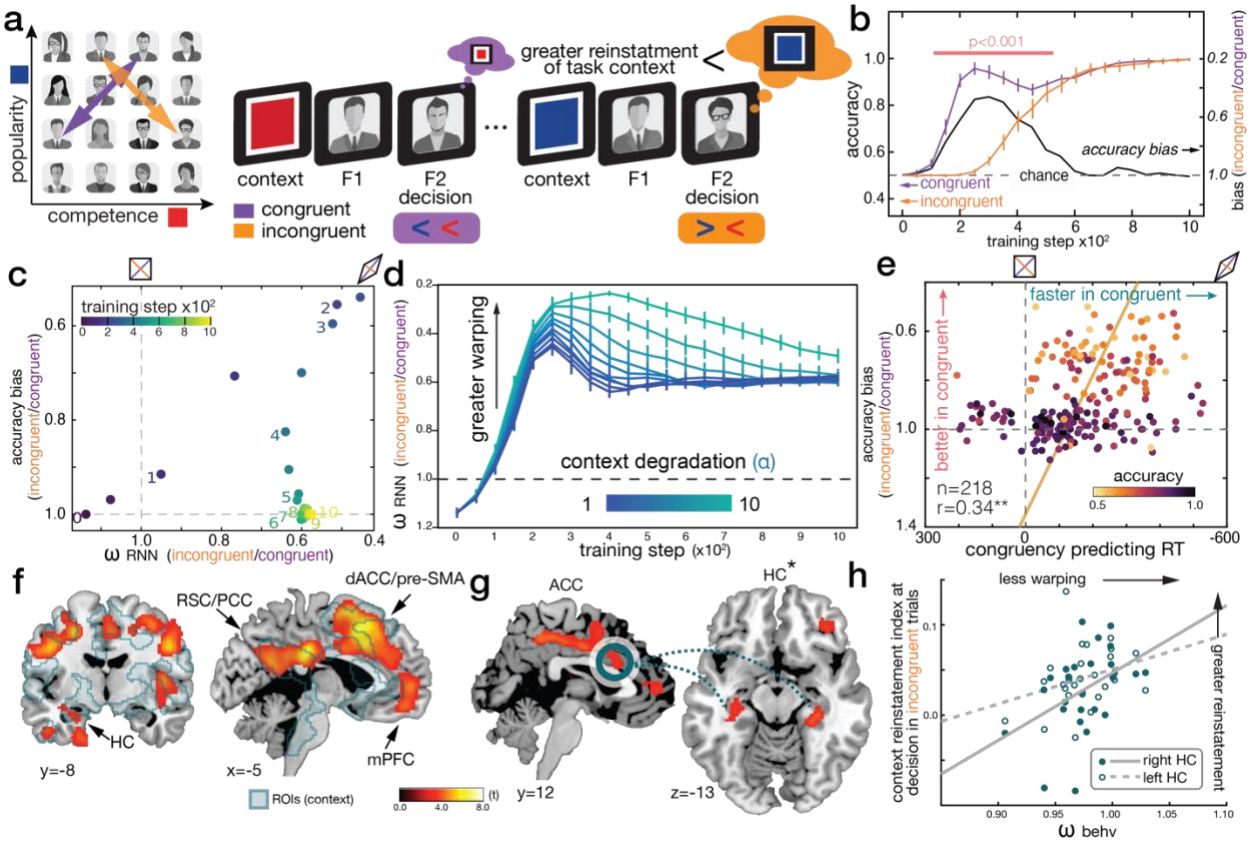


Figure 4. Relationship between context and warping in subjective representations

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 (orange line) trials. 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 step) 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 the RNN across training steps, supporting the causal influence of the efficient use of task-relevant context information on the construction of 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).

Methods

Participants

33 participants (16 female, age range: 19–23, normal or corrected to normal vision) were recruited to the fMRI experiment via the University of California, Davis online recruitment system. Six participants were excluded due to strong head movements (> 3 mm) and technical failure (2 subjects). In total, we analyzed the scans of 27 participants (mean age: 19.37±0.26, standard error mean [SEM]). The study was approved by the University of California, Davis, IRB. Participants gave written consent before the experiment. The sample size was estimated based on a power analysis assuming a medium to a large effect (d = 0.6), resulting in a sample size of 24 to achieve a statistical power of 80% (a = 0.05, two-tailed test).

fMRI experiment

In the current study, we performed novel analyses using the dataset collected in a previous study (see (Park et al., 2020) for detailed information on the experiment design). During fMRI participants made novel inferences about the relationships between pairs of individuals in one of two social hierarchy dimensions – "competence" and "popularity." In each trial of the fMRI experiment, participants were given a context cue (1 s) with the color indicating the social hierarchy dimension that relevant to the task in the current trial, and two face stimuli (F1 and F2) followed sequentially. Each face stimulus was shown for 2 s. Participants were asked to make a binary decision by selecting the face associated with a higher rank in the given social hierarchy dimension while F2 was on the screen. Before fMRI, participants learned the relative ranks between these 16 individuals in two social hierarchy dimensions. Unbeknownst to participants, F1 and F2 were selected from the pairs that had not been directly compared in the given dimension during training. This allowed us to present 12 faces while excluding 4 "hub" faces (see behavioral training) as F1 or F2 in each of two social hierarchy dimensions. To make a correct inference, therefore, participants needed to make transitive inferences using comparisons that they have been trained on. Importantly, we limited the number of pairs that participants learned about during training. This forced participants to make novel inferences through a unique trajectory. As a control experiment, at the end of each trial, a third face was shown and participants were asked to press a button according to the gender of the face. The inter-stimulus interval (ISI) was 1.5 s between the context cue to F1 and 2-5 pulses between faces. The inter-trial interval (ITI) was 2-5 s. To acquire responses to all inferences, the trial in which participants did not make a decision was presented again after a random number of trials. The fMRI experiment comprised two 104 trial blocks. The presentation order was randomized across participants.

Behavioral training

Participants were asked to learn the relationships of 16 individuals in two social hierarchy dimensions for 3 days of behavioral training. The training was comprised of learning and test blocks. For the first two days, participants learned the relationships in each of the two social hierarchy dimensions on different days. The dimension learned on the first day was counterbalanced across participants. During behavioral training, participants were presented with

a pair of individuals and asked to select the one with a higher rank in the given social hierarchy dimension. During learning blocks, participants received feedback on their decisions. No feedback was given during the test blocks. Importantly, during behavioral training and the fMRI experiment, no information hinting at the structure of the social hierarchy was given to participants. Instead, participants were only informed that the pairs shown in the learning blocks have one rank difference in the given dimension. This allowed participants, in principle, to infer the overall social hierarchy structure through transitive inferences. During test blocks, to test if participants could infer the relationships that were not directly compared during learning, participants were asked to infer relationships of pairs of individuals whose ranks differed by one or more rank differences. At the end of day 2 training, after learning/testing relationships between individuals in both dimensions, participants performed an additional test block (called “flexible test”) in which participants make inferences of relationships between pairs in the given social hierarchy dimension while the task-relevant dimensions were randomly interleaved across trials. Participants who successfully inferred the relationships in both social hierarchy dimensions and exceeded the predetermined accuracy threshold (participants who successfully distinguished the second- and third-rank individuals above chance and also reached > 85% overall accuracy in the flexible test) continued to participate in the day 3 training. This threshold was determined and used in order to exclude participants who may be using heuristics such as wins/losses regardless of the current task-relevant dimensions rather than building relational structure to infer the unlearned relationships. The 3 days of behavioral training created a relational structure of 16 faces in two social hierarchy dimensions. The inferred trajectories between individuals were determined by the learned relational structure. The details in behavioral training can be found in the protocol paper (Park et al., 2021b). Unbeknownst to participants we paired an individual only with a limited number of other individuals during the first two days of training. This inherently created two eight members groups (learning group). That is, an individual was compared only to the same group members but not to those in the other group. During day 3 training, right before the fMRI scanning, participants learned between-group relationships only from selected pairs for the first time. In particular, only certain faces called “hubs” were paired with members of their own group (on days 1 and 2) as well as members of the other group (on day 3). To infer the unlearned between group relationships, therefore, participants need to make transitive inferences through a unique hub face paired both group members. This allowed us to create a unique trajectory for a transitive inference during the fMRI experiment. Among 282 participants who initially signed up for the participation, 218 participants completed the first two days of training. Among 85 participants whose accuracy exceeded the predetermined threshold, 65 participants continued to the day 3 behavioral training.

