
Supplementary information

Abstract representations emerge in human hippocampal neurons during inference

In the format provided by the
authors and unedited

Supplementary Tables

Patient ID	Age	Sex	Patient Behavior	Session ID	Session Behavior	HPC Neurons	dACC Neurons	VTC Neurons	AMY Neurons	preSMA Neurons	vmPFC Neurons
P61CS	52	F	N/A	1	X	N/A	N/A	N/A	N/A	N/A	N/A
				2	IP	15	16	0	3	16	7
P62CS	25	F	Post	1	IA	7	1	3	24	4	13
				2	IP	7	1	3	24	4	13
				3	IA	11	2	2	38	10	6
				4	IP	4	0	4	22	4	5
P63CS	48	F	NE	1	IA	34	8	0	31	16	8
			Post	2	IA	29	3	0	32	8	6
				3	IP	33	4	0	39	9	8
P65CS	55	F	Pre	1	IP	19	0	0	35	0	9
				2	IP	19	0	0	35	0	9
				3	IA	7	0	0	27	0	5
P67CS	38	F	Post	1	IA	7	7	0	46	6	58
				2	IP	7	7	0	46	6	58
				3	IP	7	3	0	62	3	44
				4	IP	7	3	0	62	3	44
P69CS	41	F	N/A	1	X	N/A	N/A	N/A	N/A	N/A	N/A
				2	X	N/A	N/A	N/A	N/A	N/A	N/A
P70CS	30	F	N/A	1	X	N/A	N/A	N/A	N/A	N/A	N/A
				2	IP	2	0	0	12	8	1
P71CS	40	M	NE	1	IA	0	36	0	3	8	18
				2	IA	0	36	0	3	8	18
P73CS	58	F	NE	1	X	N/A	N/A	N/A	N/A	N/A	N/A
				2	IA	5	7	31	32	15	13
				3	IP	9	5	27	29	24	11
				4	IP	9	5	27	29	24	11
P74CS	23	M	Post	1	IA	0	0	0	8	15	0
				2	IP	0	0	0	8	15	0
P76CS	24	F	N/A	1	IP	28	34	8	14	10	11
				2	X	N/A	N/A	N/A	N/A	N/A	N/A
P78CS	54	F	Post	1	IA	32	0	11	22	0	0
				2	IP	32	0	11	22	0	0
P79CS	42	F	Pre	1	IP	50	15	36	57	21	30
				2	IP	50	15	36	57	21	30
				3	IP	22	26	38	36	14	23
TWH162	27	F	N/A	1	IP	2	0	0	0	0	0
TWH163	22	F	N/A	1	IP	0	37	0	0	3	1
				2	IP	0	37	0	0	3	1
TWH165	32	M	NE	1	IA	10	0	0	14	0	0
				2	IA	10	0	0	14	0	0
TWH172	37	F	Pre	1	IP	12	0	14	0	0	0
				2	IP	12	0	14	0	0	0

Table S1. Patients, Sessions, Behavior, and Neurons. Summary of patient information, the number of sessions performed, the behavioral classification at the patient and session level, and the number of recorded neurons per region per session. Patient behavior is defined with respect to instances of high-level verbal instructions (see Fig. 5), where: Pre – “pre-instruction inference achieved”, NE – “Inference not exhibited”, post – “post-instruction inference achieved”, and N/A – “did not qualify for analysis”. Session behavior is defined with respect to performance on the first available inference trial, where: IA – “inference absent”, IP – “inference present”, X – “at or below chance non-inference performance”.

