Neuron, Volume *111*

Supplemental information

Goal-seeking compresses neural codes for space

in the human hippocampus and orbitofrontal cortex

Paul S. Muhle-Karbe, Hannah Sheahan, Giovanni Pezzulo, Hugo J. Spiers, Samson Chien, Nicolas W. Schuck, and Christopher Summerfield

Supplementary Figure 1

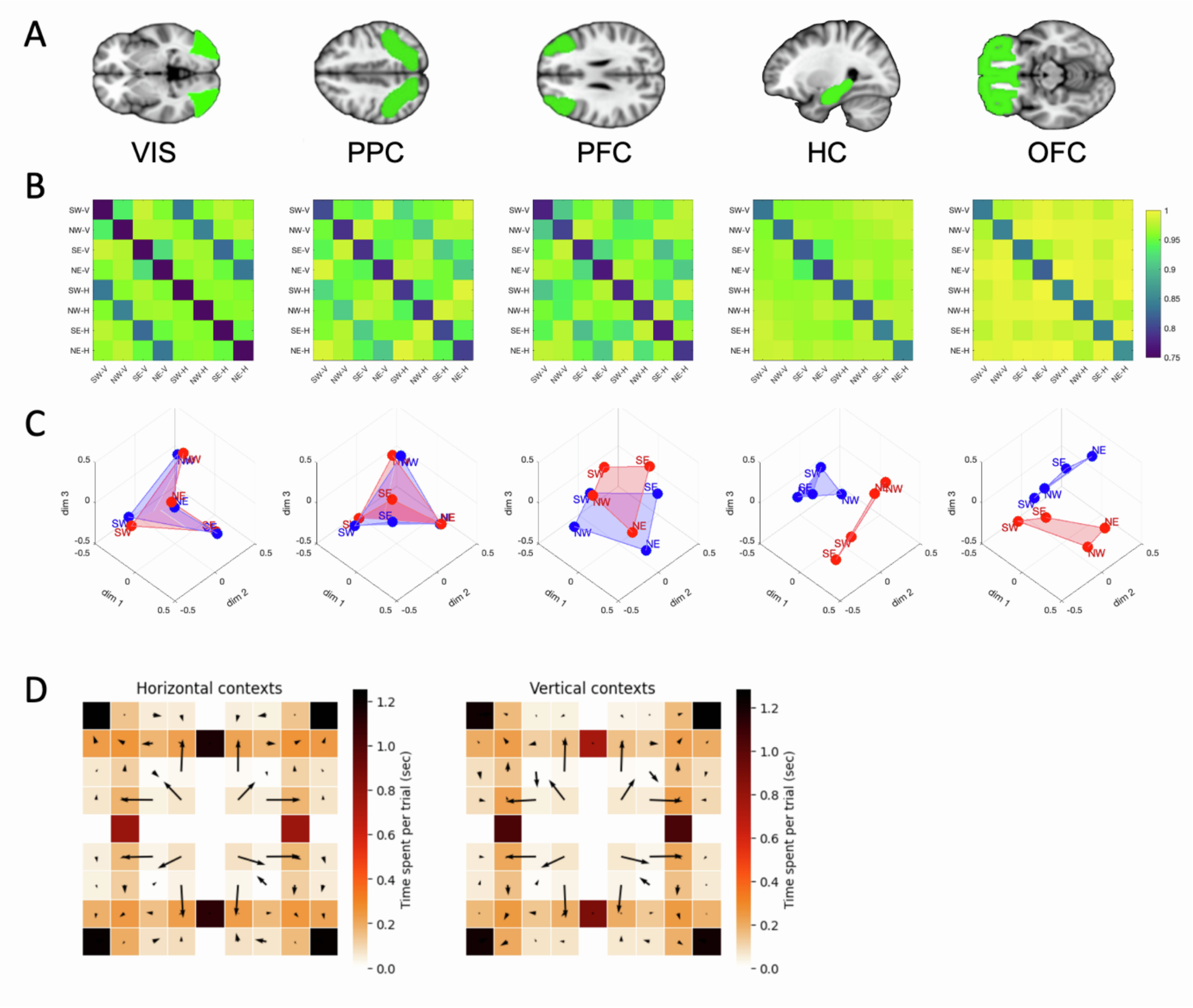

Figure S1. RDMs and neural geometry from the goal room period (related to Fig 3D) and heatmaps reflecting the grid square occupancies (related to Fig 1D). (A) Regions of interest, shown again for convenience **(B)** Group average RDMs for each ROI. Each 8 × 8 RDM is ordered {SW,NW,SE,NE} for first the vertical and then the horizontal context. Warmer colours indicate greater dissimilarity, and cooler colours greater similarity. **(C)** MDS plots (from the group average RDM) for each region. Blue dots are rooms in the vertical context and red in the horizontal context. For legibility, cardinally adjacent rooms within a context are linked by lines, which collectively form a quadrilateral when allocentric space is coded in just 2 dimensions. **(D)** Heatmaps of the average grid square occupancy per trial in each of the two contexts for human-controlled movement periods only. Black arrows show the average transition vector from each grid square. Data are averaged across participants.