Functional imaging acquisition

We acquired T2-weighted functional images using Siemens Skyra 3 Tesla scanner. We used gradient-echo-planar imaging (EPI) pulse sequence, with a multi-band acceleration factor of 2. We set the slice angle of 30° relative to the anterior-posterior commissure line to minimize the signal loss in the orbitofrontal cortex region (Weiskopf et al., 2006). The 38 slices were acquired with 3mm thickness with the following parameters: repetition time (TR) = 1200 ms, echo time (TE) = 24 ms, flip angle = 67°, field of view (FoV) = 192mm, voxel size = 3 × 3 x 3 mm^3. Contiguous

slices were acquired in interleaved order. We used a field map to correct for potential deformations with dual echo-time images covering the whole brain with the following parameters, TR = 630 ms, TE1 = 10 ms, TE2 = 12.46 ms, flip angle = 40°, FoV = 192mm, voxel size = 3 × 3 x 3 mm^3. A T1-weighted structural image was acquired using a magnetization-prepared rapid gradient echo sequence (MPRAGE) with the following parameters: TR = 1800 ms, TE = 2.96 ms, flip angle = 7°, FoV = 256mm, voxel size = 1 × 1 x 1 mm3.

Pre-processing

The preprocessing of functional imaging data was performed using SPM12 (Wellcome Trust Centre for Neuroimaging). Images were corrected for slice timing, realigned to the first volume, and realigned to correct for motion using a six-parameter rigid body transformation. Inhomogeneities created using the phase of nonEPI gradient echo images at 2 echo times were coregistered with structural maps. Images were then spatially normalized by warping subject-specific images to the reference brain in MNI (Montreal Neurological Institute) coordinates (2mm isotropic voxels).

Testing simultaneous context-invariant and context-dependent representations

We performed a whole-brain searchlight representational similarity analysis (RSA) (Kriegeskorte, 2008; Nili et al., 2014) to test whether the brain represents the 2D context-invariant and 1D orthogonal context-dependent representations simultaneously. Specifically, we estimated the mean activity patterns associated with each of the faces while they were presented at the time of F1 or F2. The mean activity patterns were separately estimated according to which social hierarchy dimension was relevant to the current goal. Based on these representations, we computed the representational dissimilarity matrix (RDM) per region of interest (ROI) across the whole brain. We quantified the representational dissimilarity using the Euclidean distance between the standardized mean activity patterns of representations estimated from two independent fMRI blocks. We subsequently predicted the RDM estimated from the patterns of neural activity in each of the ROIs using a general linear model (GLM) in which we included four idealized model RDMs (**Supplementary Fig.1** and **Eq.1**). The model RDMs included (1) pairwise Euclidean distances between faces in the context-invariant 2D social hierarchy (*2D*); (2) pairwise Euclidean distances between individuals in the context-dependent 1D orthogonal representations in which the distances representing the task-irrelevant dimension are dynamically collapsed relative to those of the task-relevant dimension (*1D*); (3) if faces were presented within the same or different task-relevant social hierarchy dimension (*Context*); (4) if the ranks of two faces were compared within the same group during behavioral training or not (*Group*).

$$RDM \sim \beta_0 + \beta_{2D} 2D + \beta_{1D} 1D + \beta_C Context + \beta_G Group$$

Eq.1

The Context and Group RDMs served as control RDMs of no interest. To identify brain areas containing the 2D context-invariant and 1D context-dependent orthogonal representations, we mapped back the β_{2D} and β_{1D}, the extent to which the brain RDM of the activity patterns of each

region of interest (ROI) was explained by the pairwise Euclidean distances in the model RDMs on the central voxel of each of the searchlight ROI across the whole brain to create continuous maps per subject. We smoothed the images using an 8-mm full-width at half maximum (FWHM) Gaussian kernel and performed the fisher-Z transform. We further performed one-sample t-tests for statistical inference in group-level analyses.

Warping in representations of the cognitive map

The effects of congruence on reaction times

Using the reaction times of each trial (RT_t) as an index of choice difficulty, we tested for effects of context congruence (C_t; 1 for congruent trials; 0 for neutral trials in which the ranks of two faces in the task-irrelevant dimension are the same; -1 for incongruent trials), while controlling for the effects of the objective Euclidean distances of the inferred trajectory (E_t), using multiple linear regression (**Eq.2**). The inferred trajectory of each inference trial was defined by a unique hub face enabling the transitive inferences of the relationship between F1 and F2 (the distance between the hub and F2 was fixed as 1 rank difference in the given dimension). In addition to C_t and E_t, the regression included the rank difference between F1 and the hub in the given task-relevant dimension (D_t), the Euclidean distance (e_t) and rank difference between F1 and F2 (d_t), the context switch (S_t, S_t was 1 when the current task-relevant context was different from that of the last trial, otherwise 0), and the subject-specific random effects (β_0). Using the one-sample t-test, we tested the effects of congruence (the mean regression coefficient β_C) on the RT.

$$RT_t \sim \beta_0 + \beta_C C_t + \beta_E E_t + \beta_D D_t + \beta_e e_t + \beta_d d_t + \beta_S S_t$$

Eq.2

Estimating levels of hypothetical warping from reaction times

The previous finding using the same data showed that the RTs of inferences decrease with increases in the Euclidean distances of inferred trajectories (Park et al., 2020). Given that, if the representation was not warped, the extent to which the Euclidean distances of the inferred trajectory between individuals explain the RTs of inferences (β_E) should not be modulated by vector angles of the inferred trajectory (θ). On the other hand, if representations are warped, the vector angles of the inferred trajectory would account for variance in the extent to which Euclidean distances of the inferred trajectory explain the RTs in a manner that depended on whether the angles correspond to congruent or incongruent pairs. We estimated the relative levels of warping across participants by approximating the Euclidean distances between individual faces on the subjective representations from the RTs of the inferences.

To infer the subjective level of warping from RTs of inferences, first, we estimated the trial-by-trial variability in the degree to which the Euclidean distance of inferred trajectories explains the RT (β_E^t) after accounting for the effects of other distance terms and the intercept, as **Eq.3** shows below:

$$\beta_E^t \approx \frac{(RT_t - \beta_0 - \beta_D D_t - \beta_d d_t)}{E_t}$$

Eq.3

where $t = \{1,2, \ \ldots \ T\}$ and T is the number of inference trials, and the effects of rank differences in the task-relevant dimension (*D* and *d*) on the RTs (β_D and β_d) and β_0 of each participant were estimated from a robust linear regression in which we inputted E_t and D_t to predict RTs. Next, we inferred the norm vector distances ($\widehat{\epsilon_t}$) over the subjective representation from β_E^t. Since the longer the distances, the faster the RTs (while $\widehat{\epsilon_t}$> 0 and β_E^t<0 in general), we approximated $\widehat{\epsilon_t}$ as follows (**Eq.4**):

$$\widehat{\epsilon_t} \approx e^{-\beta_E^t}$$

Eq.4

$\widehat{\epsilon_t}$ indicates the Euclidean distance of the unit vector of the inference trajectory of a certain vector angle (θ_t). That is, if the representation was not warped, $\widehat{\epsilon_t}$ would have consistent values across the trials regardless of the vector angle, θ_t. Consequently, the distribution of $\widehat{\epsilon_t}$ according to θ_t across trials should have a topology of the circle (**Supplementary Fig.2**). On the other hand, this distribution would have a topology of an ellipse that was elongated along the congruent axis (y=x) if the representation was warped. To test this, we examined the topology of the distribution of $\widehat{\epsilon_t}$ as a function of θ_t in polar coordinates. After excluding outlier trials of each participant (> mean $\hat{\epsilon}$ ± 3×standard deviation [SD]), if they exist, we converted the norm vector of every trial, ([$\widehat{\epsilon_t}$, θ_t]) from polar- into cartesian-coordinates as follows (**Eq.5**):

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \widehat{\epsilon_t}\ cos(\theta_t) \\ \widehat{\epsilon_t}\ sin(\theta_t) \end{bmatrix}$$

Eq.5

From the distribution of each participant, we measured the covariance between X and Y ($cov(X,Y)$) to examine the relationships between two variables composing the distribution. The covariance indicated the direction of the relationship between two variables. That is, a positive covariance indicated that the distribution is elongated along the congruent axis (y=x) while a negative covariance indicated that the distribution is elongated along the incongruent axis (y=-x). Thus, to infer the geometry of the subjective representation of each participant, we computed the covariance matrix (Σ) as below (**Eq.6**):

$$\Sigma = \begin{pmatrix} v_1 & v_2 \\ v_2 & v_3 \end{pmatrix} = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix}$$

Eq.6

Σ allowed us to capture the subjective representation from the distribution of the RTs as follows. First, $cov(X,Y)$= 0 if the representations were not warped while $cov(X,Y)$>0 if the representations were warped along the congruent axis (**Supplementary Fig.2**) (c.f., $cov(X,Y)$ <0 if the representations are warped along the incongruent axis, y=-x). Second, the ratio between $var(X)$ and $var(Y)$ indicates to what extent RTs were explained better by the rank differences in one dimension compared to those in the other dimension. That is, if $var(X) = var(Y)$, the participant

represents the competence and popularity dimensions similarly, and the distances between faces in different ranks are the same in both dimensions (otherwise the representation is elongated along the vertical or horizontal axis). By applying Σ of each participant to the positions of faces in the actual 4×4 social hierarchy (x_i^a, y_i^a), we estimated the new positions of individuals (x_i^w, y_i^w) and the Euclidean distances ($E_{i,j}^w$) between faces (*i* and *j*) in the subjective representations of the social hierarchy (***Eq.7***).