Balanced Dichotomies			Task Conditions					
#	Class 1 Conditions	Class 2 Conditions	Name	#	Stimulus Identity	Response	Outcome	Context
1	1,2,3,4	5,6,7,8	Context	1	C	Left	5¢	1
2	1,2,3,5	4,6,7,8	N/A	2	D	Right	5¢	1
3	1,2,3,6	4,5,7,8	N/A	3	A	Left	25¢	1
4	1,2,3,7	4,5,6,8	N/A	4	B	Right	25¢	1
5	1,2,3,8	4,5,6,7	N/A	5	D	Left	5¢	2
6	1,2,4,5	3,6,7,8	N/A	6	A	Right	5¢	2
7	1,2,4,6	3,5,7,8	N/A	7	B	Left	25¢	2
8	1,2,4,7	3,5,6,8	N/A	8	C	Right	25¢	2
9	1,2,4,8	3,5,6,7	N/A					
10	1,2,5,6	3,4,7,8	Outcome					
11	1,2,5,7	3,4,6,8	N/A					
12	1,2,5,8	3,4,6,7	AB vs CD					
13	1,2,6,7	3,4,5,8	N/A					
14	1,2,6,8	3,4,5,7	N/A					
15	1,2,7,8	3,4,5,6	N/A					
16	1,3,4,5	2,6,7,8	N/A					
17	1,3,4,6	2,5,7,8	N/A					
18	1,3,4,7	2,5,6,8	N/A					
19	1,3,4,8	2,5,6,7	N/A					
20	1,3,5,6	2,4,7,8	N/A					
21	1,3,5,7	2,4,6,8	Response					
22	1,3,5,8	2,4,6,7	N/A					
23	1,3,6,7	2,4,5,8	N/A					
24	1,3,6,8	2,4,5,7	AC vs BD (Stim Pair)					
25	1,3,7,8	2,4,5,6	N/A					
26	1,4,5,6	2,3,7,8	N/A					
27	1,4,5,7	2,3,6,8	N/A					
28	1,4,5,8	2,3,6,7	N/A					
29	1,4,6,7	2,3,5,8	Parity					
30	1,4,6,8	2,3,5,7	N/A					
31	1,4,7,8	2,3,5,6	AD vs BC					
32	1,5,6,7	2,3,4,8	N/A					
33	1,5,6,8	2,3,4,7	N/A					
34	1,5,7,8	2,3,4,6	N/A					
35	1,6,7,8	2,3,4,5	N/A					

Table S2. Definition of all balanced dichotomies. Class assignment and name of all 35 balanced dichotomies used in geometric balanced dichotomy analysis. Dichotomies where the name is “N/A” do not have a clear interpretation with respect to task construction. Identity of the task conditions participating in the balanced dichotomies is shown to the right.

Selected Stimuli	Training Dichotomy	Test Dichotomy	Partial Positive	Partial Negative	Full Positive	Full Negative
A,B	A1,B1	A2,B2		Response		
	A2,B2	A1,B1		Response		
A,C	A1,C1	A2,C2				Value
	A2,C2	A1,C1				Value
A,D	A1,D1	A2,D2		Response		
	A2,D2	A1,D1		Response		
B,C	B1,C1	B2,C2		Response		
	B2,C2	B1,C1		Response		
B,D	B1,D1	B2,D2			Value	
	B2,D2	B1,D1			Value	
C,D	C1,D1	C2,D2		Response		
	C2,D2	C1,D1		Response		

Table S3. Definition of all stimulus dichotomies. Task condition assignment for stimulus dichotomies. These dichotomies are used in Fig. 3b,c,e,f and associated supplements whenever there is a reference to “Stimulus CCGP” or “Stimulus Parallelism Score”. Partial and full correlations with other task variables are noted for each stimulus dichotomy.

Training Dichotomy	Test Dichotomy	Partial Positive	Partial Negative	Full Positive	Full Negative
A1,A2	B1,B2		Response		
B1,B2	A1,A2		Response		
A1,A2	C1,C2	Response	Outcome		
C1,C2	A1,A2	Response	Outcome		
A1,A2	D1,D2	Response			
D1,D2	A1,A2	Response			
B1,B2	C1,C2		Response		
C1,C2	B1,B2		Response		
B1,B2	D1,D2	Response			
D1,D2	B1,B2	Response			
C1,C2	D1,D2		Response		
D1,D2	C1,C2		Response		

Table S4. Definition of all context dichotomies. Task condition assignment for context dichotomies. These dichotomies are used in Fig. 3g,h and associated supplements whenever there is a reference to “Context CCGP” or “Context Parallelism Score”. Partial and full correlations with other task variables are noted for each context dichotomy.

Supplementary Videos

Supplementary Video 1. Transformation of hippocampal geometry shown with MDS of real data. Visualization of the transformation of the representational geometry in the hippocampus shown using MDS of condition-averaged responses for all recorded hippocampal

neurons during the stimulus period. Plotting conventions and data are identical to those used in Fig. 3i. Here, we use linear interpolation between the starting and ending geometry shown in the video, which correspond to inference absent and inference present sessions respectively. The video is meant to provide intuition for how the task conditions are represented differently in neural state space in the presence and absence of inference behavior.

Supplementary notes

S.1 Hippocampus encodes context abstractly at baseline during inference sessions.

