UPDATE ARTICLE

Social knowledge about others is anchored to self-knowledge in the hippocampal formation

Marta Rodríguez Aramendía, Mariachiara Esposito, Raphael Kaplan*

Department of Basic Psychology, Clinical Psychology, and Psychobiology, Universitat Jaume I, Castelló de la Plana, Spain

* kaplan@uji.es

Abstract

Mounting evidence suggests the human hippocampal formation (HF) maps how different people's attributes relate to each other. Yet, it's unclear if hippocampal map-like knowledge representations of other people are shaped by self-knowledge. Here, we test if a prominent heuristic involving an implicit reliance on self-knowledge when rating others, egocentric anchoring-and-adjustment, is present in the HF when relational information about different social entities is retrieved. Participants first provided likelihood ratings of partaking in everyday activities for themselves, fictitious individuals, and familiar social groups. During a neuroimaging task that doesn't require using self-knowledge, participants then learned a stranger's preference for an activity relative to one of the fictitious individuals and inferred how the stranger's preference related to the groups' preferences. Isolating the neural representation of egocentric anchoring when retrieving relational social knowledge, the HF and dorsomedial prefrontal cortex (dmPFC) represented group entities' preferences relative to the self. Furthermore, the HF selectively represented group identity over other learned entities, confirming the HF was primarily engaged by social comparisons in the more ample map-like reference frame. Taken together, these results imply that self-knowledge implicitly influences how the HF learns about others.

The Editors encourage authors to publish research updates to this article type. Please follow the link in the citation below to view any related articles.

OPEN ACCESS

Citation: Rodríguez Aramendía M, Esposito M, Kaplan R (2025) Social knowledge about others is anchored to self-knowledge in the hippocampal formation. PLoS Biol 23(4): e3003050. https://doi.org/10.1371/journal.pbio.3003050

Academic Editor: Matthew F. S. Rushworth, Oxford University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: March 5, 2024

Accepted: February 3, 2025

Published: April 2, 2025

Data availability statement: All relevant data files and code are available on Figshare (https://doi.org/10.6084/m9.figshare.28295732.v2).

Introduction

Learning via cognitive maps permits an organism to relate their current state to an internal model of the world [1]. Cognitive maps have inspired a better understanding of how neurons in the hippocampal formation (HF), including entorhinal cortex (EC), transform egocentric spatial cues into purely environment-centered allocentric coordinates supporting spatial navigation [2]. Growing evidence suggests the HF assimilates abstract knowledge similar to how it transforms spatial cues [3–6], where social learning tasks have been particularly successful in isolating this new role [7–11; see 12 for review]. In particular, Kaplan and Friston [8] investigated social learning beyond well-studied self versus other comparisons by testing how a stranger's preference was learned relative to one individual and then compared to other individuals. The transformation from learning preferences relative to one point of reference and relating that preference to others highlighted a role for the HF in social inferences in

Funding: This research is supported by grants awarded to RK from the Valencian Community's Program for the Support of Talented Researchers (CIDEGENT/2021/027), Universitat Jaume I Research Advancement Plan (UJI-B2022-45), and Spanish Science, Innovation, and University Ministry (PID2021-122338NA-I00). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations : AD, absolute distance; dmPFC, dorsomedial prefrontal cortex; EC, entorhinal cortex; fMRI, functional magnetic resonance imaging; GLM, general linear model; HF, hippocampal formation; LSA, Least Squares-All; nRDM, neural representation dissimilarity matrix; RC, rating consistency; RD, rating discrepancy; RDMs, representational dissimilarity matrices; RSA, representational similarity analysis; RSc, retrosplenial cortex; RT, response time; 2AFC, two-alternative forced choice task.

an absolute reference frame (i.e., flexibly comparing everyone in one's social environment). Although comparisons were made on a 1D number line of preferences, the authors viewed an absolute reference frame as analogous to map-like allocentric reference frames. This analogy is inspired by the similarity between the paradigm's nonlinear relative to absolute reference frame transformations of preferences and nonlinear egocentric-allocentric spatial coordinate transformations made during spatial navigation [13,14]. The extension of allocentric/absolute coding principles to social learning raises the possibility that spatial and social perspective-taking could rely on the same neural computations in the HF and beyond [15]. However, whether this form of hippocampal social learning is supported by an isometric map-like representation, or an integrated model of relational knowledge via more imprecise subjective mechanisms is unknown [16–18]. On one hand, firing fields of spatially-modulated hippocampal neurons resemble cartesian coordinates on a map [5,19], yet the self often serves as a reference point when making memory-guided inferences [20,21]. Moreover, previous work involving social reference frame transformations used the self as one of the explicit reference points in all conditions [8]. Consequently, it is unclear if subjective biases might be present in HF map-like representations of abstract knowledge about others.

One well-characterized subjective bias that could occur during map-like learning is anchoring-and-adjustment, where an individual starts with an initial idea and incompletely shifts away from their initial starting point to make an inference [22]. In egocentric anchoring-and-adjustment, people commonly begin by recruiting self-knowledge and then adjust away from this self anchor to make inferences about others (e.g., you judge someone else's preference for a food that you like to be closer to your own preference than it should be) [23,24]. Notably, the more divergent others' attributes are from a participant's, the more adjustment is needed and the longer it takes to make the inference [25,26]. These findings are consistent with the proposition that adjustment from self-generated anchors occurs serially over time [21]. Relating anchoring to neural processes, social anchoring biases, where dorsomedial prefrontal cortex (dmPFC) tracks the divergence between self and other preferences [27]. Although egocentric anchoring is well-described in social decision tasks [27], evidence of its presence in map-like knowledge representations of the social world is missing [8,15]. This dearth of evidence is due to egocentric anchoring being tested on self versus other comparisons. Such one to one comparisons don't require the type of nonlinear transformations needed to simultaneously relate different people to each other and the world in a common reference frame [8,15], or transform the position of an egocentrically viewed spatial cue into a location in an allocentric map of the environment [13,14].

Clues about how self-knowledge implicitly informs map-like representations of abstract knowledge come from work by Kaplan and Friston [8]. The authors had participants learn a stranger's preference for an everyday activity—relative to previously provided ratings for either themselves, a close friend, or a typical person on that same activity—and subsequently decide how the stranger's preference relates to the other two individuals' preferences. Highlighting how the metric nature of these abstract distances influenced behavior, they observed a steady increase in the quickness and accuracy of the participant's responses when the stranger's rating was closer to one option versus another. In parallel, they identified a HF region relating to the absolute distance between the stranger's rating and the choice options differently depending on whether a self-comparison was present in the absolute reference frame [8]. Still, due to self preferences always being needed in the transformation and/or comparison process, it is unclear whether egocentric anchoring may occur during the task. By removing the explicit need to use self-knowledge during the inference process, while keeping everything else the same, we can measure the implicit representation of self preferences tied to other social entities during relative to absolute reference frame transformations.

Here, to investigate whether there is an implicit use of the self on social inferences in an absolute map-like reference frame, we adapted the Kaplan and Friston [8] paradigm by collecting participants' preference ratings on everyday activities (e.g., eat spicy food, read a book, cycle to work) for themselves, two fictitious individuals, and two societal groups (people from cities and rural areas) from 1 to 9 on a 0–10 (ranging from impossible to happening at this moment) scale to allow for strangers with higher or lower ratings than the entities in the subsequent functional magnetic resonance imaging (fMRI) task (Fig 1A). To expedite learning of the fictitious individuals, we made the two individuals congruent with one of the two rated groups (a rural or urban person). Subsequently for the transformation phase that occurred while participants underwent fMRI, participants inferred a stranger's preference for an everyday activity—relative to previously provided ratings for one of the two individuals—and decided how a stranger's preference relates to a medium preference (normal group) and a preference of one of the two groups. More specifically, participants needed to decide whether the stranger's rating was closer to one of the two groups, or a normal group with a medium rating of 5 (Fig 1B). After deciding without any feedback, a jittered intertrial interval period (mean 2.5 s) featuring a black fixation point overlaid on a gray background appeared on the screen before moving onto a new trial. After the fMRI task, participants provided ratings for all entities, including the self, on the same scenarios to assess consistency/memory for ratings.