Supplementary Figure 2

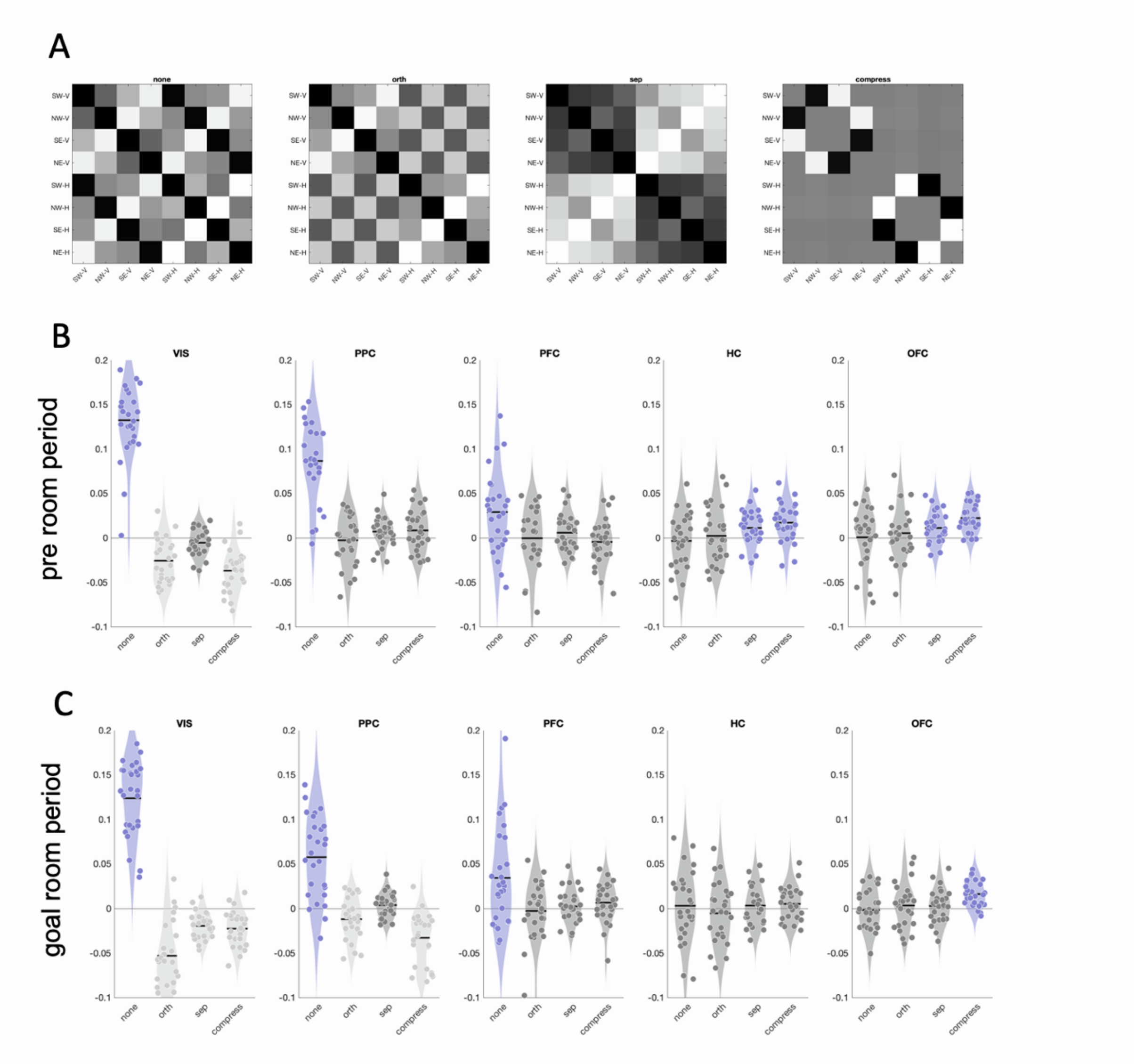

Figure S2. Correlations between RDMs from BOLD signals and RDMs from the place field model (related to Figure 4 and STAR Methods). (A) RDMs generated from the best fitting variant of the place field model, under parameterisations where (i) no parameters were allowed to vary ("none"); (ii) only the orthogonalization (β) parameter is allowed to vary; (iii) only the separation (γ) parameter is allowed to vary; and (iv) only the compression (ω) parameter is allowed to vary. Lighter colours indicate greater dissimilarity. **(B)** Coefficients from a regression on the data RDM for each region, from the pre-goal room period. **(C)** same as (B) but for the goal room period.