We analyzed the baseline period preceding stimulus onset (Fig. 2, inset) to study the geometry of task representations that might persist from the previous trial (all labels were defined by the prior just completed trial for this analysis). Context was encoded as an abstract variable in the HPC in inference present and not in inference absent sessions. Unlike the stim epoch, context was the only named dichotomy to emerge as significantly decodable at baseline in the HPC for inference present sessions (Fig. 2h, red, inference absent vs. present, $p_{RS} = 1.1 \times 10^{-33}$, $p_{Absent} = 0.35$, $p_{Present} = 6.4 \times 10^{-5}$, E3l), indicating that any task condition information from the previous trial other than the context (i.e. outcome, response) was not significantly decodable from the population during this epoch. Nevertheless, the shattering dimensionality significantly increased in inference present compared to inference absent sessions (0.51 vs 0.53, $p_{RankSum} = 0.0079$), which is attributable to the rise in decodability of context and context correlated dichotomies. Context during baseline was also encoded abstractly as shown by the increase in CCGP (Fig. 2i, Extended Data Fig. 3m, red, Inference absent vs present, $p_{RS} = 2.4 \times 10^{-34}$, $p_{Absent} = 0.45$, $p_{Present} = 7.7 \times 10^{-6}$) and Parallelism Score (Extended Data Fig. 3n,o, red, $p_{Absent} = 0.37$, $p_{Present} = 1.2 \times 10^{-10}$). These results indicate that an abstract representation of context that persisted from the previous trial also emerged in the hippocampus in sessions where patients performed inference.

S.2 Hippocampal context representation does not rely on classically tuned neurons

To determine whether the geometric findings presented here arose as a consequence of strong classical (univariate linear) tuning, two ablation analyses were conducted in which all neurons that had significant univariate linear tuning to at least one task variable were excluded and all geometric measures were re-computed. In the first analysis, neurons tuned to one or several of response, reward, or context (3-Way ANOVA, any main effect $p < 0.01$) were excluded. Of the 494 neurons in HPC, 41 were excluded for exhibiting classic tuning prior to recomputing all measures. Qualitatively, findings remained largely unchanged by the exclusion of this population of specialized neurons (Extended Data Fig. 4a-e). The Parity dichotomy was no longer decodable above chance in inference present sessions (Extended Data Fig. 4a, orange, $p_{Present} = 0.056$), but the rise in Shattering Dimensionality (average over dichotomies) was still present (Extended Data Fig. 4a, $p_{RS} = 0.0012$, Ranksum over dichotomies). A similar analysis was also conducted for the baseline representation of context, and again the findings qualitatively remained unchanged (Extended Data Fig. 4f-j). Note that, although the decodability of context at baseline during inference present sessions did decrease in significance when conducting the error trial

analysis (Extended Data Fig. 4i, red, $p_{Present} = 0.12$), it was still significantly increased compared to inference absent sessions ($p_{RS} = 5.7 \times 10^{-19}$) and was significantly reduced in error trials ($p_{Present} = 0.37$, inference present correct vs error, $p_{RS} = 0.0002$). In the second analysis, neurons tuned to stimulus identity or context (2-Way ANOVA, any main effect $p < 0.01$) were excluded. In this analysis, 82 neurons were excluded, leading to a loss of decodability of stimulus dichotomies due to the removal of visually-selective neurons as expected (Extended Data Fig. 4k-m). Context, however, remained decodable and present in an abstract format. Together, these analyses indicate that the abstract context representation in the hippocampus is a highly distributed variable whose geometry is not a simple consequence of strong univariate tuning of a small population of “context” neurons.

S.3 Hippocampal context and stimulus representations are not driven by SOZ neurons

To ensure that the reported findings were not influenced by pathology, in a further control analysis we excluded all HPC neurons (86/494) recorded from electrodes that were clinically identified to reside in medial temporal seizure onset zones (SOZs). Repeating our analyses on the neurons that remain revealed that results were qualitatively unchanged (Extended Data Fig. 4n-s), though the parallelism for context was now significantly above chance during inference absent sessions for both the stimulus (Extended Data Fig. 4p, red, $p_{Absent} = 3.5 \times 10^{-7}$) and baseline (Extended Data Fig. 4s, red, $p_{Absent} = 0.0022$) periods. These findings suggest that neurons residing in hippocampal tissue with high disease burden do not meaningfully contribute to the task representation⁷¹, thereby leading to marginally increased representation strength once these neurons are removed (note that the decodability and CCGP for context in inference absent sessions are still not significantly different from chance, confirming our finding).