$$\begin{bmatrix} x_i^w \\ y_i^w \end{bmatrix} = \Sigma \begin{bmatrix} x_i^a \\ y_i^a \end{bmatrix}$$

$$E_{i,j}^w = \sqrt{\begin{bmatrix} x_i^w \\ y_i^w \end{bmatrix} [x_j^w \quad y_j^w]}$$

Eq.7

We estimated the level of warping as the ratio (ω_{behv}) between the length of the congruent unit vector, ($\|\overrightarrow{v^c}\|$) and the length of the incongruent unit vector ($\|\overrightarrow{v^i}\|$) (**Eq.8**).

$$\omega_{behv} = \frac{\|\overrightarrow{v^i}\|}{\|\overrightarrow{v^c}\|} = \sqrt{\begin{bmatrix} M \\ m \end{bmatrix} \Sigma [M \quad m]} \Big/ \sqrt{\begin{bmatrix} M \\ M \end{bmatrix} \Sigma [M \quad M]}$$

$$where\ M = 4, \max(rank)\,; m = 1, \min\,(rank)$$

Eq.8

That is, the participant whose representation was more warped would have smaller ω_{behv}. It is important to note that we assumed that the RTs of inferences have an inverse logarithmic relationship with the distances between faces for convenience's sake (**Eq.4**). Given that, ω_{behv} should not be interpreted as the absolute level of warping in the subjective representation. Instead, ω_{behv} was used for testing for the relative level of warping and ranking participants accordingly.

To test if the representation was warped significantly along the congruent? axis, ω_{behv} was compared to the baseline (ω_{base}). This was estimated from each participant by randomly shuffling the relationship between the $\widehat{\mathrm{E}_\mathrm{t}}$ and θ_t (1000 times per subject). The difference between the levels of warping ($mean[\omega_{base}] - \omega_{behv}$) were inputted into the one-sample t test.

To validate this method, we performed a simulation. To test if we successfully infer the relative rank of warping (ω) across participants from the RTs using the above-mentioned approach, we generated synthetic reaction time data ($\widetilde{RT_t}$) according to **Eq.9**:

$$\widetilde{RT_t} = \overline{\beta_0} + \overline{\beta_E} E_t^w + \overline{\beta_D} D_t + \overline{\beta_d} d_t$$

Eq.9

where $\overline{\beta_E}$, $\overline{\beta_D}$, $\overline{\beta_d}$ and the noise ($\overline{\beta_0}$) were sampled from normal distributions ($\mathbb{N}$) using the mean and standard error of those estimated from participants' data (Eq.2) as follows: $\overline{\beta_E} = \mathbb{N}[-0.05,0.01]$, $\overline{\beta_D} = \mathbb{N}[-0.09,0.01]$, $\overline{\beta_d} = \mathbb{N}[-0.04,0.01]$, and $\overline{\beta_0} = \mathbb{N}[1.82,0.03]$. The Euclidean distances between individuals over the warped representations (E^w) were estimated from the

space to which we applied different levels of warping ($\Sigma_{target} = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}$ while varying σ ranging from 0 to 0.5 (the greater σ, the greater the warping) according to *Eq.7*. We generated 1000 sets of synthetic RT data. From the synthetic data, we estimated the level of warping ($\widehat{\omega_{behv}}$) in the same way (**Eq.3-8**). We further tested for the monotonic increase of the $\widehat{\omega_{behv}}$ estimated from synthetic RTs as a function of the level of warping (ω_{behv}) estimated from the actual structure (Σ_{target}).

Neural encoding of inferred trajectories over warped representations of the cognitive map

We tested if the brain activity is better accounted for by the distances between individuals over the subjective representations (E^w) rather than the Euclidean distances over the actual 4×4 social hierarchy (E). For the univariate analysis images were smoothed using an 8-mm full-width at half maximum Gaussian kernel (Mikl et al., 2008). To test this, we modeled the brain activity in the predefined regions of interest (ROIs) using a GLM. To deal with the inherent high correlation between E^w and E, we inputted E and E^w-E as parametric regressors into the GLM using a 2s boxcar function. The 6 motion regressors and the button press onsets were also included as regressors of no interest using a stick function. This allowed us to test for the effect of E^w in a parsimonious manner while controlling for the effects of E. Statistical inference was made within *a priori* selected ROIs that combined the anatomically defined vmPFC and EC (see ROI definitions).

Testing warping in the geometry of neural representations of the social hierarchy

We performed multivariate RSA across the whole brain to directly test the warping from the context-invariant neural representations of the social hierarchy. Unlike the analysis methods using pairwise distances to test different representational geometries, the rank correlation between the brain and model RDMs cannot prove the warping because the pairwise Euclidean distances in a warped representation highly correlate with the pairwise Euclidean distances in the actual 4×4 social hierarchy. Given that, to compute the level of warping from the neural representation, we compared the pattern dissimilarity between representations of the congruent and incongruent pairs. Specifically, first, we computed the average pattern dissimilarity across the pairs that created the same vector ($dis(\vec{v})$) (**Supplementary Fig.3b**). Second, we computed the ratio between the average pattern dissimilarity of congruent vectors, $dis(\overrightarrow{v_k^c})$ and that of equidistant incongruent vectors, $dis(\overrightarrow{v_k^i})$. The $\overrightarrow{v_k^c}$ and $\overrightarrow{v_k^i}$ were symmetric. That is, $\cos \overrightarrow{v_k^c} = -\cos \overrightarrow{v_k^i}$ (vector angles) and $\left\|\overrightarrow{v_k^c}\right\| = \left\|\overrightarrow{v_k^i}\right\|$ (vector distances). Third, the ratio between mean pattern dissimilarities was measured per pair of the congruent and incongruent vectors (9 vector pairs, *k*={1,2, …, 9}; See **Supplementary Fig.3**). Last, we estimated the level of warping (ω_{neural}) from the mean ratio of pattern dissimilarities between these symmetric vectors (**Eq.10**) per searchlight ROI across whole brain. Therefore, consistent with ω_{behv}, smaller ω_{neural} indicated greater warping in the representations. Notably, ω_{neural} was only sensitive to the shear transformation (warping) of a context-invariant representation but indifferent to other affine transformations (**Supplementary Fig.3**).

$$\omega_{neural} = 1 - mean\left(\frac{dis\left(\overrightarrow{v_k^i}\right)}{dis\left(\overrightarrow{v_k^c}\right)}\right)$$

Eq.10

We created a continuous map of ω_{neural} in the whole brain per subject by mapping back its value on the central voxel of the corresponding ROI. Each image was smoothed using an 8-mm FWHM Gaussian kernel and inputted into the group-level analysis using the one-sample t-test.

We performed a control analysis while excluding the longest vector pair ($\left\|\overrightarrow{v_9^c}\right\|$ and $\left\|\overrightarrow{v_9^i}\right\|$) which was created by the representations of the 4 faces whose ranks were at the highest or at the lowest rank in either dimension. Notably, all representations of the remaining 12 faces served to create equidistant congruent and incongruent vectors. That is, the pattern dissimilarities ($dis\left(\overrightarrow{v_k^i}\right)$ and $dis\left(\overrightarrow{v_k^c}\right)$) were estimated from different pairs of the same representations. This allowed us to test for warping while excluding the possibility that this effect was driven by a few extreme representations. The statistical inference was made within an *a priori* ROI defined in the anatomically defined right HC (see ROI definitions).

We further tested for brain areas whose level of warping estimated from neural representations (ω_{neural}) correlates with those estimated from RT (ω_{behv}) across participants by inputting ω_{behv} (demeaned) as a covariate for the second-level analysis.

Intersubject representational similarity analysis

We investigated individual differences in neural representations of the social hierarchy structure reflecting differences in behaviorally estimated levels of warping (ω_{behv}) across participants using the intersubject representational similarity analysis (IS-RSA) (Chen et al., 2020; Cohen et al., 2017; Hasson et al., 2004). The IS-RSA allowed us to examine the relationship between the neural representations and individual differences in cognitive control (RT biases) without presuming any representational geometric form. To perform this, we first constructed a warping dissimilarity matrix using the difference between the level of warping (ω_{behv}) across participants. We then build the brain RDMs for IS-RSA: We computed the RDMs by comparing the neural activity patterns of faces in two different social hierarchy dimensions estimated from different trials across blocks using Euclidean distances. We computed intersubject neural dissimilarity matrix (size=n×n, where n=27 subjects) by comparing brain RDMs estimated from the same searchlight ROI across participants from 1-Pearson correlation coefficients. We computed the intersubject neural similarity per searchlight ROI across the whole brain using the Kendall's rank correlation between the intersubject neural similarity matrix and the behaviorally estimated warping similarity matrix (**Supplementary Fig.4**).