Supplementary Figure 3

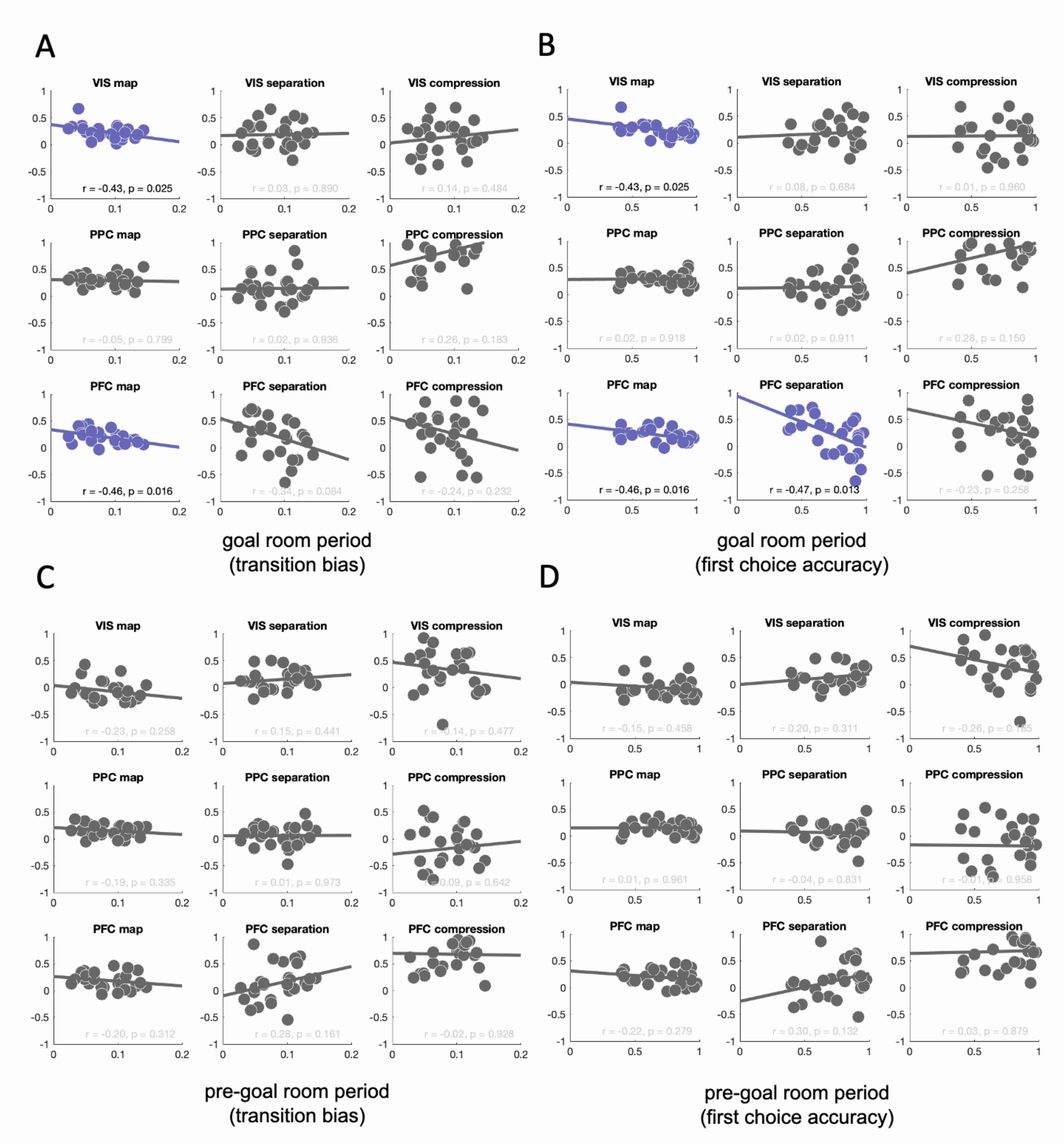

Figure S3. Correlations between neural scores (map, separation and compression) and behaviour for visual cortex, PPC and PFC (related to Fig. 4 D). (A) Correlations with transition bias for the goal room period; **(B)** Correlations with first choice accuracy for the goal room period; **(C)** Correlations with transition bias for the pre-goal room period; **(D)** Correlations with first choice accuracy for the pre-goal room period. Each dot is a single participant, and the line is the best linear fit. Blue colouring is used to highlight significant correlations ($p < 0.05$)