S.4 Hippocampal representations are not driven by differences in non-inference trial performance

To ensure that the reported geometric findings in the hippocampus did not arise due to non-inference performance differences between inference absent and inference present sessions, an additional control analysis was performed where the non-inference trial performance of included sessions was distribution-matched between sessions where inference was absent and present. Pairs of inference absent and inference present sessions with at most 7.5% difference in non-inference trial performance were selected, prioritizing sessions with more hippocampal neurons. This matching process yielded 10 inference absent sessions (152 HPC neurons) and 10 inference present sessions (187 HPC neurons) whose average non-inference performances did not statistically significantly differ (92.8% v.s. 94.7%, $p_{RS} = 0.58$, RankSum over sessions, Extended Data Fig. 4z). All main geometric findings were recapitulated using this session split during both the stimulus (Extended Data Fig. 4t-v) and baseline (Extended Data Fig. 4w-y) periods, indicating that the learning-dependent changes in the hippocampal representational geometry we observe cannot be explained by differences in non-inference trial performance.

S.5 Changes in hippocampal single-neuron tuning explain changes in representation with inference

To determine if changes in univariate tuning could partially explain the observed changes in the hippocampal representation from inference absent to inference present sessions, average tuning properties of single neurons were computed. Both the number of significant factors (main effects or interactions) per neuron (3-way ANOVA – Response, Context, Outcome, $p < 0.05$) and the tuning strength of neurons for those significant factors indexed by the F-statistic of those ANOVA factors were significantly elevated in inference present compared to inference absent (Extended Data Fig. 9d, $p_{RS} = 0.010$, E9f, $p_{RS} = 0.0089$). These effects were only observed in HPC, with either no change or a decrease in tuning observed in the other areas except for a significant increase in vmPFC (Extended Data Fig. 9e). We also separately considered the linear and non-linear (interaction) terms for the 4 (Stimulus) x 2 (Context) ANOVA and found that, while the fraction of hippocampal neurons exhibiting significant ($p < 0.05$, Main Effect) stimulus identity tuning significantly decreased from absent to present inference sessions (Extended Data Fig. 9i, 21.3% vs. 19.1% of neurons, $p = 0.002$), the fraction of neurons exhibiting significant context tuning increased (Fig. SE9i, 9.5% vs. 15.7% of neurons, $p = 2.2 \times 10^{-6}$), which could also partially explain the increase in centroid distance and resultant context decodability. Such “context neurons”, however, cannot account for this effect fully as their removal from the pseudopopulation does not qualitatively alter the hippocampal population geometry (Extended Data Fig. 4). As a control, this analysis was also performed for VTC neurons, in which we did not find a significant difference in the percentage of neurons with univariate context tuning (Extended Data Fig. 9j).

S.6 Context representations outside of the hippocampus

The dorsal Anterior Cingulate Cortex (dACC) was the only other region to exhibit significant changes in its latent context representation as a function of patients’ ability to perform inference. However, this context representation was limited to the baseline period, since geometric analysis for the stimulus period revealed an absence of significant context decodability and CCGP in inference absent and inference present sessions (Fig. 2g, E3a). However, the Parallelism Score for context in the dACC did increase significantly (Extended Data Fig. 3b, red, inference absent vs. inference present, $p_{RS} = 2.4 \times 10^{-19}$, $p_{Absent} = 0.99$, $p_{Present} = 3.8 \times 10^{-12}$), reflecting an increase in context parallelism in condition averages that was not detectable with the other metrics (which are based on single trial decoding).

During the baseline, however, context also emerged as the only decodable dichotomy in inference present sessions in the dACC (Extended Data Fig. 3k, red, $p_{Absent} = 0.37$, $p_{Present} = 0.049$). We found this context variable was also represented in an abstract format, emerging as the dichotomy with the highest CCGP (Extended Data Fig. 3p,s, red, $p_{Absent} = 0.26$, $p_{Present} = 0.018$) and Parallelism Score (Extended Data Fig. 3q,t, red, $p_{Absent} = 0.18$, $p_{Present} = 0.013$) in inference present sessions, while being at chance for both metrics in inference absent sessions.

As in HPC, the lack of decodability for previous trial outcome and response suggests that any variables encoded at earlier epochs during the previous trial (e.g. post-reply delay or outcome) were extinguished from the population by the onset time of the current baseline epoch.

Analysis of dichotomy centroid distances during the baseline period revealed context also emerged as the dichotomy with the greatest separation in the dACC during inference present session trials (3.9 vs. 5.7 Hz, $p_{Absent} = 0.48$, $p_{Present} = 0.0065$, $p_{\Delta Dist} = 0.046$, Extended Data Fig. 9m,n). However, this implementation of the context representation was achieved through significant increases in condition-wise firing rates (Extended Data Fig. 9r, dACC inference absent vs inference present, $p_{RS} = 0.049$) as opposed to the firing rate decrease observed in the hippocampus. Together, these analyses indicate that a weak, but nonetheless significant, representation of latent context encoded in an abstract format also emerged in the dACC, and that this variable was accommodated in the representation through a different implementational strategy.