The statistical significance for IS-RSA was assessed against the baseline estimated from a permutation test. We generated 1000 similarity matrices while randomly shuffling ω_{behv} across participants. We then created a null distribution of 1000 correlations between the shuffled matrix

and the neural RDM of each searchlight ROI. From the normal distribution defined by the mean and standard deviation of a null distribution, we calculated p-values of each voxel across the whole brain. We corrected for multiple comparisons using the False Discovery Rate (FDR) threshold at q=0.05 across the whole brain (Benjamini & Hochberg, 2000).

In addition, to rule out the possibility that the effects of IS-RSA were driven by any single outlier subject, we performed the IS-RSA for n-1 subjects for n times (n=27) while excluding one subject at a time (leave-one-out analysis).

Reactivation of current task-relevant context at the time of inferences

We tested the hypothesis that the participants who effectively use the task context that was relevant to the current goals to select appropriate actions in incongruent trials were more likely to build and have a more accurate social hierarchy representation, while those who were less likely to resolve incongruence might tend to have warped representations and their inferences would be modulated by the context congruence.

To test for the effect of context utilization on different levels of warping (ω_{behv}) in the representations, we performed the following behavioral and fMRI analyses. First, we tested if the accuracy biases (the tendency to make more accurate inferences on congruent than incongruent trials) related to the effects of congruence on the RTs (β_C in **Eq.2**). To capture greater variance across individuals, we test this in behavioral data acquired from a larger cohort of participants (n=218) (the flexibility test collected on day 2 training).

Next, using fMRI, we examined if participants tend to have greater reinstatement of the current context information at the time of inferences on incongruent than congruent trials. To address this, we performed a series of RSA analyses as follows: (1) We identified brain areas containing task-specific representations by comparing activity patterns at the time of the context cue presentation (See ROI definitions; **Supplementary Fig.7a**). (2) We estimated the mean activity patterns of each searchlight ROI across the whole brain at the time of inferences according to the behavioral context (competence or popularity), congruence (congruence or incongruence), and button press (right and left) (8 activity patterns per ROI); (3) We compared the activity patterns associated with context information (acquired at the time of the context cue) to the activity associated with inferences (acquired at the time of decision making) (**Supplementary Fig.7c**). This created a 10×8 RDM per searchlight ROI; (4) We created a model RDM based on a geometrical model (**Supplementary Fig.7b**) that comprised of independent vectors representing the representations of two context dimensions and those from inferences on the congruent and incongruent relationships, respectively. Based on this model, we created a model RDM in which the dissimilarity between representations was estimated from 1-cosine angle between two vectors ($1 - S_C$). (5) We used a GLM to examine to what extent the model RDM explains variance of the neural RDM across the whole brain. In addition to the model RDM of interest, we included 3 other RDMs into the GLM to control for the potential effects of the different contexts, incongruence, and the different button presses (**Supplementary Fig.7c**); (6) We created continuous maps capturing the extent to which each model RDM accounted for the neural RDM per subject by mapping back them into a central voxel of each ROI. We smoothed the map of each subject using an 8-mm FWHM Gaussian kernel, and performed one-sample t tests for a group-level analysis; (7) To

identify brain areas whose activity patterns associated with context coding was preferentially reinstated during incongruent compared to congruent trials, statistical inference was performed with TFCE within the brain areas having distinctive context representations identified from (1) (See ROI definitions).

We further tested if participants who had greater fidelity of context reinstatement during inferences would have representations that were less warped along the context congruent axis. To estimate the fidelity of context reinstatement, we computed the cosine similarity (S_C) between the neural representations associated with the inferences on incongruent trials (I_d^-) in different context of social hierarchy dimensions ($d \in$ {competence, popularity}) and those of the task-relevant context representation (R), with respect to the opposing context representation that was irrelevant to the current goals (I). We computed the fidelity of two contexts of social hierarchy dimensions separately and combine them (**Supplementary Fig.7d** and **Eq.11**).

$$\sum_{d=1}^{D} \frac{S_C(I_d^-, R)}{S_C(I_d^-, R) + S_C(I_d^-, R)}$$

Eq. 11

We estimated this fidelity by comparing vectors of neural activity patterns of an *a priori* ROI defined anatomically in the HC based on the results showing greater reinstatement of context information during inferences on incongruent compared to congruent trials. Last, we examine the relationship between this fidelity of context representation at the time of inferences and ω_{Behv} across participants using Spearman's rank correlation. We performed the same analysis using the neural activity patterns associated with inferences on congruent trials (I_d^+) as well to test if the effect was specific to inferences on incongruent trials.

Psychophysiological interactions

Using a psychophysiological interactions (PPI) analysis (Friston et al., 1997), we examined the functional connectivity between brain regions whose activity increased in response to incongruence during inferences, and brain regions showing context coding at the time of inferences, and test if the functional connectivity increased selectively on incongruent trials compared to congruent trials. First, we performed another GLM to identify brain areas showing greater activity at the time of inferences for incongruent compared to congruent trials. In addition to the level of incongruence, we included the Euclidean distance of inferred trajectories over the subjective (warped) representation (**Eq.7**) into the GLM as a regressor of no interest to control for task difficulty. Based on this analysis, we defined the seed region at the peak voxel of each participant within a 10 mm radius sphere centered on the group peak for the mid-anterior cingulate cortex ([x,y,z]=[2,12,36]). From the seed, first, we computed the physiological factor for the PPI analysis by extracting the deconvoluted time series from all voxels within a 4 mm radius sphere centered on the participant-specific peak voxel. Second, the psychological factor for the PPI model was defined at the inference onsets of incongruent and congruent trials in which we assigned a weight of 1 and -1, respectively. All other conditions were coded with a weight of 0. Third, in addition to those two regressors (physiological and psychological factors), we included

the psychophysiological interaction (PPI; [incongruent inference trials × ACC] > [congruent inference trials × ACC]) and motion regressors into the GLM for PPI analysis. Last, we estimated the individual contrasts for the PPI regressor and inputted them into the group-level analysis using one-sample t-tests. Statistical inference was performed with TFCE within *a priori* ROIs. The ROIs were defined anatomically in the HC based on the results of the independent analysis revealing the brain areas having distinctive context representations.

Statistical inferences

We perform group-level inference on whole-brain analyses unless we specified the hypothesis-driven *a priori* ROIs. We reported the results corrected for family-wise error (FWE) for multiple comparisons across the whole brain using threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009) with 1000 iterations (pTFCE < 0.05). For the ROI analysis, the multiple comparisons were corrected within each of the ROIs using TFCE with 1000 iterations (pTFCE < 0.05). All statistical parametric maps presented in the manuscript are unmasked at the threshold, pTFCE < 0.05.

ROI definitions

In the current study, we used 3 types of *a priori* ROIs: (1) the searchlight ROIs across the whole brain for multivariate analyses, (2) the anatomical ROIs defined independently from the current task based on a probabilistic map acquired from independent studies, and (3) the functional ROIs based on the results of independent time points of the current task. First, we defined the searchlight ROIs as a sphere including 100 voxels around each of the central voxels across the whole brain. Second, the anatomical ROIs were defined bilaterally in the HC combining only its anterior and middle portions (Yushkevich et al., 2015), EC (Amunts et al., 2005; Zilles & Amunts, 2010), amygdala (Tzourio-Mazoyer et al., 2002), and vmPFC/mOFC (Neubert et al., 2015). Third, the functional ROIs were defined from an independent analysis to identify the brain areas having distinctive context representations. To define these functional ROIs, we performed a whole-brain RSA in which we compared the mean activity patterns at the time of the context cue presentation acquired from different blocks. The cluster forming threshold for the functional ROIs was pTFCE<0.05 corrected over the whole brain (**Supplementary Fig.7a**).

Neural network model

The recurrent neural network (RNN) model was trained on the same task used in the fMRI experiment. On each trial, a cue indicating the relevant social hierarchy dimension (context cue) and a pair of images of faces (F1 and F2) were presented sequentially to the network, which was trained to predict which of the two faces corresponded to the individual that ranked higher on the task-relevant dimension (**Fig.1e**). The training and test sets for the model consisted of the same sets of trial samples used to train and test the participants in the human fMRI experiment: the model was trained only on limited pairs whose ranks differed by one in the given social hierarchy dimension, but was tested on pairs whose ranks differed by one or more than one - thus requiring

transitive inference. Consistent with the human participants, the RNN model did not receive any explicit indication of these groupings, nor of the fact that the faces were arranged in a latent 4x4 grid structure along the two social hierarchy dimensions.

