Neuron

Article

The neural architecture of theory-based reinforcement learning

Highlights

- Human gameplay is consistent with theory-based reinforcement learning
- Theory representations explain activity in IFG, MFG, and SFG
- IFG, FFG, and the occipital gyri show increased activity during theory updating
- Information flows top-down during gameplay and bottom-up during theory updating

Authors

Momchil S. Tomov, Pedro A. Tsividis, Thomas Pouncy, Joshua B. Tenenbaum, Samuel J. Gershman

Correspondence

mtomov@g.harvard.edu

In brief

Humans learn rich intuitive theories that allow them to plan and generalize in complex domains. Tomov et al. found evidence that such theories are represented in prefrontal regions and are updated in response to theory prediction errors computed in posterior regions.

Tomov et al., 2023, Neuron *111*, 1331–1344
April 19, 2023 © 2023 Elsevier Inc.
https://doi.org/10.1016/j.neuron.2023.01.023

CellPress

Neuron

CellPress

Article

The neural architecture of theory-based reinforcement learning

Momchil S. Tomov,[1,3,4,5,*] Pedro A. Tsividis,[2,3] Thomas Pouncy,[1] Joshua B. Tenenbaum,[2,3] and Samuel J. Gershman[1,3]
[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Motional AD, Inc., Boston, MA 02210, USA
[5]Lead contact
*Correspondence: mtomov@g.harvard.edu
https://doi.org/10.1016/j.neuron.2023.01.023

SUMMARY

Humans learn internal models of the world that support planning and generalization in complex environments. Yet it remains unclear how such internal models are represented and learned in the brain. We approach this question using theory-based reinforcement learning, a strong form of model-based reinforcement learning in which the model is a kind of intuitive theory. We analyzed fMRI data from human participants learning to play Atari-style games. We found evidence of theory representations in prefrontal cortex and of theory updating in prefrontal cortex, occipital cortex, and fusiform gyrus. Theory updates coincided with transient strengthening of theory representations. Effective connectivity during theory updating suggests that information flows from prefrontal theory-coding regions to posterior theory-updating regions. Together, our results are consistent with a neural architecture in which top-down theory representations originating in prefrontal regions shape sensory predictions in visual areas, where factored theory prediction errors are computed and trigger bottom-up updates of the theory.

INTRODUCTION

Reinforcement learning (RL) is a normative framework prescribing how agents ought to act in order to maximize rewards in the environment.[1] In the field of artificial intelligence, RL has allowed artificial agents to reach and surpass human-level performance across a variety of domains previously beyond the capabilities of computers.[2–5] In the fields of psychology and neuroscience, RL has offered a compelling account of behavioral and brain data across a number of species and experimental paradigms.[6–9] Most of this work has focused on model-free RL, a kind of RL in which the agent directly learns a mapping from different states in the environment to actions and/or values. Model-based RL, on the other hand, posits that the agent learns an internal model of the environment, which is used to simulate the outcomes of different actions. Behavioral and neural studies have found evidence for both kinds of RL,[10–13] yet model-based RL has received relatively less attention and is often studied using simple toy environments with small state spaces. This is largely owing to the relative scarcity of powerful model-based RL algorithms capable of matching human learning in complex domains,[14] leaving open the question of what the "model" in model-based RL is and how it is learned and represented by the brain.

One possible answer from cognitive science is theory-based RL,[15–17] a strong form of model-based RL in which the model is an intuitive theory—an abstract causal model of world dynamics rooted in core cognitive concepts such as physical objects, intentional agents, relations, and goals (Figure 1). Building on findings in developmental psychology, theory-based RL posits that the agent learns the theory from experience using probabilistic inference and uses it together with an internal simulator to predict and evaluate the outcomes of different action sequences generated by an internal planner. Theory-based RL has captured patterns of human learning,[16,17] exploration,[16] and generalization[15] in complex domains where model-free and simpler model-based RL approaches fail or learn rather differently. This has provided strong support for theory-based RL as a concrete realization of human model-based RL.

Building on this work, our study aims to identify brain regions involved in theory-based RL and how they map to its constituent processes. To achieve this, we used a particular formalization of theory-based RL[16] to analyze functional magnetic resonance imaging (fMRI) data collected from human participants while they learned to play Atari-style games designed to mirror some of the richness and complexity of real-world tasks. Our analyses revealed evidence that theory representations in inferior frontal gyrus (IFG) and other prefrontal regions are activated and updated in response to theory prediction errors—discrepancies between theoretical predictions and actual observations—which are in turn computed in occipital and ventral stream regions such as the fusiform gyrus (FFG). We also found evidence that, much like in our theory-based RL model, theory updating in the brain is factored into updating of objects, relations, and goals,