Supplementary results

Behavior (Fig. 1)

Block-wise estimates of task performance for inference absent (Extended Data Fig. 1e) and inference present (Extended Data Fig. 1f) sessions reveal that during inference absent sessions, patients exhibit poor inference performance after every context switch throughout the task, although the performance at the end of every block is high. In contrast, during inference present sessions, inference performance rapidly rose over the first few blocks and remained high throughout the duration of the session (Extended Data Fig. 1f). Note that within a given session, the two latent contexts had identical stimuli, responses, and outcomes; the only difference was which stimulus was associated with which response and outcome. Correspondingly, subject-level accuracies (Extended Data Fig. 1c) and reaction times (Extended Data Fig. 1d) for the two contexts (arbitrarily labeled 1 and 2 across sessions) were not significantly different, indicating that there was no systematic performance bias for one of the two contexts.

Single neuron analysis (Fig. 1)

When neurons were separated into those recorded from inference absent and inference present sessions, 5-15% of neurons were significantly tuned for each of the main and interaction effects of the 3-Way ANOVA, with significant reductions in the proportion of neurons tuned to Outcome, Response x Context, and Response x Context x Outcome ($p_O = 0.0007$, $p_{RxC} = 0.0395$, $p_{RxCxO} = 0.0048$, two sample z-test) in inference present sessions compared to inference absent sessions (Extended Data Fig. 1q). Similar analyses conducted on a separate 2-Way ANOVA for Stimulus Identity and Context (Fig. 1l, SC column, Extended Data Fig. 1r, $p_S = 0.0165$, two sample z-test comparing inference absent with inference present sessions, Extended Data Fig. 1u shows each brain area separately), and for Stimulus Identity and Response (Fig. 1l, SR column, Extended Data Fig. 1s, $p_S = 0.0287$), revealed a significant decrease in the fraction of neurons tuned to Stimulus Identity in inference present compared to inference absent

sessions, again with no significant changes in the proportion of neurons coding for Context, Response, or interactions.

Decodability and format of variables in areas other than hippocampus (Fig. 2)

In vmPFC, stim pair (Fig. 2g, purple, inference absent vs. present, $p_{RS} = 6.9 \times 10^{-20}$, $p_{Absent} = 0.21$, $p_{Present} = 0.0097$) and in preSMA, response (Fig. 2g, green, inference absent vs. present, $p_{RS} = 5.2 \times 10^{-13}$, $p_{Absent} = 0.091$, $p_{Present} = 0.0057$) increased significantly in decodability in inference present compared to inference absent sessions, as expected (see discussion). The stim pair variable was also significantly decodable in AMY both during inference absent and present sessions, with no significant difference between the two (Fig. 2g, purple, inference absent vs. present, $p_{RS} = 0.88$, $p_{Absent} = 0.0016$, $p_{Present} = 0.0018$). Elevated CCGP was also observed for the stim pair variable in vmPFC during inference present and not inference absent sessions (Extended Data Fig. 3a, purple, $p_{Absent} = 0.45$, $p_{Present} = 0.014$), the response variable in preSMA (Extended Data Fig. 3a, green, $p_{Absent} = 0.050$, $p_{Present} = 0.0010$), and the stim pair variable in AMY during both inference absent and inference present sessions (Extended Data Fig. 3a, purple, $p_{Absent} = 0.050$, $p_{Present} = 0.039$).

Additional analysis on format of hippocampal representations (Fig. 2)

In addition to CCGP (see main text), we also confirmed the abstract format of hippocampal representations using the PS. Following stimulus onset, the PS was significantly larger than expected by chance for the variables context and stim pair in the inference present but not inference absent sessions (Extended Data Fig. 3g, red, $p_{Absent} = 0.55$, $p_{Present} = 1.4 \times 10^{-15}$ and Extended Data Fig. 3g, purple, $p_{Absent} = 0.17$, $p_{Present} = 1.7 \times 10^{-8}$).