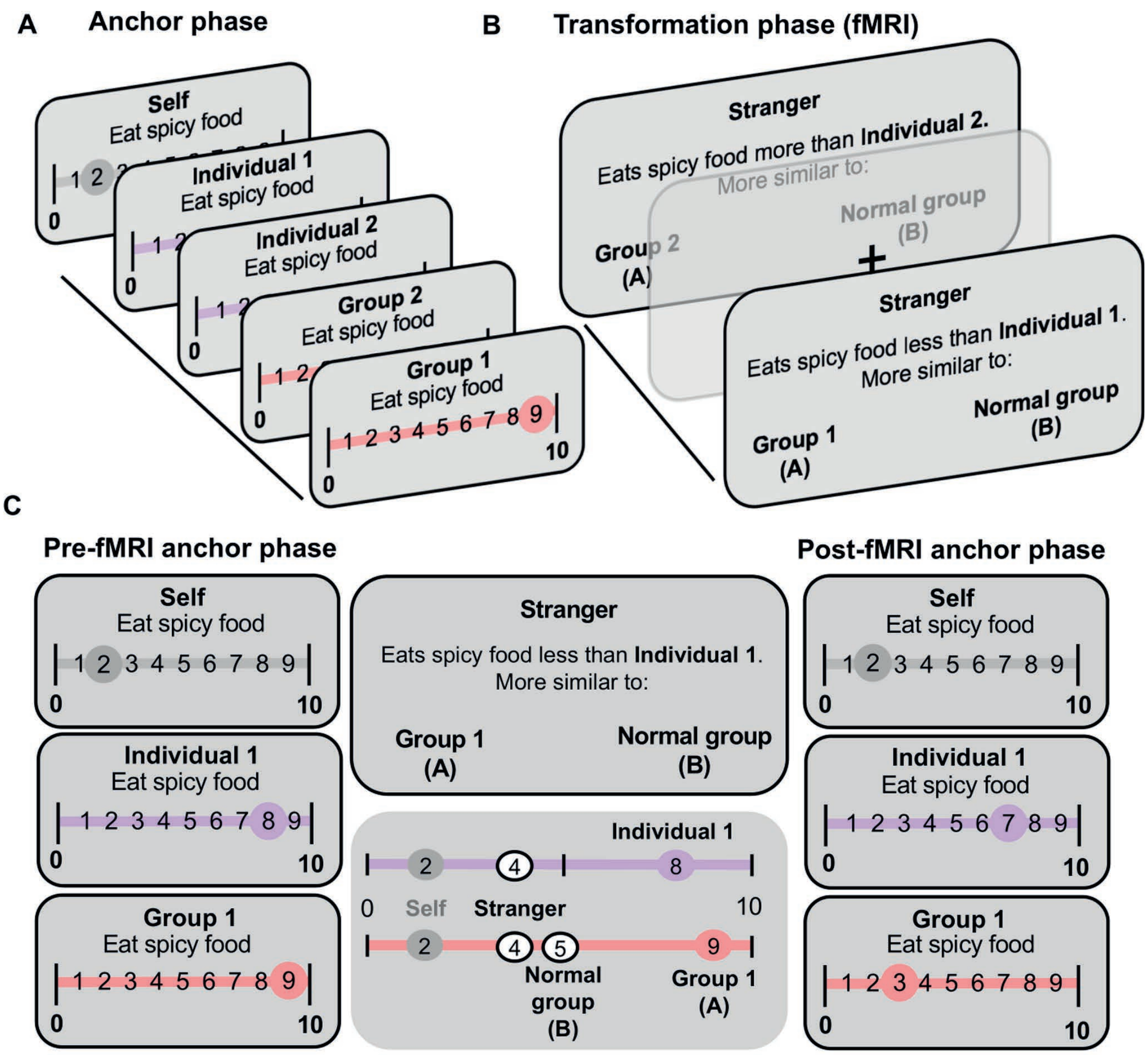
Including two different types of both individuals and groups ensured that the rated entities, the stranger, and the participant didn't have overlapping ratings too often (see Methods for more details). Furthermore, separating the groups and individuals in different stages of a transformation trial allowed us to test what task representations were most important to the HF and dmPFC (Fig 1C). Crucially, self ratings weren't necessary to perform the transformation phase of the task and unlike the previous version of the task, no spatial number line cues of preferences were provided to remove potential spatial scaffolding effects that could enhance performance. This experimental design allows us to study how the absolute distance between the stranger's rating and the ratings of the groups that are being compared with the stranger (absolute distance, AD), the differences between self and others' preferences (rating discrepancy, RD, with either groups or individuals), the cognitive demand of how far the stranger's preference is away from the individual's rating (rescaling), and the memory for self and others' preferences (rating consistency, RC, for groups or individuals) are represented during reference frame transformations of social knowledge (see Methods and Fig 1D).

The motivations behind our experimental design were 3-fold: First, to replicate both the egocentric anchoring-and-adjustment rating response time (RT) and accuracy findings in the social domain [25,26]; Second, to replicate the linear relationship between accuracy and the absolute distance between the stranger's rating and the compared groups' in the transformation phase [8]; Third, to isolate the behavioral and neural representation of egocentric anchor biases during reference frame transformations of social knowledge. We predicted that egocentric anchoring-and-adjustment would be present in map-like representations of social knowledge in the hippocampal-entorhinal system and dmPFC. We employed representational similarity analysis (RSA) here because we were interested if trial-level differences in self-other (group or individual) rating discrepancy, independent of preference type (i.e., food, transportation, music, etc.), would be represented in a common format in the aforementioned regions during our social inference task.

Results

Behavioral results

During the anchor phase, the average time that participants took to provide ratings of others was 3.12 s (SEM = 0.085). We ran a linear mixed effects model to test the influence of various



D

Behavioral Variable	Equation	Example
Stranger Preference (SP)	$\text{preference}_{\text{individual}} - \lvert((0 - \text{preference}_{\text{individual}}) \times 0.5)\rvert$	$SP = 8 - \lvert(8 * 0.5)\rvert = 4$
Absolute Distance (AD)	$\lvert SP - \text{preference}_{\text{group}}\rvert - \lvert SP - \text{preference}_{\text{normal group}}\rvert$	$AD = \lvert\lvert 4-9\rvert - \lvert 4-5\rvert\rvert = 4$
Rating Discrepancy (RD)	$\lvert \text{preference}_{\text{self}} - \text{preference}_{\text{others}}\rvert$	$RD_{\text{individual}} = \lvert 2-8\rvert = 6$ $RD_{\text{group}} = \lvert 2-9\rvert = 7$
Rating Consistency (RC)	$\lvert \text{preference}_{\text{pre-fMRI anchor}} - \text{preference}_{\text{post-fMRI anchor}}\rvert$	$RC_{\text{self}} = \lvert 2-2\rvert = 0$ $RC_{\text{individual}} = \lvert 8-7\rvert = 1$ $RC_{\text{group}} = \lvert 9-3\rvert = 6$
Rescaling	$\lvert SP - \text{preference}_{\text{individual}}\rvert$	$Rescaling = \lvert 4-8\rvert = 4$

Fig 1. Experimental paradigm. (**A**) Anchor phase. Before and after fMRI scanning, participants provide a likelihood rating for themselves in different everyday scenarios, as well as confidence ratings about each preference. After reading a description of two different fictive individuals and being presented with two different social groups, participants rate the likelihood of these entities from 1 to 9 (very unlikely to very likely) on a scale of 0 to 10 to partake in a given everyday scenario. (**B**) Transformation phase. During fMRI scanning, participants infer a stranger's preference relative to one of the previously rated individuals in a particular scenario and in a two-alternative forced choice (2AFC) determine whether that stranger's preference is more similar to a normal group (participants were informed beforehand that this group always has a preference rating of 5), or one of the two groups they previously rated (using the preference rating provided by the participant for that entity). The 2AFC was self-paced (mean = 4.48s) and the intertrial interval

duration lasted a mean of 2.5 s (range 1–4 s) over four runs. (**C**) Example fMRI trial where participants' pre- and post-fMRI task ratings for themselves (the rating in gray) and the other relevant entities (individuals in purple, groups in pink) are highlighted. Ratings for the self aren't needed to accurately perform the fMRI task. (**D**) Table describes how the experimental variables of interest would be calculated for the given preference ratings in this example trial in C. Stranger preferences (SP) are only used to inform the other behavioral variables and aren't analyzed in isolation in this study. Crucially, post-fMRI task ratings are only used for the Rating Consistency (RC) variables.

https://doi.org/10.1371/journal.pbio.3003050.g001

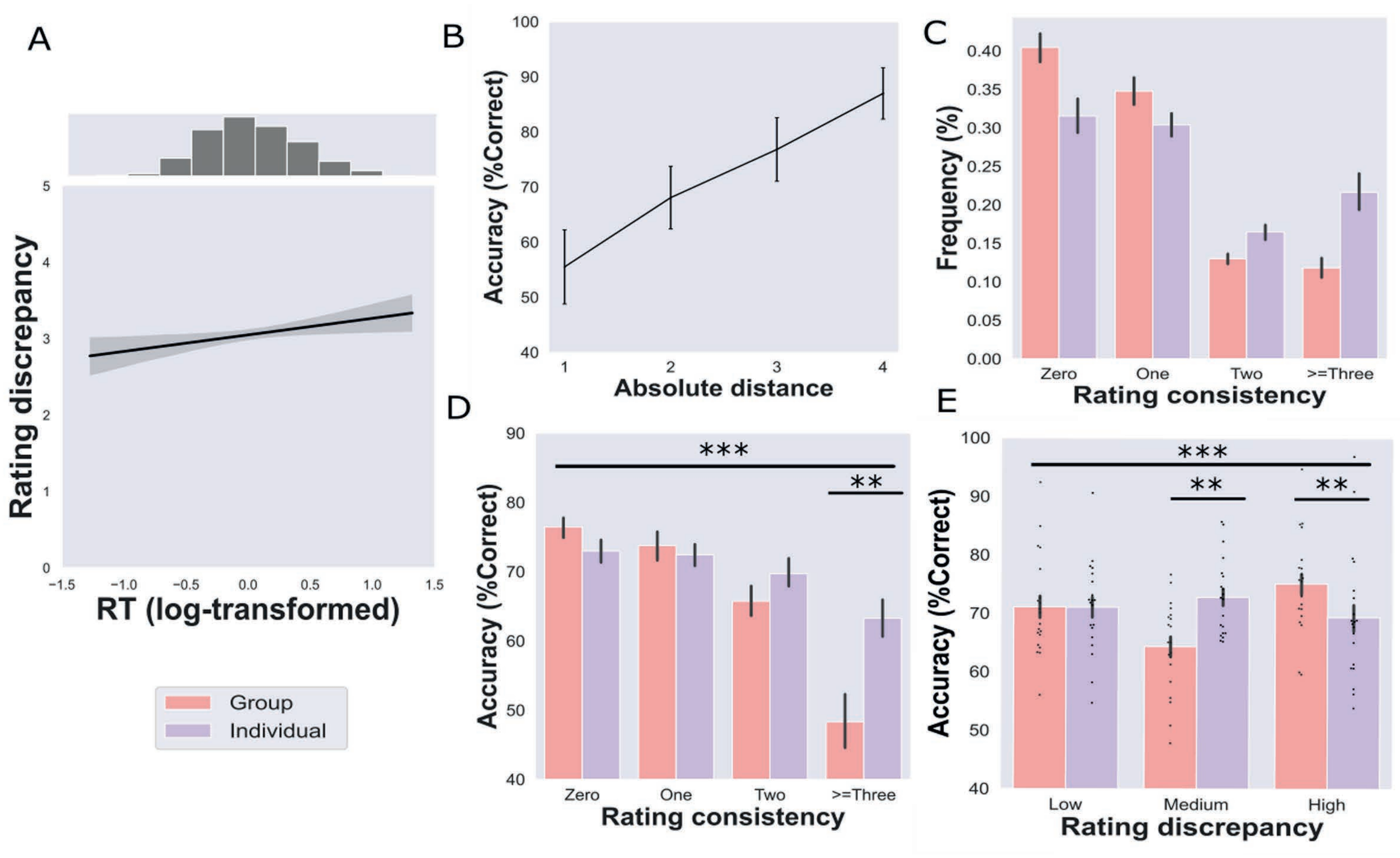


Fig 2. Behavioral results: (**A**) Positive relationship between self vs. other entity rating discrepancy (RD) and reaction time (RT: centered by group mean). (**B**) Significant relationship between accuracy and absolute distance between the stranger's rating and two choice options for each trial. (**C**) Distribution of group and individual rating consistency (RC). (**D**) Significant difference in accuracy for rating discrepancy for groups and individuals. (**E**) Significant differences how individual and group rating discrepancy relate to accuracy. All error bars showing mean ± SEM with dots representing individual participants. Asterisks showing significant differences: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. The underlying data for this figure can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2 under "Data".

https://doi.org/10.1371/journal.pbio.3003050.g002

factors on determining self versus other discrepancy for each participant. We observed that the anchor entity, individuals and groups, influenced how participants judged the target. A significant effect of entity was observed ($b = 0.077$, $t(19) = 2.10$, $p = 0.036$), indicating that participants rated the preferences of individual entities closer to their own preferences than those of group targets. Examining the hypothesized positive linear relationship between RT and self versus other discrepancy, we found a significant effect of RT on self-other discrepancy scores ($b = 0.21$, $t(19) = 2.18$, $p = 0.030$). This result suggests that increases in RT correspond to increases in the magnitude of correction away from a self-anchor—evidence of the serial adjustment present in anchoring-and-adjustment (Fig 2A).