Supplementary Figure 4

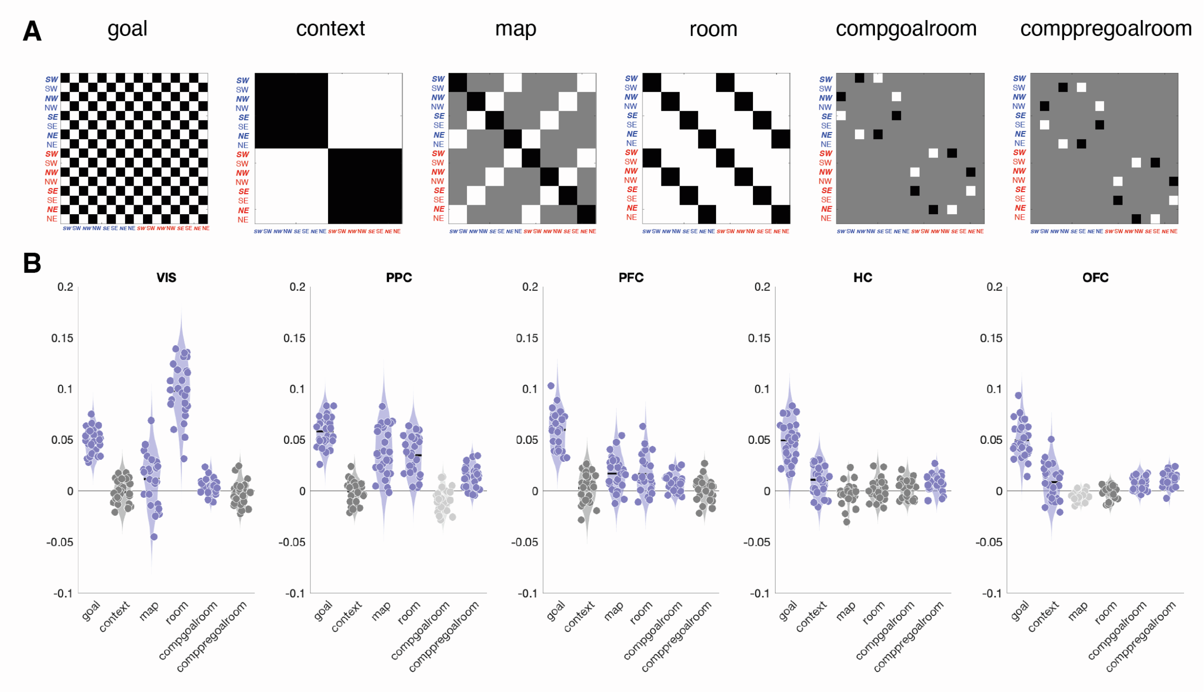

Figure S4. Inverted neural geometry across trial periods (related to Fig. 5). (A) Model RDMs used for the analysis shown in Fig. 5. **(B)** Coefficients for the regression of model RDMs for the full 16 × 16 (period × room × context) analysis described in Fig. 5. Blue dots show significant ($p < 0.01$) predictors.

Supplementary Figure 5:

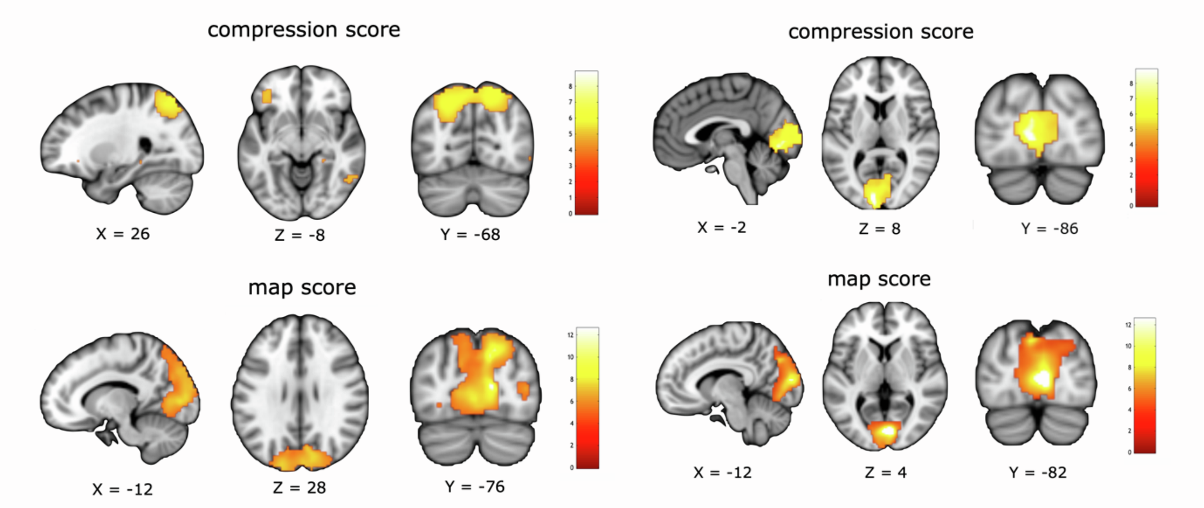

Figure S5. Whole-brain searchlight analyses with family-wise error correction (related to Fig. 5 E and F). Whole-brain effects of *compression score* and *map score* for the pre-goal room period (left) and the goal-rom period (right), rendered onto a template brain after familywise error correction at $p < 0.05$. During the pre-goal room period, we observed significant *compression score* in the posterior parietal cortex (peak t = 8.47, FWE $p < 0.001$), the right inferior temporal gyrus (peak [52 -63 -3], t = 4.83, FWE p = 0.003), the orbitofrontal cortex (peak [-38 36 -12], t = 4.79, FWE p = 0.005); the right putamen (peak [30 3 6], t = 5.45, FWE p = 0.024), the right hippocampus (peak [24 -39 -6], t = 5.36, FWE p = 0.032), and the right middle temporal gyrus (peak [-20 57 -18], t = 5.23 FWE p = 0.045). Significant correlations with *map score* were observed bilaterally in occipital and parietal cortices (peak [18 -84 6], t = 12.14, FWE $p < 0.001$), the precentral gyrus (peak [18 -84 6], t = 6.44, FWE p = 0.007), and the right insula (peak [48 3 -9], t = 6.08, FWE p = 0.015). During the goal room period, significant correlations with compression score were observed only in the medial portion of the visual cortex (peak [-8 -90 6], t = 8.42, FWE $p < 0.001$), and significant correlations with map score in bilateral visual and parietal cortices (peak [18 -66 51], t = 6.55, FWE $p < 0.001$). No significant correlations with *separation score* were observed in either period at the chosen threshold.