Error trial analysis (Extended Data Fig. 3)

Shattering dimensionality was significantly higher in correct trials in the stimulus period (Extended Data Fig. 3e, inference present vs. inference present (error), 0.59 vs 0.54, $p_{RS} = 0.0048$), as was decodability of the parity dichotomy (Extended Data Fig. 3e, inference present vs. inference present (error), orange, $p_{RS} = 0.029$). Furthermore, the dichotomies context and stimulus pairing were significantly more decodable in correct compared to incorrect trials (Extended Data Fig. 3e, inference present vs. inference present (error), red, $p_{RS} = 1.2 \times 10^{-20}$, purple, $p_{RS} = 1.0 \times 10^{-9}$). Furthermore, Parallelism Score for context but not other variables was significantly elevated in correct trials but not in incorrect trials (Extended Data Fig. 3f, red, inference present vs. inference present (error), $p_{Present} = 1.1 \times 10^{-16}$, $p_{Present (error)} = 0.083$) in inference present sessions. Similarly, the baseline representation of context also showed this effect, being decodable during correct trials but not during incorrect trials (Extended Data Fig. 3i, red, inference present vs. inference present (error), $p_{RS} = 3.5 \times 10^{-12}$, $p_{Present} = 0.012$, $p_{Present (error)} = 0.47$). The baseline shattering dimensionality did not significantly differ between correct and error trials (0.52 vs 0.51, $p_{RS} = 0.31$). Context was present in an abstract format only in correct trials based on the Parallelism Score for context being significantly larger than chance during correct but not incorrect trials

(Extended Data Fig. 3j, red, inference present correct vs. error $p_{Present} = 0, p_{Present (error)} = 0.94$).

Across-session generalization (Fig.2)

We examined whether the geometry of the context representation was preserved such that context decoding could generalize across different inference present sessions. To do so, we aligned the geometries in the two sessions to each other in neural state space using a subset of task conditions and then examined whether decoding context generalized from one session to the other on held-out conditions (see methods). This analysis revealed high context parallelism between random subsets of different inference sessions during both the baseline and stimulus periods (Extended Data Fig. 3z, aa). Such cross-session context parallelism was not found when performing the same analysis for the inference absent sessions (Extended Data Fig. 3ab, ac). This indicates that the geometry of the hippocampal context representation generalizes across inference sessions.

Controls for univariately tuned neurons, seizure-onset zones, and non-inference performance

We performed three sets of control analyses. First, to determine the relationship of our population-level findings to classical (univariate) tuning, we repeated our analyses after removing subsets of neurons from the population. We removed neurons with significant main effects in a (i) a 2x2x2 ANOVA for Response, Context, and Outcome, and (ii) a 4x2 ANOVA for Stimulus Identity and Context (Supplement S.2). In both analyses, context remained significantly decodable and in an abstract format. Also, as expected, in (ii), stimulus pair representations were no longer decodable. These control analyses indicate that the abstract representation of context in the hippocampus did not arise only due to the emergence of classically context tuned neurons. Rather, context was represented by broadly distributed context modulation at the level of the population.

Second, to assess whether our results were influenced by pathology, we repeated our analysis after excluding hippocampal neurons that were located within clinically confirmed seizure onset zones (Supplement S.3). We found no quantitative changes in our results, suggesting that hippocampal pathology did not influence our results.

Lastly, we examined whether our results were sensitive to behavioral accuracy in non-inference trials (Fig. 1d, trial indicated as ‘last’). We repeated our analysis in a subset of inference absent and inference present sessions that were chosen such that non-inference trial performance was matched (Supplement S.4). This control analysis revealed no qualitative changes in our results, suggesting that differences in non-inference trial performance cannot explain our results.

Context representations outside of the Hippocampus (Fig. 2)

The only other area of the brain that we examined in which we found a representation of latent context that correlated with inference behavior was in the dACC, but only during the baseline and not the stimulus period (Supplement S.6). Interestingly, context emerged in the task representation through a different implementation strategy, namely an increase in firing rates rather than a decrease as in the hippocampus.

Further controls for firing rate properties underlying geometry changes in hippocampus (Fig. 4).

Further analysis revealed a small number of inference absent sessions with high firing rate that biased the mean inference absent firing rate. Repeated geometric analysis after removing these sessions (Extended Data Fig. 9x) revealed a more modest 32% firing rate difference (1.67 ± 0.05 to 1.13 ± 0.03 Hz, Extended Data Fig. 9y). Excluding these sessions did not alter the geometric results (Extended Data Fig. 9z-ab).

Though the increase in distances between dichotomy centroids for context appears to be a distributed, population-level phenomenon (see Extended Data Fig. 4a-e), we sought to determine if a signature of this increase could be detected in the tuning of individual neurons (Supplement S.5). We found that the proportion of neurons exhibiting univariate context tuning increased from inference absent to inference present sessions, thus partially explaining the increased representational distance. However, the hippocampal population geometry did not exclusively rely on these neurons because after excluding all neurons with significant univariate coding, we found no qualitative change in the population geometry (Extended Data Fig. 4). Furthermore, the reduction in hippocampal firing rate from inference absent to present sessions did not bias any geometric measure determined through a firing rate distribution-matched control analysis (Extended Data Fig. 9s-w).