To test if serial adjustment from self-knowledge occurred exclusively when the target was an individual or a group, we explored whether the relationship between RT and self–other discrepancy depended on entity type (i.e., individuals or groups). There was no significant interaction of the target entity on RT ($b = -0.07$, $t(19) = -0.68$, $p = 0.49$). These findings replicate prior social anchoring-and-adjustment studies [25,26], aligning with the idea that participants anchor on self-knowledge and serially adjust away from this anchor when evaluating both individuals and groups.

To study whether self-other discrepancy was consistent across the anchor phase, we calculated and compared pre-fMRI and post-fMRI RD using the ratings provided in each of the anchor phases (Fig 1D). We observed significant differences between pre-fMRI and post-fMRI RD ($\mathrm{RD}_{\mathrm{individual}}$ $t(19) = -8.99$, $p < 0.001$ and $\mathrm{RD}_{\mathrm{group}}$ $t(19) = -9.6$, $p < 0.001$). This result suggests that the RD between participants' ratings and the rated social entities decreased after participants performed the fMRI task.

Following previous work [8], participants' ratings for groups and individuals were generally consistent before and after scanning (Fig 2C); participants were on average within ±1 of their original ratings for 75.2% (SEM = 1.84%) of group rating trials and only made large deviations from their original ratings ≥3 on 11.8% (SEM = 1.28%) of trials. For individual rating trials, they were on average within ±1 of their original ratings for 61.9% (SEM = 1.94%) and only made large deviations from their ratings ≥±3 on 21.6% (SEM = 2.56%). For self preferences, participants had high consistency in recalling their own preferences, with a 74.40% of trial consistency within 1 of their original ratings, and 5% (SEM = 1.17%) of trials displayed large deviations ≥3. When relating subjective self-confidence to self rating consistency, confidence ratings significantly correlated with rating consistency ($t(19) = -5.09$; $p < 0.001$), where higher confidence ratings yielded more consistent ratings.

During the fMRI experiment, participants made 70.9% of decisions correctly (SEM = 2.98%, $n = 20$), after eliminating trials for which either answer was correct (i.e., same or equidistant ratings), while mean RT was 4.48s (SEM = 0.11). Replicating previous findings in Kaplan and Friston [8], there was a positive correlation between the absolute distance of the choice options from the stranger and task performance (group-level statistics: $t(19) = 12.54$, $p < 0.001$; Fig 2B), and a negative correlation with RT (group-level statistics: $t(19) = -2.65$, $p = 0.01$). Notably, we identified a negative correlation between RT and accuracy (group-level statistics: $t(19) = -5.33$, $p < 0.001$), indicating quicker RTs for accurate choices.

Furthermore, our investigation focused on the impact of egocentric anchoring-and-adjustment during the transformation phase, particularly in the absolute reference frame. Examining the effect of RD and entities on choice accuracy, we observed a significant interaction between RD and entities ($F(2,38) = 11.30$, $p = 0.0015$, partial eta squared ($\eta^2 p$) $= 0.37$), where significant differences on performance were observed in the medium and high levels of RD between individuals and groups (mean differences in the medium bin = 8.5%, $p = 0.004$; mean differences in high bin = 5.73%, $p = 0.007$) (Fig 2E). These significant differences highlight the presence of a u-shaped effect for $\mathrm{RD}_{\mathrm{group}}$ with fMRI task performance that wasn't present with $\mathrm{RD}_{\mathrm{individual}}$. A u-shaped relationship between $\mathrm{RD}_{\mathrm{group}}$ and accuracy means that egocentric biases in the absolute reference frame hurt task performance when the self rating and the presented groups' ratings on a given trial weren't too similar or different (Fig A in S1 Text). Next, we examined the influence of rating discrepancy on accuracy separately for individuals and groups. For individuals, a one-way ANOVA with three levels of RD (low, medium and high) showed no significant differences in accuracy ($F(2,38) = 1.08$, $p = 0.34$, $\eta^2 p = 0.054$, Fig 2E). Additionally, no significant quadratic relationship was found ($t(19) = -1.16$, $p = 0.87$), indicating that self versus individual rating discrepancy ($\mathrm{RD}_{\mathrm{individual}}$) didn't significantly impact task performance. In contrast, a one-way ANOVA for performance by self-group rating discrepancy (RD_{group}) revealed significant differences in accuracy ($F(2,38) = 10.8$, $p \leq 0.001$, $\eta^2 p = 0.36$, Fig 2E). Moreover, a significant quadratic effect was observed ($t(19) = 3.39$, $p = 0.0015$, and Fig A in S1 Text), providing further support that the relationship between RD_{group} and performance follows a nonlinear u-shaped pattern.

Exploring the effect of memory on choice accuracy, we observed a significant interaction between rated entities and RC ($F(3,57) = 6.59$, $p < 0.001$, $\eta^2 p = 0.25$). In a post-hoc analysis, we observed significant differences between individuals and groups for the very inconsistent ratings ≥±3 bin (mean differences = 14.97%, $p = 0.004$) (Fig 2D). Then, we examined the relationship between performance and rating consistency (including memory for individuals and groups). We observed a significant correlation between these two variables ($\rho = -0.59$, $p = 0.005$). Next, we separately analyzed the correlation between accuracy and RC_{group} ($\rho = -0.54$, $p = 0.013$) and $RC_{individual}$ ($\rho = -0.52$, $p = 0.017$). For both entities, we observe a significant negative correlation between accuracy and memory for the rating of others. In other words, the more consistent participants rated others' preferences, the better they performed the fMRI task.

In our behavioral analyses, we observed an intricate relationship between egocentric anchoring-and-adjustment and flexible social comparisons requiring reference frame transformations. Replicating previous behavioral findings [8], we confirmed that AD negatively relates to the difficulty of a trial and egocentric anchoring-and-adjustment is present when rating entities in the task. In a subsequent attempt to disentangle the effects of memory (RC) and egocentric anchoring-and-adjustment (RD) on transformation trials, we found that memory was better for groups' preferences (RC_{group}). Notably, memory for both entity types ($RC_{individual}$ and RC_{group}) impacted how well participants flexibly compared a stranger's preferences to different social entities that didn't include themselves. Additionally, egocentric anchoring-related social comparison performance effects were selective to entities in the absolute reference frame (groups), where performance was lowest on mid-level rating discrepancy bins. Notably, this effect was absent when analyzing the relationship between accuracy and egocentric anchoring to the entity in the relative reference frame ($RD_{individual}$). Taken together, these behavioral findings point towards egocentric anchoring exerting the most influence on social comparisons in the map-like absolute reference frame during the transformation task.

fMRI analysis

RSA searchlight. To test which regions represented egocentric anchor biases and other social inference demands during the transformation fMRI task, we performed whole-brain searchlight RSA. Neural representational dissimilarity matrices (RDMs) were computed in a sphere centered on each voxel of the brain, and their relationships to behavioral RDMs were investigated (Fig 3A). This allowed us to determine whether the neural representation of anchor biases in both relative (self rating discrepancy with individual rating: $RD_{individual}$) and absolute (self rating discrepancy with compared groups: RD_{group}) reference frames is different after explaining variance related to other important cognitive aspects in the transformation phase. The additional cognitive predictors were the absolute distance between the stranger and choice options (AD), that allow us to control for trial difficulty, rescaling of the relative position of the stranger's rating in relation to the individual to its absolute position on the rating scale, as well as rating consistency for groups (RC_{group}) and individuals ($RC_{individual}$).

In a whole-brain general linear model (GLM) analysis, we observed a relationship between group rating discrepancy (RD_{group}) and left hippocampus pattern dissimilarity ($x = -22$, $y = -18$, $z = -18$; $t(19) = 3.78$, bilateral hippocampus small-volume corrected peak-voxel $p = 0.046$; Fig 3B), as well as left EC pattern dissimilarity ($x = -18$, $y = -24$, $z = -22$; $t(19) = 4.37$, bilateral EC small-volume corrected peak-voxel $p = 0.015$; Figs 3A and, Fig B and Table B in S1 Text).