Model architecture

The RNN model was built from standard building blocks including a convolutional neural network (CNN) for processing the images of faces and a Long Short-Term Memory (LSTM) network for processing the sequence of stimuli on each trial (Cue, F1, and F2). A recurrent network was chosen in order to better replicate the experience of the participants, who received the stimuli sequentially, and to capture the emergent dynamics of representations unfolding over the course of each trial. The Cue was represented as a one-hot vector (x_c in **Eq.12** and **Fig.1e**) indicating whether the relevant context/dimension was popularity or competence, and was embedded in a 32-dimensional vector (e_c) using a linear layer (W_c). We used the same face images with the fMRI experiment to train the neural network model. The two faces (F1 and F2) were given to the network as 64x64 grayscale images (x_1 and x_2), which were embedded in two 32-dimensional vectors (e_1 and e_2 in **Eq.12** and **Fig.1e**) using the same CNN:

$$e_c = W_c x_c$$
$$e_1 = CNN(x_1)$$
$$e_2 = CNN(x_2)$$

Eq. 12

The LSTM then processed these three embeddings in sequence ($e_c \rightarrow e_1 \rightarrow e_2$):

$$h^{(t)} = LSTM(e^{(t)}, h^{(t-1)})$$

Eq. 13

where $h^{(t)}$ is the hidden state of the LSTM at time step (the sequential processing steps of the RNN) t and $e^{(t)}$ is the embedding (e_c, e_1, or e_2) at time step t **(Eq.13)**. The model produced a prediction ($\hat{y}$) for whether the first or second face ranked higher along the relevant dimension using a linear output layer (W_o) applied to the final hidden state **(Eq.14)**:

$$\hat{y} = W_o h^{(3)}$$

Eq. 14

Training procedure

The RNN was trained with standard procedures, including a cross-entropy loss function and the backpropagation algorithm. The model was trained in a supervised manner on batches of 32 trials

randomly sampled from the entire training set. Unlike the human participants, training the neural network on one dimension at a time in a blocked manner would have caused the network to forget during the second block what had been learned in the first (well-known problem of catastrophic forgetting (McClelland et al., 1995; McCloskey & Cohen, 1989). The catastrophic forgetting phenomenon represents a relatively well-studied discrepancy between the learning behavior of neural networks as compared with humans. To avoid catastrophic forgetting, the training set for the RNN was randomly sampled from batches combining all pairs presented during learning blocks thoughout 3 days of behavioral training of human participants. Therefore, the task relevant dimensions were randomly intermixed across trials. While this is not central to the issues under investigation in the current study, other studies (Flesch et al., 2018; Flesch, Nagy, et al., 2022) including ours (Russin et al., 2022) are exploring the issue of catastrophic forgetting.

Implementation details

All neural network simulations were implemented using PyTorch (Paszke et al., 2019). The model was optimized with Adam (Kingma & Ba, 2014) with a learning rate of 0.001 and a batch size of 32 for 1000 gradient steps. For each simulation, 20 runs were performed with different random initializations. All parameters were initialized to be small (i.e., in the "rich" regime) using the PyTorch default option. The CNN included two convolutional layers with kernel sizes (3, 3), strides (2, 2) and number of channels (4, 8). Max pooling followed each convolutional layer with kernel sizes (2, 2) and strides (2, 2). The CNN also included a final linear layer to map the output of the last pooling operation to a single vector with 32 dimensions. The hidden layer of the LSTM was 128 dimensions.

Representational geometry in the RNN

To test our first two hypotheses about the nature of the representations learned by the model, we conducted an RSA analogous to the one used with the fMRI data. Averaged representations for each face in each context were collected from the activity patterns of the RNN hidden layer (*h*), and an RDM was computed using pairwise Euclidean distances. This created a 32×32 RDM, by including the patterns associated with 16 faces in two social hierarchy dimensions, considering that the test trials in the RNN were randomly sampled from the pairs presented in either test blocks in the behavioral training or in the fMRI. The representational geometries of the hidden layers of neural network model was tested using the same GLM with the fMRI experiment (**Eq.1**) while excluding the group RDM since the pairs were intermixed during training across learning groups in RNN. The GLM was performed after every 50 training steps, allowing us to visualize the emergence of these phenomena throughout the course of training. We further performed one-sample t-tests for group-level analysis across 20 runs.

Testing warping in the RNN representations

To test the hypothesis that the map-like representations learned by the model were warped along the congruent diagonal, we measured the ratio (**Eq.15**; analogous to **Eq.8** in behavioral data of

human participants) of average Euclidean distances between RNN representations of incongruent pairs of faces ($\|\overrightarrow{v^i}\|$) to the average distances between representations of congruent pairs ($\|\overrightarrow{v^c}\|$). We also conducted a two-sample t-test measuring whether the average distances between congruent pairs of faces were significantly larger than the average distances between incongruent pairs of faces. This analysis was also conducted every 50 training steps.

$$\omega_{RNN} = \frac{\|\overrightarrow{v^i}\|}{\|\overrightarrow{v^c}\|}$$

Eq. 15

Ablation simulation: testing the effects of context information on representational geometry

To better understand the warping we observed in both the human brain and the neural network model, we performed a simulated "ablation" on the model where we degraded the incoming information from the context cue indicating the current task-relevant social hierarchy dimension (e_c in **Fig.1e** and **Eq.16**). This was achieved by multiplying the cue's embedding (e_c) by a weight (α^{-1}) ranging from 0 to 1.

$$e_c' = \frac{e_c}{\alpha} \; where \; \alpha \in [1,10]$$

Eq. 16

We subsequently measured the warping in the model's representations after learning using the same method described above (**Eq.15**). In addition, we performed a linear regression analysis (**Eq.17**) testing whether there was a significant effect of the ablation weight (α^{-1}) on ω_{RNN}.

$$\omega_{RNN} \sim \beta_0 + \beta_a \alpha^{-1}$$

Eq. 17

Relationship between congruence, warping, and accuracy

To test the hypothesis that the model learned to perform well at congruent trials before it learned to perform well at the incongruent trials, we conducted a logistic regression after every 50 training steps to test whether congruence significantly predicted the task performance accuracy.

Relationship between congruence, warping, and accuracy

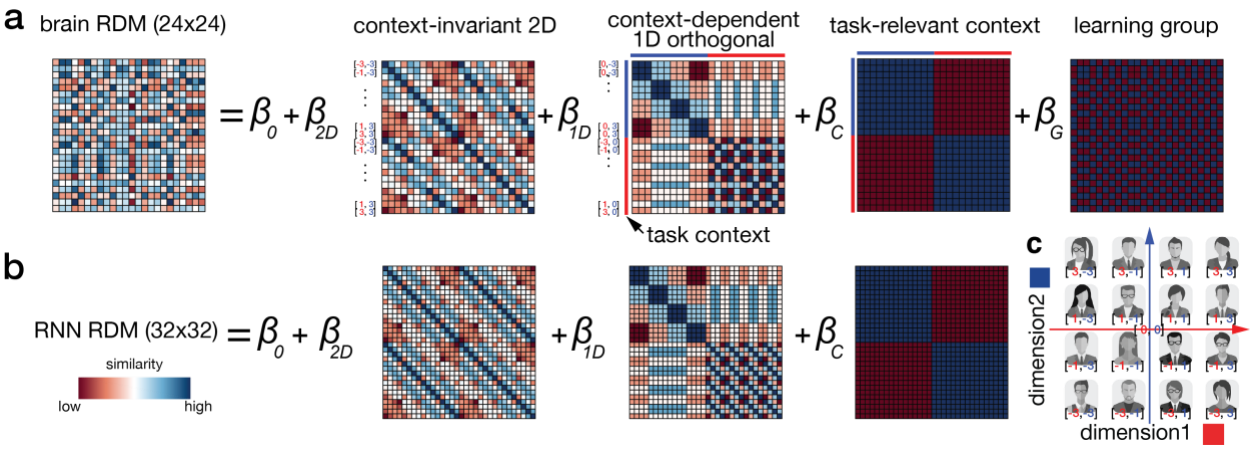
To test the hypothesis that the model learned to perform well at congruent trials before it learned to perform well at the incongruent trials, we conducted a logistic regression after every 50 training steps to test whether congruence significantly predicted the task performance accuracy.