Supplementary Figure 6:

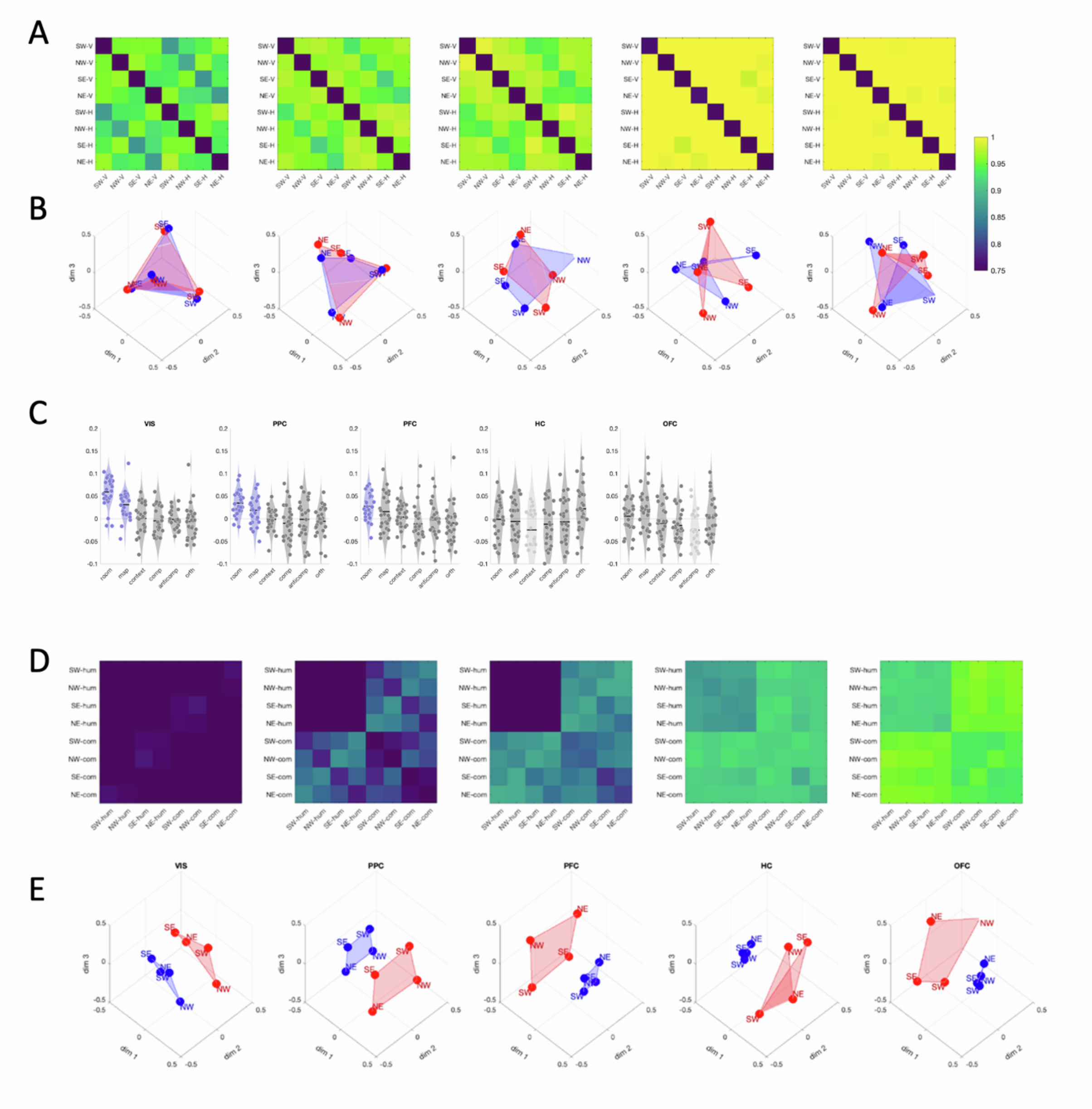

Figure S6. Control analyses examining brain activity before the onset of the first feedback signal (related to STAR Methods), and separating brain activity with human-based and computer-based control of the agent. (A) Neural geometries from an analysis focusing on brain activity during the first movement period of the trial (i.e., before any feedback has been received and agents have no knowledge of reward locations). Group average RDMs are shown for each ROI. Each 8 × 8 RDM is ordered by room {SW,NW,SE,NE} for first the vertical and then the horizontal context. Warmer colours indicate greater dissimilarity, and cooler colours greater similarity. **(B)** MDS plots from the group average RDM for each region. Blue dots are rooms in the vertical context and red in the horizontal context. For legibility, cardinally adjacent rooms within a context are linked by lines, which collectively form a quadrilateral when allocentric space is coded in just 2 dimensions. **(C)** Violin plots showing coefficients

for a competitive regression of model RDMs against each data RDM for the pre-goal room period. Each participant is an individual dot. Blue dots and shading (positive values) and light grey dots and shading (negative values) indicate $p < 0.05$. (**D)** Neural geometries from an analysis focusing on brain activity during movement periods after the first feedback, with separate predictors for events that were controlled by the human participant and the computer. The top row displays group average RDMs for each ROI. Each 8×8 RDM is ordered by room {SW,NW,SE,NE} for first the human and then the computer controlled events. Warmer colours indicate greater dissimilarity, and cooler colours greater similarity. **(E)** MDS plots from the group average RDM for each region. Blue dots are rooms in the human-controlled events and red dots are rooms in the computer-controlled events. For legibility, cardinally adjacent rooms within a context are linked by lines, which collectively form a quadrilateral when allocentric space is coded in just 2 dimensions.