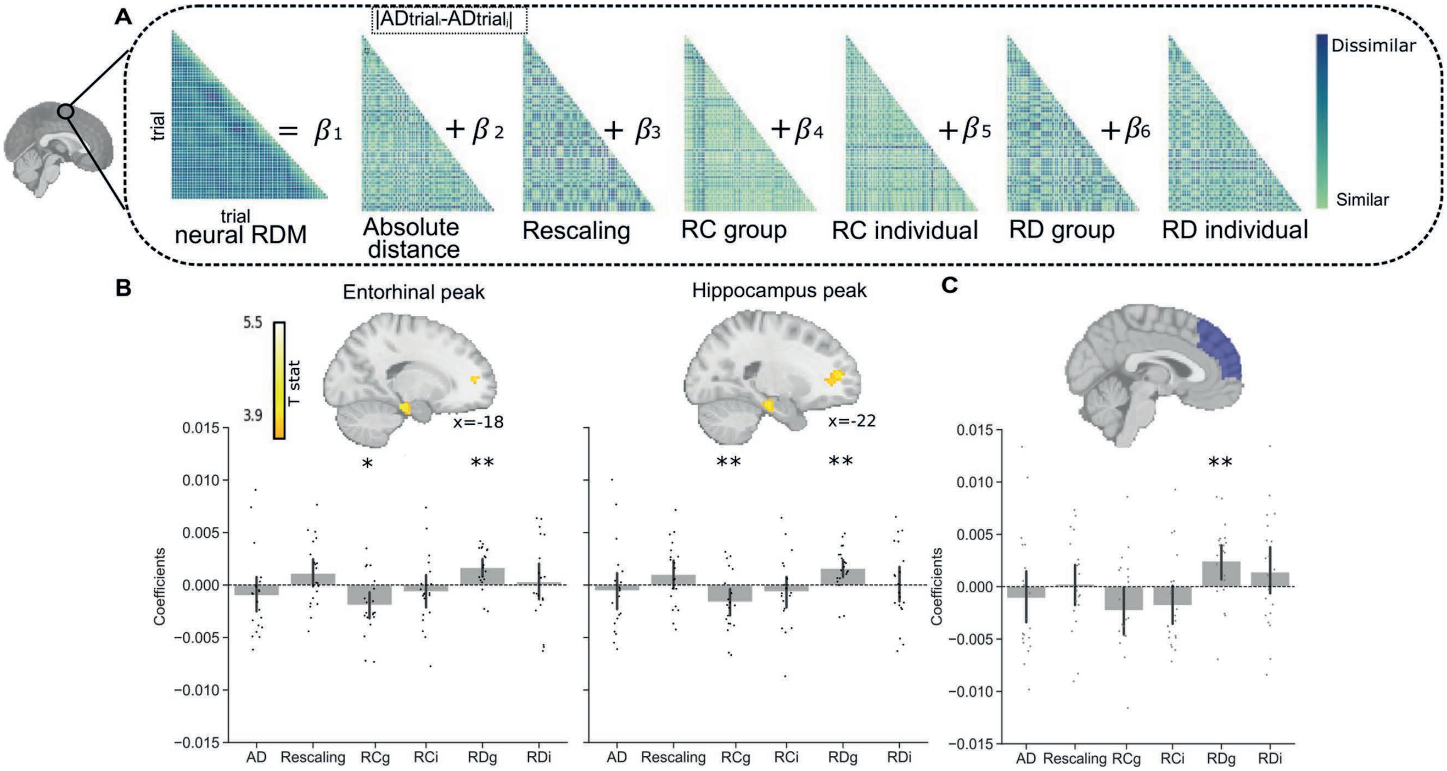


Fig 3. Egocentric anchoring-and-adjustment in the hippocampal formation. (**A**) Neural and behavioral representational dissimilarity matrices (RDMs) for the general linear model (GLM) searchlight analysis for one representative participant. The neural RDM was used as a dependent variable with six predictor behavioral variables. These six variables were the absolute distance between the stranger's rating and the ratings of the groups that are being compared with the stranger (absolute distance, AD), the differences between self and others' preferences (rating discrepancy, RD, with either groups or individuals), the cognitive demand of how far the stranger's preference is away from the individual's rating (rescaling), and the memory for self and others' preferences (rating consistency, RC, for groups or individuals). To construct the behavioral matrices, absolute differences between these variables on each trial pair comparison were computed. Matrices were standardized before performing regression. (**B**) Top: Group activation maps for self vs. group rating discrepancy (RD_{group}). Differences in RD_{group} were significantly associated with left entorhinal cortex (right) (small-volume corrected family-wise error (FWE) for bilateral entorhinal cortex mask, peak-voxel corrected $p = 0.015$) and left hippocampus pattern dissimilarity (left: small-volume corrected FWE for bilateral hippocampus mask, peak-voxel corrected $p = 0.046$). Bottom: Barplot showing coefficients (mean ± SEM) for a 10-mm sphere around entorhinal and hippocampal peaks for each behavioral variable of interest. (**C**) Beta coefficients of RSA GLM in ROI of dmPFC. Bars showing beta coefficients resulting from the GLM for each behavioral variable in the dmPFC, where there is a significant dmPFC effect for RD_{group} ($t(19) = 2.92$, $p = 0.008$). Each dot represents an individual participant. $^{**}p < 0.01$, $^{*}p < .05$. The underlying data for Fig 3B and 3C can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2 under "Data".

https://doi.org/10.1371/journal.pbio.3003050.g003

Testing whether egocentric anchoring to the group in the hippocampus and EC was also present in individuals, we observed no similar effect in this hippocampal region ($t(19) = 0.35$, $p = 0.72$) or EC ($t(19) = 0.10$, $p = 0.91$) for $RD_{individual}$. In a subsequent t test, we tested whether this effect was specific to RD_{group}, where there was not a significant difference between hippocampal pattern dissimilarity for RD_{group} versus $RD_{individual}$ ($t(19) = 1.48$, $p = 0.15$), nor for EC pattern dissimilarity ($t(19) = 1.71$, $p = 0.10$). We then tested whether the effect found in the hippocampus peak and entorhinal peak was exclusive to the RD_{group} predictor. Using a 10 mm sphere, we examined in a post-hoc region of interest (ROI) analysis if this effect was present in the beta maps of other predictors. We found a significant effect for RC_{group} in both the hippocampus peak ($t(19) = -3.00$, $p = 0.007$) and the entorhinal peak ($t(19) = -2.53$, $p = 0.02$), but not for the other predictors (Fig 3B). Notably, these RC_{group} effects were also measurable at the whole-brain level in both the left EC ($x = -18$; $y = -34$; $z = -6$; cluster-level FWE $p = .021$) and hippocampus ($x = -16$; $y = -34$; $z = -4$; cluster-level FWE $p = .038$; Fig B and Table B in S1 Text), A HF representation of RC is consistent with its putative role in mnemonic function [3].

Given our previous hypothesis about the role of dmPFC in egocentric anchoring-and-adjustment, we investigated whether there were any significant effects there. Despite not finding a significant dmPFC cluster that represented any variable of interest, we checked whether any effects were present in an ROI of the dmPFC. In the ROI, we observed a significant effect for RD_{group} ($t(19) = 2.92$, $p = 0.008$, Fig 3C). There were no significant dmPFC results for the other predictors. Testing whether the dmPFC rating discrepancy effects were selective to RD_{group}, a subsequent t test revealed that there was no significant difference between dmPFC RD_{group} and $RD_{individual}$ pattern dissimilarity ($t(19) = 0.74$, $p = 0.46$). Checking whether any other brain region significantly represented any variables in our GLM, we observed a significant peak for RD_{group} in the left ventrolateral prefrontal cortex ($x = -54$, $y = 20$, $z = 2$; $t(19) = 5.55$, peak-level FWE $p = 0.039$, Fig B and Table B in S1 Text related to Fig 3B), but no other significant effects for any variables anywhere else in the brain.

To rule out that the HF, dmPFC, or anywhere else in the brain represented the self as the center of the 0–10 absolute preference rating scale on each transformation trial, we conducted a control RSA. The control analysis consisted of a whole-brain RSA searchlight informed by the proportion of 'self recentering' (i.e., the amount of recentering needed to move a self preference rating to the center of the rating scale) on each trial. Specifically, we calculated self recentering as the absolute difference between the self rating and 5 (i.e., the normal entity's rating on every trial) divided by the absolute difference between the self rating and the furthest extreme of the scale (0 or 10) from the self rating. Crucially, we didn't observe any significant RSA effect related to self recentering in the hippocampus peak ($t(19) = 0.13$; $p = 0.89$), entorhinal peak ($t(19) = 0.17$; $p = 0.86$), dmPFC ROI ($t(19) = 0.59$; $p = 0.55$), or anywhere else in the brain.

Pattern dissimilarity and RD_{group}. We then tested whether the behavioral influence of egocentric anchoring on our task was also present in HF and dmPFC representations. To investigate whether pattern dissimilarity in the HF and dmPFC followed the u-shaped effect found in the behavioral results between RD_{group} and Accuracy (Fig 2E and Fig A in S1 Text), we conducted quadratic regression analyses. These analyses let us determine if there was a linear or quadratic effect between pattern dissimilarity and RD_{group}. However, we didn't find a significant u-shape effect in any region examined (hippocampus peak $t(19) = -2.05$, $p = 0.97$, entorhinal peak $t(19) = -1.08$, $p = 0.85$) and dmPFC ($t(19) = -1.26$, $p = 0.88$). We also investigated the possibility of a linear effect between pattern dissimilarity and RD_{group}. However, no significant linear effects were found in the hippocampal peak ($t(19) = -2.41$, $p = 0.98$), entorhinal peak ($t(19) = -0.97$, $p = 0.82$), and the dmPFC ROI ($t(19) = -0.44$, $p = 0.67$). Additionally, we investigated whether participants who changed their pre- to post-fMRI ratings closer to their own preferences (increased egocentric anchoring) exhibited different neural representations compared to those who exhibited less of this distortion. However, we didn't observe any significant individual differences in neural representations in the hippocampal peak (($\rho = -0.27$; $p = 0.26$), entorhinal peak ($\rho = -0.19$; $p = 0.42$), or dmPFC ($\rho = -0.23$; $p = 0.33$) related to increased egocentric anchoring over the course of the task.

RSA social entity model comparison. To examine whether the hippocampus, EC, and dmPFC preferentially represented group identity, individual identity, or trial congruence during transformation trials, we generated trial-by-trial correlation matrices and contrasted them with the three aforementioned conceptual matrices. A trial was considered congruent when the individual and the social group involved in the transformation trial generally aligned to the same social category (i.e., an individual with an urban or rural background being paired with a respective city or village dweller group in the condition). This analysis allowed us to explore the impact of various task features on the regions' pattern dissimilarity. We hypothesized that trials featuring the same groups would exhibit greater similarity in their hippocampal representation than trials featuring the same individuals or congruent individual-group entities. For all of the regions studied, we observe significant differences in model fits (hippocampus peak, $F(2,38) = 17.47$, $p < 0.001$; entorhinal peak, $F(2,38) = 21.70$, $p < 0.001$ and dmPFC $F(2,38) = 12.88$, $p < 0.001$). Following a significant ANOVA for each brain region, a paired T test was performed to test pairwise differences between the three models. In the hippocampus peak, there was a significant difference between the group versus congruence models ($t(19) = 2.40$, $p = 0.03$) and group versus individual models ($t(19) = 4.89$, $p < 0.001$). Lastly, for the entorhinal peak, we observed a significant difference between group versus congruence models ($t(19) = 3.85$, $p = 0.001$) and group versus individual models ($t(19) = 5.41$, $p < 0.001$). These results demonstrate a consistent pattern across the regions, where the group model consistently outperformed both the individual and consistency models, suggesting that the group model better captures the underlying neural representations. In slight contrast for the dmPFC, there was no significant difference between the group versus congruence models ($t(19) = 1.97$, $p = 0.06$), but there was still a significant difference between group versus individual models ($t(19) = 5.15$, $p < 0.001$). Notably, there remains a possibility that the negative results for the individual identity model could be a byproduct of representational overlap with the group model when a condition for a trial has both the same individual and group identity.