Further analysis for verbal instructions (Fig. 5)

We also examined firing rate changes of hippocampal neurons separately for the post-instruction inference, pre-instruction inference, and inference not-exhibited groups (Extended Data Fig. 10v). This analysis revealed significant reductions in firing rate across all conditions from Session One to Session Two for the post-instruction inference patients alone (-0.39 ± 0.15 Hz, $p_{\text{Post-Instruction}} = 1.4 \times 10^{-4}$), confirming that reductions in hippocampal firing rate in the same neurons recorded across adjacent sessions were associated with increases in inference performance. With a cell-by-cell comparison, 22/46 neurons (48%) show a significant decrease in firing rate with an average reduction of 1.3 Hz. Hippocampal neurons from the inference not-exhibited group showed an increase in firing rate (0.10 ± 0.04 Hz, $p_{\text{Not-Exhibited}} = 1.2 \times 10^{-4}$) and pre-instruction inference group firing rates did not significantly change (0.06 ± 0.08 Hz, $p_{\text{Pre-Instruction}} = 0.08$).

Supplementary discussion

The focus of our study was to examine how representations of context, stimulus identity, response, and predicted outcome change as a function of learning. In a prior study in macaques¹¹, the representation of the same variables in two very well trained animals was examined in HPC, dlPFC, and ACC in a similar task after the completion of training. Several notable differences exist between the two studies. First, context was encoded in an abstract format at baseline and was decodable after stimulus onset in all three brain areas examined in the macaques. In contrast, in humans, context is only strongly decodable in the HPC. We hypothesize that the wide-spread encoding of context in the macaque study was due to the extensive training the animals received

before recordings commenced. In contrast, our patients had no prior task experience. It is possible that early on during learning, latent context representations are present only in the hippocampus and are propagated to the cortex (dACC) with extensive task experience. This hypothesis is supported by prominent direct and indirect projections from the hippocampus to dACC in primates^{72–74}, and flexible, context-dependent interactions between medial frontal cortical neurons and hippocampal outputs^{59,75}. Second, the human hippocampus exhibited abstract stimulus representations, unlike the abstract response or “choice” representation in the macaques (in the interval during the presentation of the stimulus). Notably, the abstract stimulus pair and response dichotomies are constructed such that high CCGP for one will necessarily lead to below-chance CCGP for the other, which was indeed the case for both our study (high stim pair, low response) and the primate study (low stim pair, high response). One potential reason for these differences is a species difference: human HPC neurons are strongly modulated by the identity and semantic category of presented images^{28,59,63,76,77}, making it natural to organize representations of context relative to this existing representation. Similarly, representations of choices are not prominent in the human HPC⁵⁹. Another potential reason is a difference in task construction: our task employed semantically identifiable images, whereas the prior experiment with macaques used fractals. Third, unlike macaque HPC, human HPC did not encode predicted outcome. We note that in our task, outcome prediction was not necessary to perform the task because context switches were signaled by the accuracy of the response (correct or incorrect), which was independent of predicted outcome received for a correct response. Furthermore, all possible task-states were uniquely indexable using stimulus identity and context, rendering outcome prediction representation unnecessary for unambiguously defining the current task state. Finally, another possibility is that the reward for the macaques (juice volume) was more motivationally salient than the small monetary reward (25¢ or 5¢) patients received. We hypothesize that these reasons obviated the need for a predictive representation of outcome to complete the task in our patients. It remains an important question whether representations similar to those seen in macaques emerge in the other brain areas we examined following extensive training. Our data indicates that on short experiential timescales, the human hippocampus generates a representation that encodes the minimum set of variables required to solve the task.

Apart from the hippocampus, abstract representations also emerged in two other brain areas we studied: stim-pair and response representations emerging in the vmPFC and preSMA, respectively (Fig. 2g, E3a). While interesting in their own right, these variables were the only encoded variables in each respective region, thus preventing us from studying the geometry of multiple simultaneously abstract variables in these two areas. The hippocampus was also not unique in its representation of context. A weaker representation of context was also found in the dACC, but only during the baseline period. This finding aligns with work implicating the dACC in the representation of task rules and task sets^{78–83}. Following stimulus onset, however, dACC did not contain a representation of latent context (Fig. 2g). In contrast, previous studies in tasks with explicitly cued context switches^{59,60,66,84} find that neurons in the medial frontal cortex (dACC and preSMA) are tuned to task context following stimulus onset. We hypothesize that this might be due to differences in task demands: context switches were uncued in our task and had to be inferred from outcomes. It remains an open question to examine whether cued vs.

inferred context switches engage different mechanisms of switching between contexts and/or different context encoding schemes in the hippocampus.