Supplementary Table 1:

	Score matrix (pre-goal room)			***Score matrix (goal room)***		
	Map	***Separation***	***Compression***	***Map***	***Offset***	***Compression***
VIS	6.56***	0.16	-1.00	-3.13	0.30	2.27
PPC	10.82**	1.19	7.05***	3.44**	1.62	-4.33
PFC	4.45***	0.98	0.58	3.27**	0.39	4.33***
HC	2.11	2.66**	4.81***	-0.30	1.18	1.54
OFC	3.01**	3.08**	7.84***	1.37	0.35	5.98***

Table S1. Model-free analyses of neural geometries (related to Fig 4 A-C). In this analysis, we constructed “score matrices” indicating which pairs of vertices were compared with each other. First, we asked whether the planes for each context were roughly quadrilateral (reflecting the spatial layout of the environment). Here, we compared distances between rooms that were spatially adjacent (e.g., NE and NW) to those that were not (e.g., NE and SW), yielding a single *map score* which was zero under the null, but for which positive scores provided evidence for quadrilateral structure. In **Fig. 4B** (see also **Table 3**) we can see that there is a significant map score in all regions except HC during the pre-goal room period, and in PPC and PFC during the goal room period. Secondly, we computed a *separation score* by comparing neural distances between each room and every other room within and between contexts. Whilst the effect of separation was only marginal in the regression analysis, the separation score was reliable during the pre-goal room period for both HC and OFC. Finally, we computed a *compression score* by comparing distances between N and S and E and W rooms in each context; this score was positive if E and W rooms were neurally more proximal in the horizontal context and N and S rooms were more proximal in the vertical context, and negative for the converse. T-values for a test of each score in each period. Each row is a brain region. Asterisks: ** p < 0.01, *** p < 0.001 after FDR correction.