We then ran a searchlight RSA to test if there were other brain regions that were significantly modulated by the difference between the three models. We observed two significant peaks: one in the thalamus ($x = -7$, $y = -16$, $z = 9$; $F(2,38) = 13.42$, pFWE < 0.001, Fig C and Table D in S1 Text) and another in the ventral striatum ($x = 27$, $y = 18$, $z = 6$; $F(2,38) = 12.38$, pFWE < 0.001, Fig C and Table D in S1 Text). For the thalamus, we observed significant differences for the individual versus congruence model ($t(19) = 3.88$, $p < 0.001$) and group versus congruence model ($p < 0.001$, $t(19) = 3.63$), while not finding any significant difference between individual and group models ($t(19) = -0.12$, $p = 0.9$). For the ventral striatum peak, we observed significant differences between the group versus individual models ($t(19) = -2.58$, $p = 0.018$), individual versus congruent model ($t(19) = 3.35$, $p = 0.0034$), and group versus congruence models ($t(19) = 4.25$, $p = 0.002$).

In sum, we tested whether neural representations of implicit egocentric anchoring biases are present in the dmPFC, HF, or anywhere else in the brain when making social inferences that require flexibly comparing a stranger's preference to multiple social entities in a map-like absolute reference frame. We found that the HF including the EC, the dmPFC, and ventrolateral PFC maintained representations of preferences in the absolute reference frame that were implicitly anchored to self preferences. Additionally, HF representations signaled how well preferences in the absolute reference frame were remembered, where preferences that were better remembered were represented more distinctly. Lastly, the HF and dmPFC were found to prioritize the identity of social preferences in the map-like reference frame.

Discussion

We tested whether map-like representations of abstract social knowledge are susceptible to egocentric anchoring-and-adjustment. Replicating findings from social anchoring [25,26], we observed egocentric anchoring-and-adjustment when participants rated entities (Fig 2A) and map-like social inference behavioral performance effects from the Kaplan and Friston [8] paradigm (Fig 2B). Highlighting the presence of egocentric anchoring in the HF and dmPFC, hippocampal-entorhinal and dmPFC pattern similarity was related to self versus group rating discrepancy (Fig 3). Notably, the hippocampus and EC also represented how well group preferences were remembered, which was a key indicator of task performance

(Fig 3). Testing whether the HF and dmPFC preferentially represent group identity, individual identity, or trial congruence during transformation trials, we find that the HF selectively represents group identity (Fig 4). In what follows, we speculate on how egocentric anchoring potentially shapes abstract knowledge codes and relate our findings to the wider literature.

The generality of egocentric biases on allocentric coding

Our experiment doesn't test whether HF and dmPFC representations of egocentric anchor biases during map-like inferences are selective to the social domain [12], or are more generally applicable to abstract knowledge [28]. Comparable work has uncovered the co-existence of egocentric and allocentric representations in different brain structures for abstract knowledge [29], which allows for the possibility that these parallel representations could be flexibility modulated depending on task demands. The switch between egocentric and allocentric processing could be flexibly adapted depending on task demands, which is supported by recent evidence showing personal biases can facilitate emotion prediction [30] and learning more generally [31]. Future work can study the relationship between egocentric anchoring and hippocampal generalization to uncover when self-projection is beneficial or detrimental in spatial and non-spatial inference tasks [21,32]. Despite observing significant egocentric anchoring representations in the HF and dmPFC, we can't confirm that this subjective computation occurs on every trial across all participants. Future work with larger samples can more closely examine how individual differences in egocentric anchoring-and-adjustment changes HF and dmPFC representations during flexible social comparisons.

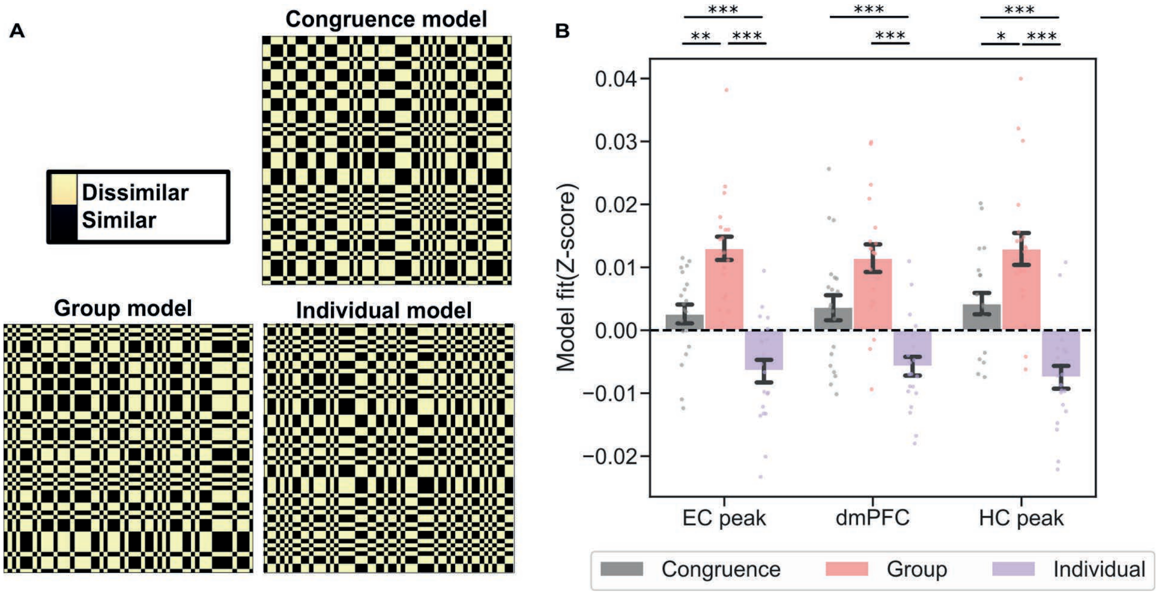


Fig 4. Neural representational model comparisons. (**A**) Model matrices based on hypothesized representational conditions: congruent vs. incongruent individual and group entities (congruence), group trial identity (group), and individual trial identity (individual). (**B**) Correlation between model RDMs and neural RDM in the left hippocampus peak (HC), left entorhinal peak (EC), and dmPFC, where group trial identity is the winning model. All error bars showing mean ± SEM with dots representing individual participants. *$p < .05$, **$p < 0.01$, ***$p < 0.001$. The underlying data for Fig 4B can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2 under "Data".

https://doi.org/10.1371/journal.pbio.3003050.g004

Effects of social cognition and relationship with the hippocampal formation

Our findings add to a growing literature highlighting the importance of the HF in assimilating social knowledge [7–10,33,34], although an important caveat to these findings is that no study has demonstrated they are specific to cognitive map-like or non-isometric/cartesian representations of social knowledge. Further investigation of the effect of subjective biases on cognitive maps of social space [7,9,10] and assimilation of knowledge in social networks [35–38] could yield insights into whether there are differences in subjective and objective map-like coding. Moreover, investigating subjective confidence and its relationship to egocentric anchoring in the HF could be a promising avenue of investigation. Lastly, we don't specifically investigate spatial or social perspective switching computations that likely rely on regions like the retrosplenial cortex (RSc) [13]. Future work can address whether social and spatial perspective transformations require similar underlying computations and how they might differ in areas like dmPFC or RSc [15,39].

The specificity of anchoring to groups versus individuals in allocentric reference frames

For neural representations of rating discrepancy (egocentric anchoring) and consistency (i.e., memory strength), we find that our ROIs represent these behavioral variables for group, but not individual ratings. Moreover, these RSA effects are paralleled by behavioral performance results (Fig 2D and 2E), where the accuracy of social inferences is most strongly determined by self-group rating discrepancy and group rating consistency. Taken together, these results suggest that the most crucial aspects of flexible map-like social inferences are how well attributes of the social groups in our task are learned and remembered in the absolute reference frame. An important caveat to these HF and dmPFC findings is that we cannot determine whether our effect is specific to egocentric anchoring for groups, or more generally relevant to any social entity in the absolute reference frame. This is especially important because groups or any collective minded thinking can be seen as more allocentric [40], independent of any explicit reference frame transformation. Moreover, rating consistency was much lower for individuals than groups. This was likely because the groups, city and village dwellers, were more familiar than specific fictive individuals to participants. To resolve this issue, future studies can alternate whether individuals or societal groups are in the initial relative position to the stranger in our task and test how that changes task performance. Notably, different individuals were interchangeably used as the initial reference point and choice options in Kaplan and Friston (2019), which had similar behavioral results as this study. Therefore, we predict that our fMRI results would hold if the stranger was learned relative to a group and subsequently compared to individuals in the absolute reference frame.

Relating dmPFC effects to the social anchoring literature

Building on previous work by Tamir and Mitchell [27], we observed a relationship between dmPFC representational similarity and social anchoring (self versus group rating discrepancy) in our task. Previous work isolated dmPFC fMRI signals that related to self-other discrepancy when inferring mental states of one individual [27]. In contrast, our dmPFC self-other discrepancy results extend to nonlinear social inferences that involves learning a stranger's preference relative to an individual and then extrapolating whether the stranger's preference is more similar to one of two groups' preferences. Consequently, we show that dmPFC isn't limited to its putative role of guiding one to one self-other comparisons [27,41] and can extend to more complex social inference of relating multiple social entities to each other.

dmPFC versus hippocampal-entorhinal contributions to social learning

Despite the dmPFC and HF being commonly linked to social [27,41] and spatial [19] learning, respectively, we find that both regions represent egocentric anchor biases during social inference. One of the key differences between the dmPFC and hippocampal-entorhinal system in our task is that the HF also represented how consistently preferences were rated. Finding hippocampal-entorhinal representations of how well groups' preferences are remembered is consistent with the region's putative role in long-term memory [3]. Additionally, we find that the HF selectively represents group identity during transformation trials. In contrast, we observed that the dmPFC represented group identity over individual identity, but not individual-group identity congruence. These results are consistent with the idea that the HF maintains a map-like representation of one's social network [15], while dmPFC has a more general role in integrating multiple sources of information [42]. Taken together with work showing that knowledge storage of these self preference ratings relies on ventral mPFC [43], our results imply that dmPFC is ideally positioned in the brain to interface self-knowledge with map-like knowledge of others in the hippocampal-entorhinal system.