We also tested whether the difference between model's performance on congruent ($accuracy^+$) and incongruent trials ($accuracy^-$) could be predicted by the degree of warping in the model's representations. To do this, we performed a linear regression predicting the accuracy ratio at each time step from (1) levels of warping, (2) a linear time variable, and (3) their interaction (**Eq.18**). Here, we included the linear time predictor (T=[0,20], every 50 training steps) in addition

to the warping predictor, as well as their interaction to capture the possibility that the effects of warping could vary over the course of training.

$$\frac{accuracy^{-}}{accuracy^{+}} \sim \beta_0 + \beta_T T + \beta_w \omega_{RNN} + \beta_x (T \times \omega_{RNN})$$

Eq. 18

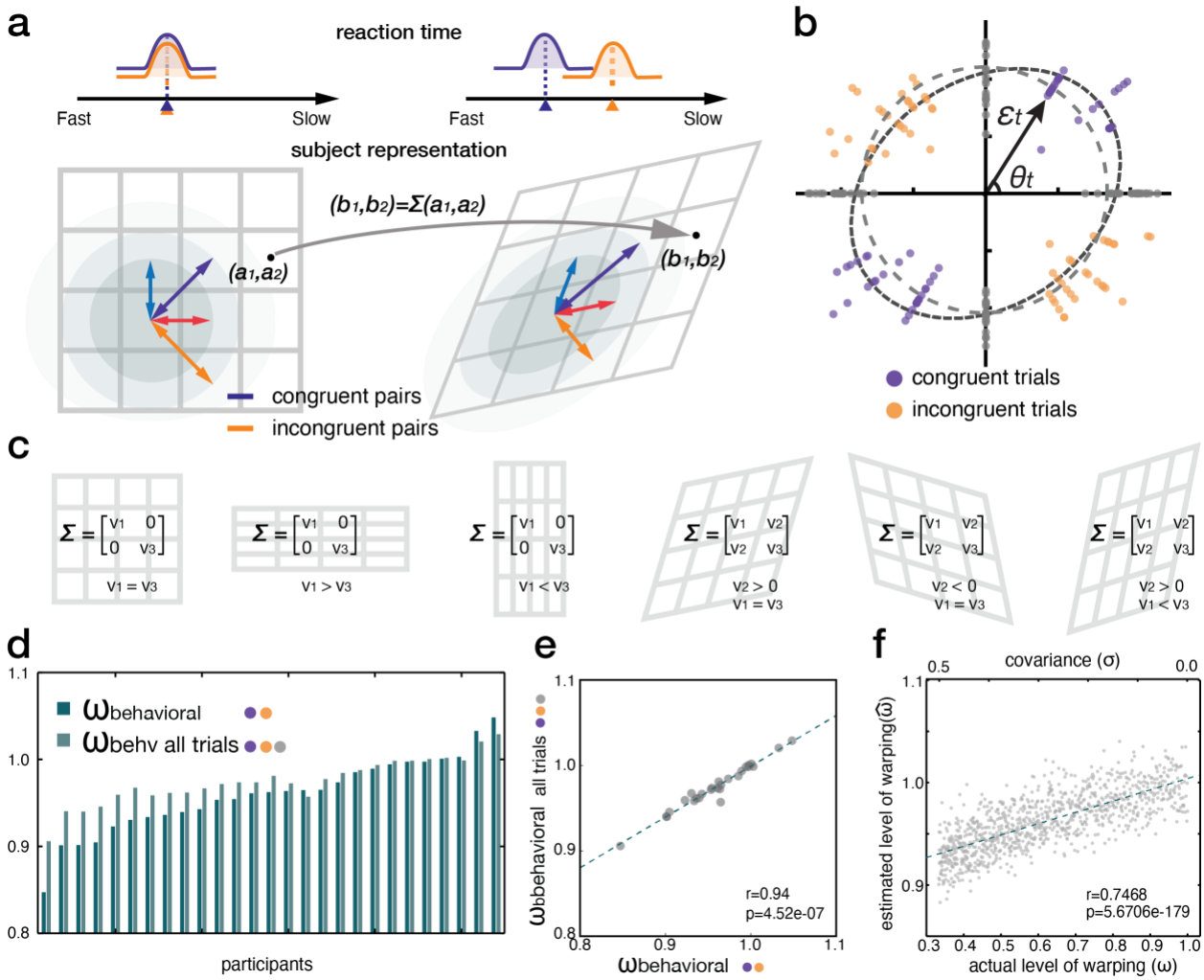


Supplementary Figure 1. The multivariate representational similarity analysis (RSA) using a general linear model (GLM) to examine the representational geometry in the brain and RNN

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 for approximating 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 whether RTs varied as a function of the angles of the 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. Otherwise, the circular distribution should be elongated along the context-invariant axis. From Σ of each participant, we computed the level of warping in

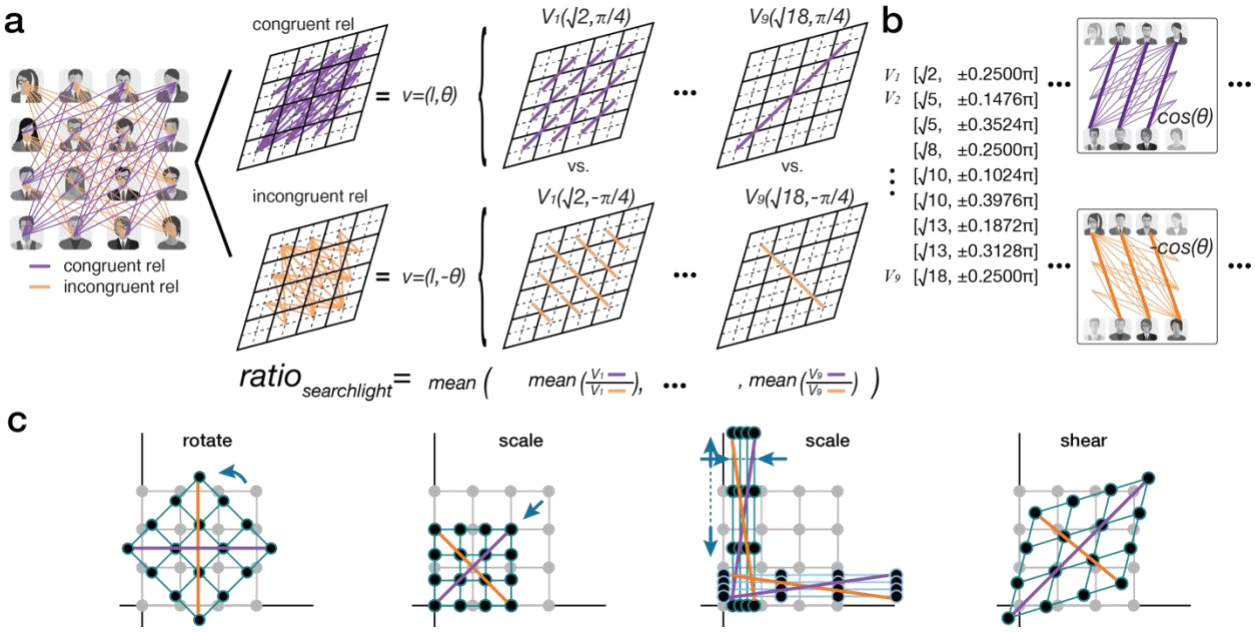
the shared representation ($\omega_{\text{behavioral}}$) as the ratio of the distances of incongruent pairs to those of congruent pairs.

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

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

f. Simulation results to validate the current methods. The level of warping estimated from RTs ($\omega_{\text{behavioral}}$) 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 $\omega_{\text{behavioral}}$ using the same method.

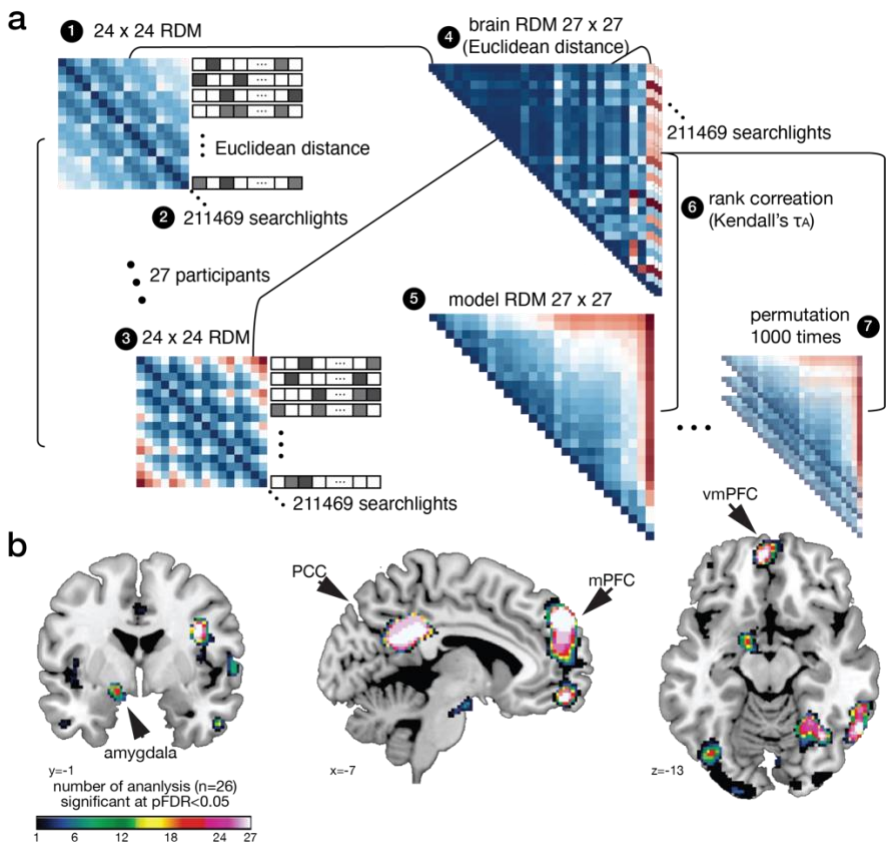


Supplementary Figure 3. The procedure to measure the level of warping (ω_{neural}) in the shared representation from activity patterns in the brain

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 us to test whether warping in the representation was not driven by some representations having extreme ranks.