The process of instruction-dependent restructuring of hippocampal representations is likely cortical-dependent, given the role of the cortex in language comprehension, but the exact mechanism by which this process occurs remains to be explored^{31,85}.

In our experiment, pre and post-instruction inference groups were mutually exclusive and we did not assign subjects to either group by design. Further experiments are needed to directly test how, if any, differences exist between experientially learned and instructed task representations.

Supplementary methods

Instructions given to patients:

----- Instruction set 1 (before first session) -----

In this task, we will show you a series of images, 4 of them in total. Your objective is to learn the correct response for each image (either left or right). In the beginning, you will not know what the correct answer is, so take a guess. The correct answer for an image may occasionally change, so pay close attention. For every correct answer you will receive a reward of either 25 or 5 cents. For an incorrect answer you will receive 0 cents. This is real money that you will receive before you leave the hospital in the form of a gift card to your favorite place (ex. Starbucks). You will have the opportunity to take a break halfway through.

----- Instruction set 2 (before second session) -----

You may have noticed that some images have the same correct response and some images have the same reward. Even when the correct response changes, they usually change together. In this experiment, we are going to try a different strategy. Pay attention to which images go together (i.e. have the same correct response and similar reward). This should make it a lot easier to perform the task. To make the task a little more difficult, now the correct response for each image will change a little more frequently.

Supplementary References

71. Lee, S. J. *et al.* Single-neuron correlate of epilepsy-related cognitive deficits in visual recognition memory in right mesial temporal lobe. *Epilepsia* **62**, 2082–2093 (2021).
72. Insausti, R. & Muñoz, M. Cortical projections of the non-entorhinal hippocampal formation in the cynomolgus monkey (*Macaca fascicularis*). *Eur J Neurosci* **14**, 435–451 (2001).
73. Aggleton, J. P., Wright, N. F., Rosene, D. L. & Saunders, R. C. Complementary Patterns of Direct Amygdala and Hippocampal Projections to the Macaque Prefrontal Cortex. *Cereb Cortex* **25**, 4351–4373 (2015).

74. Wang, J., John, Y. & Barbas, H. Pathways for Contextual Memory: The Primate Hippocampal Pathway to Anterior Cingulate Cortex. *Cereb Cortex* **31**, 1807–1826 (2020).
75. Spellman, T. *et al.* Hippocampal–prefrontal input supports spatial encoding in working memory. *Nature* **522**, 309–314 (2015).
76. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. *Proceedings of the National Academy of Sciences* **105**, 329–334 (2008).
77. Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat Neurosci* **18**, 1041–1050 (2015).
78. Duncan, J. & Owen, A. M. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci* **23**, 475–483 (2000).
79. Dosenbach, N. U. F. *et al.* A Core System for the Implementation of Task Sets. *Neuron* **50**, 799–812 (2006).
80. Ma, L., Chan, J. L., Johnston, K., Lomber, S. G. & Everling, S. Macaque anterior cingulate cortex deactivation impairs performance and alters lateral prefrontal oscillatory activities in a rule-switching task. *PLOS Biology* **17**, e3000045 (2019).
81. Monosov, I. E., Haber, S. N., Leuthardt, E. C. & Jezzi, A. Anterior cingulate cortex and the control of dynamic behavior in primates. *Curr Biol* **30**, R1442–R1454 (2020).
82. Economides, M., Guitart-Masip, M., Kurth-Nelson, Z. & Dolan, R. J. Anterior Cingulate Cortex Instigates Adaptive Switches in Choice by Integrating Immediate and Delayed Components of Value in Ventromedial Prefrontal Cortex. *J Neurosci* **34**, 3340–3349 (2014).
83. Cohen, Y., Schneidman, E. & Paz, R. The geometry of neuronal representations during rule learning reveals complementary roles of cingulate cortex and putamen. *Neuron* **109**, 839–851.e9 (2021).
84. Wang, S., Mamelak, A. N., Adolphs, R. & Rutishauser, U. Abstract goal representation in visual search by neurons in the human pre-supplementary motor area. *Brain* **142**, 3530–3549 (2019).
85. Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: a new window into the human brain’s unique capacity for flexible cognitive control. *Cogn Affect Behav Neurosci* **13**, 1–22 (2013).