Conclusion

We highlight how personal preferences are present in the HF and dmPFC when making inferences about how others relate to each other. Future work can determine whether similar egocentric biases are transferable to cognitive map-like representations in spatial and other non-spatial domains [17]. By further disentangling the role of the self in shaping hippocampal models of the world, a better understanding of variability in allocentric representations will be formed.

Methods

Ethics statement

Study participants were compensated and gave informed written consent to participate. The study was approved by the local research ethics committee at Universitat Jaume I (ethics reference: CD/10/2022). The study was conducted in accordance with Declaration of Helsinki protocols.

Participants

Thirty participants (16 female, with a mean age of 21.16 years, SD = 3.46) took part in an experiment conducted at Universitat Jaume I, using an online recruitment system. Four participants were excluded due to poor performance (<60% accuracy). A further four participants were excluded due to excessive movement during the experiment and two participants were omitted due to technical issues during image acquisition resulting in a final sample of 20 participants. All participants were right-handed, had normal or corrected-to-normal vision, and had no history of neurological or psychiatric disorders.

Task

Stimuli were presented using PsychoPy (v 2020.1.1) toolbox running in Python 3.9. The task consisted of three distinct phases: anchor phase, transformation phase fMRI task, and rating consistency phase (Fig 1).

Anchor phase

Participants first provided likelihood ratings for 105 everyday activities adapted from [8] (Table A in S1 Text) on a scale from 0 to 10 (0 indicating impossibility, 10 indicating extreme likelihood). Following each question, participants provided a confidence rating on a scale from 1 to 5 (from no idea to very sure), reflecting their confidence regarding their own preferences. To achieve a more varied distribution of ratings, we excluded the 30 trials that were nearest to the mean of each participant's preference, resulting in 75 trials for the subsequent phase.

Participants were then introduced to two individuals with a brief paragraph description and the name of two stereotypical groups (city dwellers and village dwellers). To make the individuals easier to learn, each fictitious individual was created to have preferences either more compatible with city or rural village living and were gender matched to the participant. Participants were given as much time as they needed to read the full description of the two individuals. Subsequently, participants judged the entities' preferences. Using the same 0–10 scale for self rating, participants expressed what they felt was the likelihood of each entity partaking in a variety of everyday scenarios within a 10 second time limit [8]. Participants only provided preference ratings for a given scenario once before and after the fMRI task.

Transformation fMRI phase

Participants then performed a brief practice version of the fMRI social decision-making task (adapted from Kaplan and Friston [8]) outside of the scanner, until they achieved a performance level of at least 70%. During the transformation task, participants inferred a stranger's preference relative to one of the familiar individuals using their previous ratings for the individual entities. Subsequently, in a two-alternative forced choice task (2AFC) they determined whether that stranger's preference is more similar to a normal group (always a 5), or one of the two groups they previously rated [8] within 10 s. Participants estimated the stranger's preference using a scenario, and a cue of less or more related to the preferences of the previously presented individual (e.g., the stranger eats spicy food more/less than the individual). If the cue was "more", participants had to add half of the difference between the individual preference and the high extreme of the scale (10) to the individual entity's preference. If the cue was "less", participants had to subtracted half of the difference between individual preference and the low extreme of the scale (0 to the individual entity's preference). In contrast with the Kaplan and Friston study that had participants compute the stranger as slightly (¼) or very (¾) less/more than the highly familiar initial reference individual [8], we limited it here to only halfway between the reference individual and the closest extreme. This change permitted us to reduce the cognitive demand for participants since they needed to remember their ratings for four distinct entities. Participants were instructed that it was a different stranger on each trial. Crucially, participants never received any feedback during the task.

We had four conditions, each based on the individual entity used to infer the stranger's value and the group entity presented as a choice option in the 2AFC (Fig 1A). The four different conditions were used to control for different levels of similarity between the rated entities and the participant, as well as schematic congruence between individuals and groups (level of attribute similarity between an individual and group entity in a given trial). Given the different compatibility between each rated individual and group, there were two conditions with a congruent rated individual and group (e.g., the urban oriented individual with city dwellers; the rurally oriented individual with village dwellers) and two involving an incongruent rated individual and group (i.e., the rurally oriented individual with city dwellers; the urban oriented individual with village dwellers). Due to time constraints for fMRI scanning, we excluded an additional 15 scenarios from the remaining 75 scenarios for the fMRI task, specifically those where participants rated the group preferences as 5 to avoid overlap with the normal group. Participants were therefore tested on the same 60 scenarios for all four conditions during the transformation phase, resulting in a total of 240 trials. These trials were distributed over four fMRI runs, with each run consisting of 60 trials that were assigned completely randomly. Trials were self-paced, allowing participants to determine the time for each trial (mean = 4.48 s) with a maximum allowed RT of 10s. As in the Kaplan and Friston study [8], any transformation trial where both choice options were correct wasn't included in the analyses.

All of the information needed to perform the task was presented at once on the screen, and the aspects of the task that varied from trial to trial were the egocentric reference frame (individual), the cue (that connects individuals preferences and stranger preferences), and the choice options (see Fig 1A). Participants had a maximum of 10s to make a decision, followed by a jittered ITI period (mean = 2.5 s; range = 1–4 s), where a white fixation cross on a gray background was presented. After completing the fMRI task, participants were asked to give ratings for themselves and the four entities again to test the consistency/memory for their preferences.

Anchoring-and-adjustment analysis

To study the anchoring and adjustment effect during the anchor phase, we applied a linear mixed effects model (equation 1). Our focus was on studying the relationship between participant RTs and RD, calculated as the absolute differences between participants' self-preferences and others' preferences, when inferring others' preferences. To assess differences between egocentric anchoring and adjustment for individuals and groups, we included the anchor entity and the interaction with RT as predictive variables in the model.

$$\text{RD} \sim \text{RT} + \text{anchor entity} + \text{RT}^{*}\text{anchor entity} + \left(1 \middle| \text{ participant}\right) \qquad \text{(equation 1)}$$

We tested for egocentric anchoring-and-adjustment by incorporating the fixed effect of RT, which is the positive correlation between RT and self–other discrepancy [25,26]. We added participants as nested variables in the model. Trials in which participants did not provide a response within the 10-s window during the self-rating or anchor phase were excluded from the analysis. This exclusion was necessary due to the impossibility of calculating the rating discrepancy variable in such instances. RTs were log-transformed to address their right-skewed distribution, and we excluded responses whose log RTs were ≥2.5 SD from the grand mean [44].

fMRI acquisition

Whole-brain structural and fMRI data were acquired using 3T General Electric Signa Architect magnetic resonance imaging (MRI) scanner using a 32-channel head coil. Four whole-brain functional runs were acquired using a multi-band multi-echo EPI sequence (voxel size = 3 mm isotropic, TR = 2 s; TEs = 18.5, 34.16, and 49.82 ms.; flip angle θ = 80; resolution matrix = 80 × 80, and a multi-band acceleration factor = 2). We used a functional sequence that sets the slice angle of 30° relative to the anterior-posterior commissure line to minimize signal loss in the medial temporal and orbitofrontal regions [45]. Structural T1-weighted images were acquired using an MPRAGE sequence (TR = 2.5 s; flip angle = 8, and slice thickness = 1 mm).

Beta-series modeling

Pre-processed BOLD time series data was analyzed using a Least Squares-All (LSA) GLM that afforded estimation of the specific activation patterns elicited by each trial and run. The model includes every single trial as a regressor, and various confound regressors [46] composed of six motion regressors, six physiological noise regressors, and a global signal regressor. The last regressor was included to help control for global effects or potential confounding factors that might affect the entire fMRI brain signal [47]. Analysis was performed using Nilearn v 0.10.1, a Python library for statistical learning on neuroimaging data (https://nilearn.github.io).

fMRI data analysis

Calculation of behavioral dissimilarity matrices

For each participant and run, we calculated trial by trial dissimilarity matrices for each of the behavioral variables of interest: absolute distance between choices (AD), rescaling, self versus individual rating consistency ($RC_{individual}$), self versus group rating consistency (RC_{group}), self versus group rating discrepancy (RD_{group}), and self versus individual rating discrepancy ($RD_{individual}$). These matrices were constructed by computing the absolute differences between the values of these variables across trials.

Representational similarity analysis (RSA)

We performed a whole-brain searchlight-based multiple regression RSA to examine brain regions in which changes in dissimilarity between neural activity patterns in the trials is explained by the behavioral variables of interest. To create the neural representation dissimilarity matrix (nRDM) for each voxel, a spherical searchlight was run by defining a sphere with a radius of four voxels that was moved across the brain (spherical searchlight radius = 12 mm). For each sphere, we extracted the beta coefficients for the voxels within that sphere, which are derived from the LSA-GLM model (as detailed in the Beta-series modeling section). This extraction was repeated across all trials, producing a vector of beta coefficients for each sphere location [48]. Following previous work [49], we then performed a singular value decomposition analysis to reduce the dimensionality to 10 dimensions, where these first 10 dimensions capture 76.04% variance for all participants and searchlights. These reduced vectors are then correlated to generate an nRDM, with dimensions determined by the number of trials in the transformation phase. This neural RDM is then used as the dependent variable in the model for each voxel. To investigate how changes in behavioral variables relate to neural fMRI signal patterns in different brain regions, we applied a voxel-wise GLM to predict the nRDM with six behavioral predictors (Fig 3A). Behavioral matrices were standardized before performing regressions. For each participant, the multiple regression-based RSA provided a beta map for each of the six predictors. These beta maps indicate the relative influence of each predictor on the dissimilarity of neural activity patterns within the searchlight spheres. These maps were smoothed using an 6 mm FWHM Gaussian kernel at the participant level. In a second-level analysis, the resulting beta estimates for each participant and voxel were testing the regression coefficients against zero using a one sample *t* test.

In addition to the searchlight-based RSA, we ran a ROI-based RSA in dmPFC. We computed the trial by trial nRDM, extracting the beta coefficients for all voxels within the ROI and computed Pearson correlations between all transformation trials. We also applied the aforementioned dimensionality reduction. The resulting nRDM was then introduced as the independent variable in the same GLM (Fig 3A) employed for the searchlight analysis.