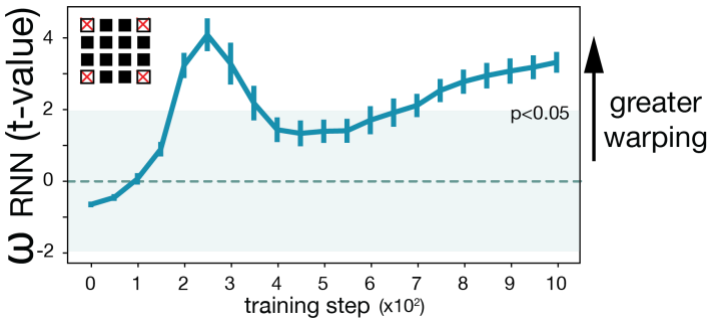
c. Measuring ratio between lengths of two diagonals allows testing of warping, defined by a shear transformation rather than other affine transformations (translation, rotation, or scaling in either dimension).



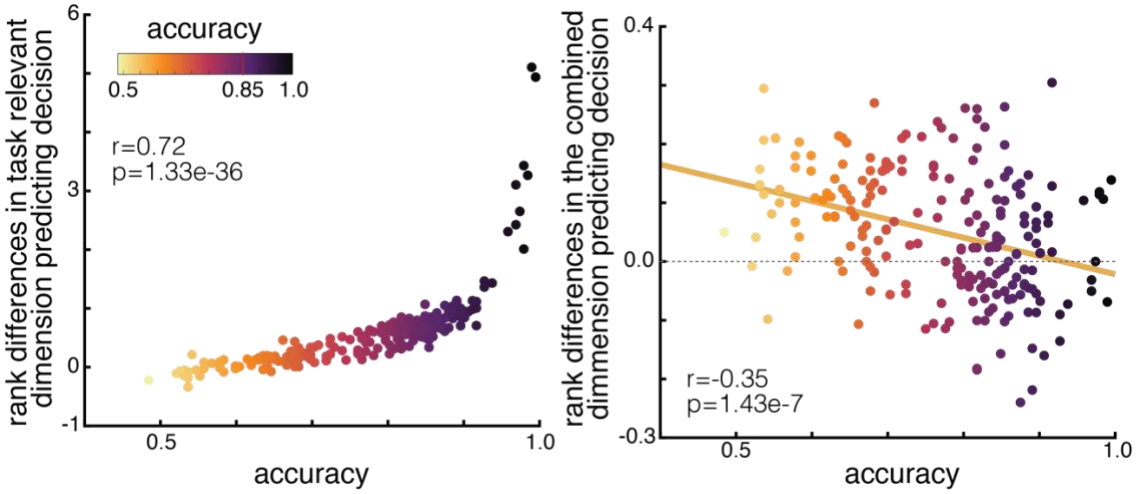
Supplementary Figure 4. Intersubject representational similarity analysis (IS-RSA)

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 each of the two task contexts using Euclidean distances; (2) An RDM was computed from each of 211,469 searchlight ROIs across the whole brain; (3) This procedure was repeated across participants (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 leave-one-subject-out IS-RSA included data from 26 participants while excluding one participant at a time. We repeated this 27 times. The color map indicates the number of analyses showing significant intersubject similarity across 26 participants at the threshold $p_{FDR}<0.05$

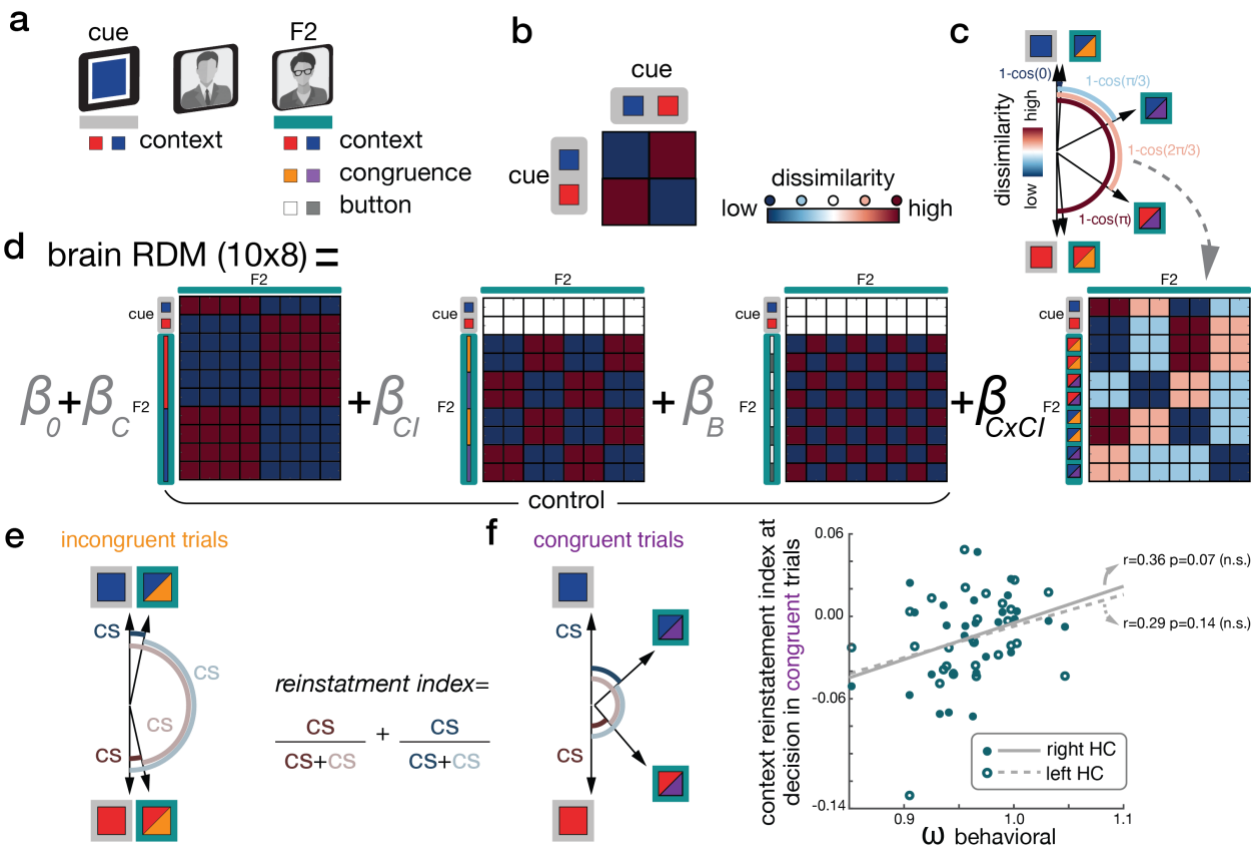


Supplementary Figure 5. Time course changes of level of warping in RNN representations while excluding the representations of faces who had the highest or lowest rank in either social hierarchy dimension. This result shows that the warping was not driven by the representations of certain faces but observed systematically across all face representations.



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

In the current task, participants should use rank differences in the task-relevant dimension (context-dependent 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 (context-independent 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 greater task context reinstatement at the time of inferences on the incongruent compared to congruent trials

a. For an additional RSA, we estimated the mean activity patterns of each task context (red: competence; blue: popularity) at the time of 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 regions containing task context-specific representations were identified by comparing the mean activity patterns associated with the two task contexts at the time of context cue presentation.

c. A graphical model for creating a model representational dissimilarity matrix (RDM) to test the hypothesis that context reinstatement at the time of inferences is greater on the incongruent trials compared to congruent ones. Each arrow indicates the vector of representation of two task contexts (gray outline) and that of F2 grouped by context congruence in each of two task-relevant contexts (green outline). The model RDM reflects $1-cos(angle)$ between two arrows. 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-interest 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 a 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 ($\omega_{behavioral}$) (similar to **Fig.4h**. using activity patterns associated with inferences on congruent relationships)

Brain areas	Laterality	cluster size (k)	Peak activity (MNI)				
			x	y	z	T	P_{TFCE}
2-D context-invariant representation							
Ventromedial prefrontal cortex / Medial orbitofrontal cortex (vmPFC/ mOFC)	Bilateral (B)	844	14	42	-22	5.83	<0.01
Hippocampus (HC)	Right (R)	832	28	-2	-18	4.67	0.01
Subgenual area	R	832	12	8	-20	4.15	0.02
lateral OFC	Left (L)	1844	-24	26	-16	4.55	0.01
lateral OFC	R	514	28	22	-22	3.96	0.02
Entorhinal cortex (EC)	R		20	0	-36	4.02	0.02
1-D orthogonal context-dependent representation							
Dorsomedial prefrontal cortex (dmPFC)	B	16680	-4	54	24	5.99	0.01
Dorsomedial frontal cortex/ Pre-supplementary motor area (dmFC/pre-SMA)	B		4	14	48	5.59	<0.01
Lateral PFC	L		-42	24	20	13.48	<0.01
Lateral PFC	R		42	26	18	9.03	<0.01
Precuneus/ posterior cingulate cortex (PCC)	B	15321	-4	-64	28	13.6	<0.01
Inferior parietal lobule/ Temporoparietal junction (IPL/ TPJ)	R		46	-64	22	5.2	<0.01
IPL/ TPJ/ Posterior superior temporal sulcus (STSp)	L		-48	-72	18	6.72	<0.01
Fusiform gyrus	L		-40	-60	-20	6.44	<0.01
Fusiform gyrus	R		30	-84	-10	5.54	<0.01

Table S1. Shared context-invariant and orthogonal context-dependent representations using whole-brain searchlight representational similarity analysis (RSA)

The upper table, associated with **Fig. 2a**, shows brain areas where the dissimilarity between activity patterns for faces significantly correlates with their pairwise Euclidean distances computed from the idealized context-invariant representation (i.e. the true 2D social hierarchy).