To control for the presence of any type of self recentering representation, we ran a control whole-brain searchlight RSA. This analysis was informed by the proportion of 'self recentering' (i.e., the amount of recentering needed to move a self preference rating to the center of the rating scale) on each trial. We also checked for the presence of these effects in 10 mm ROIs around our hippocampal and entorhinal RD_{group} peaks, as well as in the dmPFC ROI.

Pattern dissimilarity by RD$_{group}$

We extracted the beta values within a 10 mm radius sphere around the hippocampal peak, entorhinal peak, and dmPFC ROI for each participant from the corresponding beta image for each trial. Subsequently, we concatenated all trials from different runs and constructed correlation matrices for trials that share the same RD$_{group}$ value, resulting in nine RDMs. Trials with a singular representation for an RD$_{group}$ category were excluded from the analysis. To quantify dissimilarity, we convert correlation values to dissimilarity values using the 1 – Pearson correlation. Following this transformation, we calculated the mean pattern dissimilarity value for each nRDM.

To study the relationship between accuracy and pattern dissimilarity with RD$_{group}$, we calculate two regression models for each participant. We add a quadratic term $\left(RD_{group} - \overline{RD_{group}}\right)^2$ to account for potential symmetrical quadratic relationships, where the $\beta 2$ coefficient expresses the strength of the quadratic relationship between dependent variables and RD$_{group}$.

$$Accuracy = b + \beta 1 * RD_{group} + \beta 2 * \left(RD_{group} - \overline{RD_{group}}\right)^2$$

$$Pattern\ dissimilarity = b + \beta 1 * RD_{group} + \beta 2 * \left(RD_{group} - \overline{RD_{group}}\right)^2$$

For the group-level analysis, we test if the $\beta 2$ coefficients were significantly different from zero, which allows us to study the quadratic relationship. To study the relationship between these two models, we correlated the beta values for each regression between them. This correlation between coefficients across participants helped us identify consistent patterns in the direction of the quadratic effects.

Model matrix comparisons

To test specific hypotheses about how groups, individuals, and trial congruency are represented, we compared trial by trial correlation matrices for pattern dissimilarity data to model matrices (Fig 4A). Beta images for each trial, resulting from the preprocessing step in the *Beta series modeling* section, were masked with a 10 mm radius spherical mask on significant peaks identified in the RSA searchlight analysis and in the dmPFC ROI. The beta values underwent pairwise comparisons resulting in a trial by trial nRDM, for each participant and run. We computed three model matrices: (1) congruence model, (2) group model, and (3) individual model.

For the congruence model, transformation trials were classified as congruent when the egocentric reference frames (individuals) and allocentric reference frames (groups) belong to the same social category (i.e., trials involving city individual and city group). Conversely, trials were considered incongruent when the egocentric and allocentric reference frames belong to different categories (i.e., city individual and rural people group). To calculate congruence matrices, we hypothesize that trials within the same congruence condition will exhibit higher correlations with other trials sharing the same condition. To calculate the group model matrix and the individual model matrix, we hypothesized that trials involving the same social groups or individuals will contain more similar representations. Model matrices were constructed assigning a value of 0 to trials with a hypothesized high correlation (similarity), and a value of 2 to those with a hypothesized low correlation (dissimilarity) (Fig D in S1 Text related to Fig 4).

To evaluate model fit, we computed a partial correlation between participant average trial-wise pattern dissimilarity data and each of the 3 model matrices using the lower half of the matrix (excluding comparisons involving the same trials) after eliminating the effect of condition. At the group level, we computed one-way ANOVAs and a paired *T* test. These statistical analyses allowed us to compare fits across models.

In addition to the model matrix comparison RSA, we performed a whole-brain searchlight-based multiple regression RSA to examine the brain areas in which the change in dissimilarity between neural activity patterns in the trials is explained by the different three models. Subsequently, we performed an ANOVA to compare the difference between the coefficients for each model and a paired *t* test to determine which model was driving the observed effects.

fMRI statistical analysis

For the whole-brain results, the statistical analyses are performed using 10 mm radius spheres in Nilearn toolbox (v.0.10.2) [50] around the respective peak-voxel specified in the voxel-wise GLM analysis. We report peak-voxels in the hippocampus, dmPFC, and EC that survive small-volume correction for multiple comparisons ($p < 0.05$) based on bilateral hippocampal, dmPFC, and EC masks. We defined the bilateral hippocampal region using the Wake Forest University (WFU) PickAtlas, integrated with SPM12 [51,52]. The EC mask was created based on prior literature [8,53,54]. For the ROI analysis, we define the dmPFC mask [55] as the medial wall of the prefrontal cortex that forms part of Brodmann areas 9 and 10. Additionally, medial parts of Brodmann area 8 adjacent to area 9, as well as the dorsomedial portion of area 32, have been included. For all analyses outside the ROIs, we report activations surviving an uncorrected statistical threshold of $p < 0.001$ and correction for multiple comparisons at the whole-brain level (FWE $p < 0.05$) either at the peak-voxel or cluster-level. Coordinates of brain regions are reported in MNI space. Analysis was performed using a Python open-source RSA-toolbox v.3.0 (https://github.com/rsagroup/rsatoolbox, [56]) and statsmodels toolbox v. 0.14.0 (https://github.com/statsmodels/statsmodels, [57]).

Supporting information

S1 Text. Fig A in S1 Text. Related to Fig 2. Significant quadratic relationship between accuracy and RD_{group} ($t(19) = 3.39$, $p = 0.0015$).The underlying data for Fig A in S1 Text can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2 under "Data". **Fig B in S1 Text. Whole-brain searchlight analysis related to Fig 3**. (**A**) Whole-brain group effect of rating discrepancy group (RD_{group}) centered on left ventrolateral prefrontal cortex peak ($x = -54$; $y = 20$; $z = 2$; pFWE = .039). (**B**) Whole-brain group effect of group rating consistency (RC_{group}) centered on the left hippocampus peak ($x = -16$; $y = -34$; $z = -4$; pFWE = .038). For visualization purposes, the images are displayed at an uncorrected statistical threshold of $p < 0.005$. The underlying data for Fig B in S1 Text can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2. **Fig C in S1 Text. Whole-brain searchlight analysis for model matrices related to Fig 4**. (**A**) Left Thalamus peak ($x = -7$, $y = -16$, $z = 9$, pFWE < 0.001). (**B**) Ventral striatum peak ($x = 27$, $y = 18$, $z = 6$). For visualization purposes, the images are displayed at an uncorrected statistical threshold of $p < 0.001$.The underlying data for Fig C in S1 Text can be found in DOI: https://doi.org/10.6084/m9.figshare.28295732.v2. **Fig D in S1 Text. Explanation of Model-RDM Calculations. Related to Fig 4**. (**A**) Example trial types during transformation, each coded with specific color, involving a particular individual and group. (**B**) Entries of the matrix, where each cell of the matrix represents a comparison between two trials. (**C**) Each row represents a distinct model, while each column represents which pair of trials is coded as similar or dissimilar based on the model criteria. For the individual model, trials involving the same individual are considered similar, while trials involving different individuals are considered dissimilar. In the group model, trials involving the same group are

computed as similar, while those involving different groups are dissimilar. For the congruence model, trials with the same congruence condition (e.g., individual 1-group 1 versus individual 2-group 2) are assigned high similarity, while incongruent conditions are treated as dissimilar. Finally, in the condition model, only pairs of trials within the same condition are computed as similar, all other comparisons are dissimilar. For matrix entries, a 0 is assigned to comparisons with similar hypothesized representation and a 2 to those considered dissimilar across all models. **Table A in S1 Text.** 105 Everyday Scenarios related to Fig 1. **Table B in S1 Text.** Brain areas that were significantly modulated by group rating discrepancy related to Fig B in S1 Text. * = Small-volume corrected FWE for bilateral entorhinal cortex and bilateral hippocampus mask, respectively. **Table C in S1 Text.** Brain areas that were significantly modulated by differences in group rating consistency related to Fig B in S1 Text. * = Cluster-level correction for bilateral entorhinal cortex and bilateral hippocampus mask respectively. **Table D in S1 Text.** Brain areas that were significantly modulated by the difference between the three models (group, individual, and congruence) related to Fig C in S1 Text.
(PDF)

Acknowledgments

We thank María Baena Pérez and Esteban Villar Rodríguez for assisting with fMRI piloting, Maria Picó Pérez for fMRI preprocessing advice, and Ameer Ghouse for statistical and programming assistance. We thank Shahar Arzy and Maya Visser for feedback on a previous version of this manuscript. We thank General Electric for providing a pilot version of their multi-echo fMRI sequence package.

Author contributions

Conceptualization: Raphael Kaplan.

Data curation: Marta Rodríguez Aramendía.

Formal analysis: Marta Rodríguez Aramendía.

Funding acquisition: Raphael Kaplan.

Investigation: Marta Rodríguez Aramendía, Mariachiara Esposito, Raphael Kaplan.

Methodology: Raphael Kaplan.

Project administration: Raphael Kaplan.

Supervision: Raphael Kaplan.

Writing – original draft: Marta Rodríguez Aramendía, Raphael Kaplan.

Writing – review & editing: Raphael Kaplan.