The lower table, associated with **Fig. 2b**, shows brain areas where the dissimilarity between activity patterns for faces significantly correlates with their pairwise Euclidean distances computed from the idealized 1D orthogonal context-dependent representation.

		1D orthogonal representations			
		All (pTFCE<0.05)	dmFC/ pre-SMA	lPFC	Precuneus/ PCC
2D map-like representations	All (pTFCE<0.05)	0.277 (0.171)	0.238 (0.243)	0.312 (0.121)	-0.247 (0.225)
	vmPFC/mOFC	-0.050 (0.810)	-0.202 (0.322)	0.078 (0.704)	-0.182 (0.372)
	EC	0.127 (0.536)	0.095 (0.644)	0.119 (0.564)	0.143 (0.485)
	HC	0.022 (0.915)	0.283 (0.161)	-0.045 (0.827)	0.228 (0.262)

Table S2. The relationship between shared context-invariant and orthogonal context-dependent representations in the brain across individuals

The correlation coefficient (r) and associated P-values (in parentheses) between mean brain activity (β) of brain regions that represent 2D map-like and 1D orthogonal representations across participants. The β values were estimated either from all brain areas that significantly represent 2D or 1D representations at a cluster-corrected threshold, $P_{TFCE}<0.05$, or from one of the selected clusters. The lack of a significant negative correlation (P>0.05) suggests that the brain represents both types of representations, rather than some individuals only showing one or the other representation.

Brain areas	Laterality	cluster size (k)	Peak activity (MNI)				
			x	y	z	T	P_{TFCE}
Group level							
HC	R	105	28	-12	-22	4.5	0.05
Supramarginal gyrus	R	259	52	-30	42	5.51	0.02
Inferior temporal gyrus	R	149	64	-12	-30	4.79	0.05
Fusiform/ Parahippocampal gyrus	R	238	44	-38	-14	4.6	0.04
Dorsolateral prefrontal cortex	R	104	42	34	24	4.56	0.05
Individual differences							
Dorsomedial prefrontal cortex (dmPFC)	R	357	22	48	16	9.9	<0.01
Amygdala	R	4962	28	0	-12	6.67	0.01
Amygdala	L		-20	-8	-16	3.3	0.03
Lateral OFC	R		34	32	-18	4.68	0.02
Fusiform gyrus	L		-36	-32	-18	4.24	0.02
Ventromedial/ Medial prefrontal cortex (vmPFC/mPFC)	B		-5	34	2	4.52	0.01
Fusiform gyrus	R	3027	28	-54	-6	5.88	0.01

Table S3. Warped map-like representations using whole-brain searchlight RSA

The upper table, associated with **Fig. 3d**, shows brain areas showing significant warping in their map-like representations.

The lower table, associated with **Fig. 3e**, shows brain areas with map-like representations whose level of warping correlates with individual differences in behaviorally estimated warping from reaction times (RTs) (ratio between diagonals on congruent and incongruent relationships between individuals, where their distances were approximated from the relationship between Euclidean distance and RTs)

Brain areas	Laterality	cluster size (k)	Peak activity (MNI)			
			x	y	z	T
Posterior cingulate cortex (PCC)	B	1728	-8	-48	28	7.28
Fusiform gyrus	R	1367	24	-50	-6	7.05
Fusiform gyrus	L	103	-20	-54	-10	4.3
Temporoparietal junction (TPJ)	L	900	-34	-74	16	6.72
TPJ	R	606	40	-64	18	6.48
Medial/ Dorsomedial prefrontal cortex (mPFC/dmPFC)	B	940	-8	56	38	6.28
Parahipocampal gyrus	R	117	50	-18	-22	5.82
Amygdala	L	178	-16	-2	-12	5.73
Inferior frontal gyrus	R	336	38	0	26	5.56
Precentral gyrus	R	453	36	-18	40	5.34
Dorsolateral frontal cortex (dlFC)	R	172	38	24	40	5.01
Medial orbitofrontal cortex (mOFC)	B	248	-4	56	-14	4.87
Inferior frontal gyrus/ Frontal inferior operculum	L	370	-46	14	16	4.73
Inferior temporal gyrus	R	237	54	-46	-26	4.39

Table S4. Intersubject RSA (ISRSA)

Associated with **Fig. 3f**, the brain areas where the similarity between representational geometry is associated with individual differences in behaviorally estimated warping from reaction times across participants. T-values compared to baseline are computed from 1000 permutations with random shuffling at the threshold $P_{FDR}<0.05$.

Training step	T	P
0	0.00E+00	1.00E+00
50	1.21E+00	2.25E-01
100 ⁎	3.53E+00	4.10E-04
150 ⁎	1.59E+01	4.12E-57
200 ⁎	3.05E+01	1.94E-204
250 ⁎	3.07E+01	2.32E-206
300 ⁎	2.89E+01	4.19E-183
350 ⁎	2.34E+01	7.18E-121
400 ⁎	1.46E+01	3.05E-48
450 ⁎	8.16E+00	3.30E-16
500 ⁎	4.32E+00	1.56E-05
550 ⁎	3.39E+00	6.97E-04
600	-1.43E+00	1.54E-01
650	-1.54E+00	1.25E-01
700	3.46E-01	7.29E-01
750	2.91E+00	3.57E-03
800	3.06E+00	2.18E-03
850	3.09E+00	1.97E-03
900	-8.72E-01	3.83E-01
950	9.26E-01	3.54E-01
1000	1.93E-01	8.47E-01

Table S5. Timecourse of the effect of context congruency on RNN inference accuracy.

Associated with Fig.4b, logistic regression analyses show that the RNN was more likely to make correct inferences on congruent trials than on incongruent trials, especially in the early training phase (⁎ indicates the training step where P<0.001). As training progressed, accuracies on both trial types improved and the effect of context congruency on accuracy decreased.

Brain areas	Laterality	cluster size (k)	Peak activity (MNI)				
			x	y	z	T	P_{SVC} with TFCE
SMA	B	8778	-6	12	44	6.59	<0.01
PCC	B		-4	-28	28	6.35	<0.01
Anterior insula	L		-30	18	-8	5.48	<0.01
Lateral OFC	L		-42	30	0	5.3	0.02
Inferior temporal gyrus	L		-48	-6	-32	3.7	0.02
Inferior frontal gyrus	L		-58	6	14	3.42	0.02
HC	L		-26	-8	-22	3.25	0.02
Insula/ Lateral OFC	R	5021	36	22	-2	6.44	<0.01
Inferior frontal gyrus	R		58	4	18	5.73	<0.01
Lateral PFC	R		32	32	16	4.89	<0.01
vmPFC	B	409	-12	48	-4	3.19	0.03

Table S6. Reinstatement of current task-relevant context during inferences on incongruent trials.

We examined brain areas where the activity patterns elicited by the task-relevant context cue were reinstated to a greater extent during inferences on incongruent trials than congruent trials. This effect was tested within an independently defined region of interest (ROI) that combined all brain areas showing context-specific activity patterns at the time of context cue presentation, using small volume correction (SVC). Multiple comparisons were corrected within the ROI using TFCE.

Brain areas	Laterality	Peak coordinate (MNI)			cluster size (k)	T	P(TFCE)
		x	y	z			
PCC	r	14	-32	28	494	4.47	0.04
HC	r	30	-28	-14	97	2.93	0.01 ⁎
HC	l	-28	-24	12	70	2.78	0.02 ⁎

Table S7. Psychophysiological interaction (PPI) analysis

Associated with Fig.3g, the brain areas showing increased functional connectivity to the ACC seed ROI when making inferences on incongruent pairs compared to congruent pairs. The ACC seed ROI was a 4 mm radius sphere centered on the peak activity of each individual participant at the time of inferences on incongruent trials, within a 10 mm radius sphere centered on the group peak for the mid-anterior cingulate cortex at [x,y,z]=[2,12,36]. The HC effects were corrected within an anatomically defined ROI (denoted by ⁎) using TFCE.