References

1. Tolman EC. Cognitive maps in rats and men. Psychol Rev. n.d.;55(4):189–208. https://doi.org/10.1037/h0061626
2. O'Keefe J, Nadel L. The hippocampus as a cognitive map. Oxford University Press; 1978.
3. Schiller D, Eichenbaum H, Buffalo EA, Davachi L, Foster DJ, Leutgeb S, et al. Memory and space: towards an understanding of the cognitive map. J Neurosci. 2015;35(41):13904–11. https://doi.org/10.1523/JNEUROSCI.2618-15.2015 PMID: 26468191
4. Kaplan R, Schuck NW, Doeller CF. The role of mental maps in decision-making. Trends Neurosci. 2017;40(5):256–9. https://doi.org/10.1016/j.tins.2017.03.002 PMID: 28365032
5. Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, et al. What is a cognitive map? Organizing knowledge for flexible behavior. Neuron. 2018;100(2):490–509. https://doi.org/10.1016/j.neuron.2018.10.002 PMID: 30359611

6. Bellmund JLS, Gärdenfors P, Moser EI, Doeller CF. Navigating cognition: spatial codes for human thinking. Science. 2018;362(6415):eaat6766. https://doi.org/10.1126/science.aat6766 PMID: 30409861

7. Tavares RM, Mendelsohn A, Grossman Y, Williams CH, Shapiro M, Trope Y, et al. A map for social navigation in the human brain. Neuron. 2015;87(1):231–43. https://doi.org/10.1016/j.neuron.2015.06.011 PMID: 26139376

8. Kaplan R, Friston KJ. Entorhinal transformations in abstract frames of reference. PLoS Biol. 2019;17(5):e3000230. https://doi.org/10.1371/journal.pbio.3000230 PMID: 31048835

9. Park SA, Miller DS, Nili H, Ranganath C, Boorman ED. Map making: constructing, combining, and inferring on abstract cognitive maps. Neuron. 2020;107(6):1226-1238.e8. https://doi.org/10.1016/j.neuron.2020.06.030 PMID: 32702288

10. Park SA, Miller DS, Boorman ED. Inferences on a multidimensional social hierarchy use a grid-like code. Nat Neurosci. 2021;24(9):1292–301. https://doi.org/10.1038/s41593-021-00916-3 PMID: 34465915

11. Zhang L, Chen P, Schafer M, Zheng S, Chen L, Wang S, et al. A specific brain network for a social map in the human brain. Sci Rep. 2022;12(1):1773. https://doi.org/10.1038/s41598-022-05601-4 PMID: 35110581

12. Schafer M, Schiller D. Navigating social space. Neuron. 2018;100(2):476–89. https://doi.org/10.1016/j.neuron.2018.10.006 PMID: 30359610

13. Bicanski A, Burgess N. A neural-level model of spatial memory and imagery. Elife. 2018;7:e33752. https://doi.org/10.7554/eLife.33752 PMID: 30176988

14. Alexander AS, Robinson JC, Stern CE, Hasselmo ME. Gated transformations from egocentric to allocentric reference frames involving retrosplenial cortex, entorhinal cortex, and hippocampus. Hippocampus. 2023;33(5):465–87. https://doi.org/10.1002/hipo.23513 PMID: 36861201

15. Arzy S, Kaplan R. Transforming social perspectives with cognitive maps. Soc Cogn Affect Neurosci. 2022;17(10):939–55. https://doi.org/10.1093/scan/nsac017 PMID: 35257155

16. Tversky B. Cognitive maps, cognitive collages, and spatial mental models. Psychol Rev. 1993.

17. Tversky B. Mind in motion: how action shapes thought. Hachette UK; 2019.

18. Peer M, Hayman M, Tamir B, Arzy S. Brain coding of social network structure. J Neurosci. 2021;41(22):4897–909. https://doi.org/10.1523/JNEUROSCI.2641-20.2021 PMID: 33903220

19. Buzsáki G, Moser EI. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nat Neurosci. 2013;16(2):130–8. https://doi.org/10.1038/nn.3304 PMID: 23354386

20. Sui J, Humphreys G. The ubiquitous self: what the properties of self-bias tell us about the self. Ann New York Acad Sci. 2017;1396(1):222–35.

21. Todd AR, Tamir DI. Factors that amplify and attenuate egocentric mentalizing. Nat Rev Psychol. 2024.

22. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185(4157):1124–31. https://doi.org/10.1126/science.185.4157.1124 PMID: 17835457

23. Epley N, Keysar B, Van Boven L, Gilovich T. Perspective taking as egocentric anchoring and adjustment. J Pers Soc Psychol. 2004;87(3):327–39. https://doi.org/10.1037/0022-3514.87.3.327 PMID: 15382983

24. Epley N, Gilovich T. The anchoring-and-adjustment heuristic: why the adjustments are insufficient. Psychol Sci. 2006;17(4):311–8. https://doi.org/10.1111/j.1467-9280.2006.01704.x PMID: 16623688

25. Tamir DI, Mitchell JP. Anchoring and adjustment during social inferences. J Exp Psychol Gen. 2013;142(1):151–62. https://doi.org/10.1037/a0028232 PMID: 22506753

26. Wang YA, Simpson AJ, Todd AR. Egocentric anchoring-and-adjustment underlies social inferences about known others varying in similarity and familiarity. J Exp Psychol General. 2023;152(4):345–56.

27. Tamir DI, Mitchell JP. Neural correlates of anchoring-and-adjustment during mentalizing. Proc Natl Acad Sci U S A. 2010;107(24):10827–32. https://doi.org/10.1073/pnas.1003242107 PMID: 20534459

28. Boorman ED, Sweigart SC, Park SA. Cognitive maps and novel inferences: a flexibility hierarchy. Curr Opin Behav Sci. 2021;38:141–9. https://doi.org/10.1016/j.cobeha.2021.02.017

29. Viganò S, Bayramova R, Doeller CF, Bottini R. Mental search of concepts is supported by egocentric vector representations and restructured grid maps. Nat Commun. 2023;14(1):8132. https://doi.org/10.1038/s41467-023-43831-w PMID: 38065931

30. Zhao Z, Sened H, Tamir DI. Egocentric projection is a rational strategy for accurate emotion prediction. J Exp Soc Psychol. 2023;109:104521. https://doi.org/10.1016/j.jesp.2023.104521 PMID: 37663408

31. Tarantola T, Kumaran D, Dayan P, De Martino B. Prior preferences beneficially influence social and non-social learning. Nat Commun. 2017;8:817.

32. Sui J, Humphreys GW. The integrative self: how self-reference integrates perception and memory. Trends Cogn Sci. 2015;19(12):719–28. https://doi.org/10.1016/j.tics.2015.08.015 PMID: 26447060

33. Kumaran D, Melo HL, Duzel E. The emergence and representation of knowledge about social and nonsocial hierarchies. Neuron. 2012;76(3):653–66. https://doi.org/10.1016/j.neuron.2012.09.035 PMID: 23141075

34. Kumaran D, Banino A, Blundell C, Hassabis D, Dayan P. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. Neuron. 2016;92(5):1135–47. https://doi.org/10.1016/j.neuron.2016.10.052 PMID: 27930904

35. Peer M, Brunec IK, Newcombe NS, Epstein RA. Structuring knowledge with cognitive maps and cognitive graphs. Trends Cogn Sci. 2021;25(1):37–54. https://doi.org/10.1016/j.tics.2020.10.004 PMID: 33248898

36. Du M, Basyouni R, Parkinson C. How does the brain navigate knowledge of social relations? Testing for shared neural mechanisms for shifting attention in space and social knowledge. Neuroimage. 2021;235:118019. https://doi.org/10.1016/j.neuroimage.2021.118019 PMID: 33789132

37. Son J-Y, Bhandari A, FeldmanHall O. Cognitive maps of social features enable flexible inference in social networks. Proc Natl Acad Sci U S A. 2021;118(39):e2021699118. https://doi.org/10.1073/pnas.2021699118 PMID: 34518372

38. Son J-Y, Bhandari A, FeldmanHall O. Abstract cognitive maps of social network structure aid adaptive inference. Proc Natl Acad Sci U S A. 2023;120(47):e2310801120. https://doi.org/10.1073/pnas.2310801120 PMID: 37963254

39. Tversky B. Spatial thought, social thought. In: Spatial dimensions of social thought. Vol. 18, Berlin, Boston: De Gruyter Mouton; 2011, 17–38.

40. Triandis HC, Leung K, Villareal MJ, Clack FI. Allocentric versus idiocentric tendencies: convergent and discriminant validation. J Res Pers. 1985;19(4):395–415. https://doi.org/10.1016/0092-6566(85)90008-x

41. Amodio DM, Frith CD. Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci. 2006;7(4):268–77. https://doi.org/10.1038/nrn1884 PMID: 16552413

42. Kolling N, O'Reilly JX. State-change decisions and dorsomedial prefrontal cortex: the importance of time. Curr Opin Behav Sci. 2018;22:152–60. https://doi.org/10.1016/j.cobeha.2018.06.017 PMID: 30123818

43. Stendardi D, Giordani LG, Gambino S, Kaplan R, Ciaramelli E. Who am I really? The ephemerality of the self-schema following vmPFC damage. Neuropsychologia. 2023;188:108651. https://doi.org/10.1016/j.neuropsychologia.2023.108651 PMID: 37481034

44. Harald Baayen R, Milin P. Analyzing reaction times. Int J Psychol Res. 2010;3(2):12–28. https://doi.org/10.21500/20112084.807

45. Weiskopf N, Hutton C, Josephs O, Deichmann R. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3 T and 1.5 T. Neuroimage. 2006;33(2):493–504. https://doi.org/10.1016/j.neuroimage.2006.07.029 PMID: 16959495

46. Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, et al. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage. 2013;64:240–56. https://doi.org/10.1016/j.neuroimage.2012.08.052 PMID: 22926292

47. Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced?. Neuroimage. 2009;44(3):893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036 PMID: 18976716

48. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2008;2:4. https://doi.org/10.3389/neuro.06.004.2008 PMID: 19104670

49. Muhle-Karbe PS, Sheahan H, Pezzulo G, Spiers HJ, Chien S, Schuck NW, et al. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. Neuron. 2023;111(23):3885-3899.e6. https://doi.org/10.1016/j.neuron.2023.08.021 PMID: 37725981

50. Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform. 2011;5:13. https://doi.org/10.3389/fninf.2011.00013 PMID: 21897815

51. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage. 2003;19(3):1233–9. https://doi.org/10.1016/s1053-8119(03)00169-1 PMID: 12880848

52. Maldjian JA, Laurienti PJ, Burdette JB, Kraft RA. An automated method for neuroanatomic precentral gyrus discrepancy in electronic versions of the Talairach Atlas. Neuroimage. 2004;21(1):450–5. https://doi.org/10.1016/j.neuroimage.2003.09.032 PMID: 14741682

53. Chadwick MJ, Jolly AEJ, Amos DP, Hassabis D, Spiers HJ. A goal direction signal in the human entorhinal/subicular region. Curr Biol. 2015;25(1):87–92. https://doi.org/10.1016/j.cub.2014.11.001 PMID: 25532898

54. Garvert MM, Dolan RJ, Behrens TE. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife. 2017;6:e17086. https://doi.org/10.7554/eLife.17086 PMID: 28448253

55. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. Neuroimage. 2012;62(2):782–90. https://doi.org/10.1016/j.neuroimage.2011.09.015 PMID: 21979382

56. Schütt H. Representational similarity analysis 3.0, version swh:1:rev:01e767c432e-77633fe31304201718afce6a6ff9c. Software Heritage; 2023.

57. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference. 2010;57(61):10–25080.