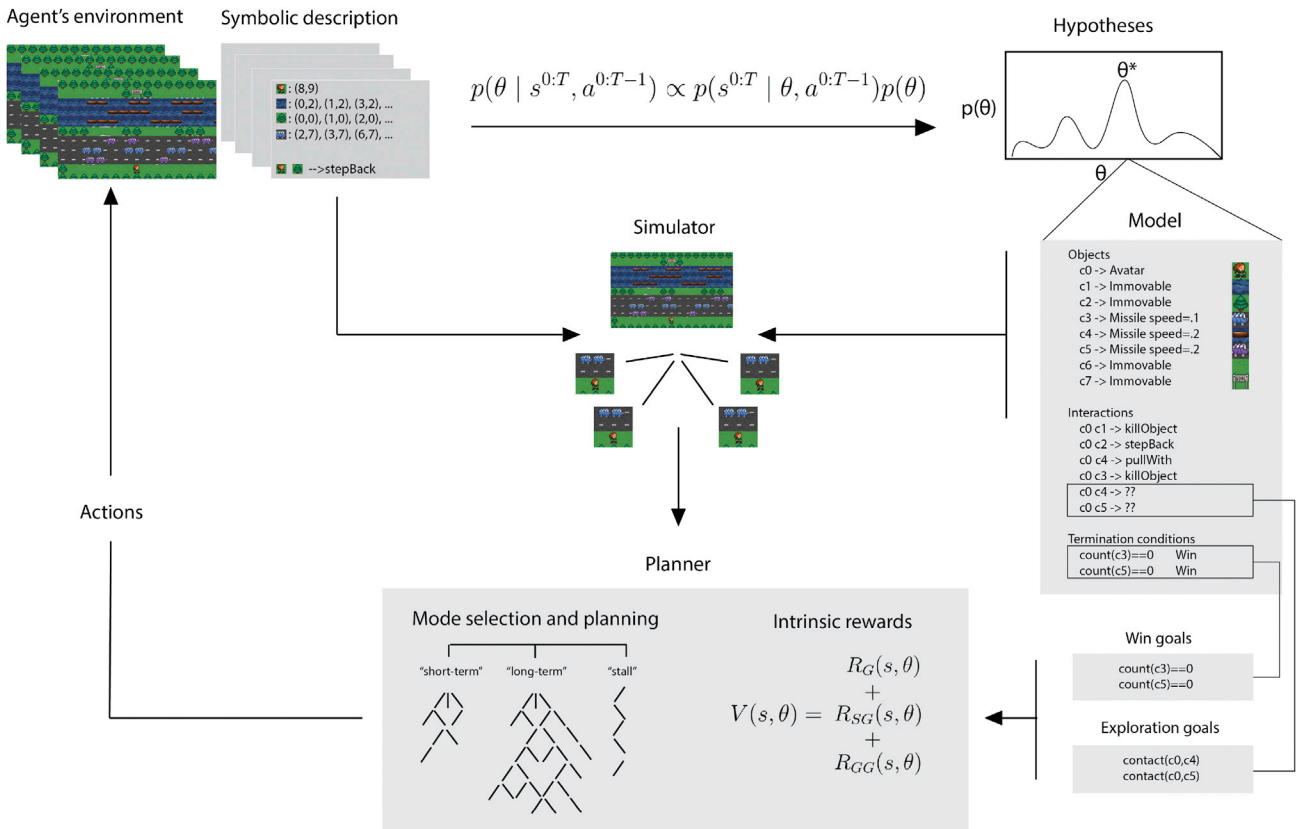

Figure 1. EMPA architecture

Symbolic descriptions of game frames are fed to an inference engine which updates the most-likely theory, θ^*, using an approximation of Bayesian inference. The theory consists of objects (sprites), relations (interactions), and goals (termination conditions). Exploitative (win) and exploratory goals based on the theory are fed to a planner which uses a theory-based internal simulator and an intrinsic reward function to search for rewarding action sequences. The agent then takes actions in the environment according to the best plan. Reused with permission from Tsividis et al.[16]

suggesting key differences between these cognitive components. Finally, analyses of effective connectivity suggest that theory inference involves both feedforward and feedback processing reminiscent of hierarchical predictive coding.[18,19] Together, these results present the first direct evidence for theory-based RL in the brain and establish a foundation for understanding its underlying neural processes.

RESULTS

We scanned 32 human participants using fMRI while they played six Atari-style games (Figure 2A; Table S4). Each game had nine levels of increasing complexity and had to be learned from experience, without any visual hints or prior information about the rules. For data analysis purposes, games were interleaved and balanced across pairs of runs (Figure 2B).

As a particular instantiation of theory-based RL, we used the explore, model, plan agent (EMPA; Figure 1) proposed by Tsividis et al.[16] Theories are formalized as symbolic, probabilistic program-like descriptions of game dynamics that specify the different object kinds, the outcomes of interactions between them, and the win/loss conditions. EMPA performs Bayesian inference over the space of theories and uses the most likely theory to run internal simulations and search for rewarding action sequences. Tsividis et al.[16] showed that EMPA exhibits human-level learning efficiency in a large suite of Atari-style games, including those used in our study. They also showed that EMPA exhibits human-like object-oriented exploratory behaviors. In contrast, model-free RL agents failed on both counts, learning orders of magnitude more slowly and exploring much more randomly than humans.

Consistent with these results, we found that EMPA performed similarly to our participants (Figure 2C; no significant difference, two-sided Wilcoxon rank-sum test based on simulated and actual expected bonus payouts), while both humans and EMPA performed significantly better than a pretrained deep RL network, the double DQN (DDQN; $p < 10^{-10}$), a powerful model-free RL algorithm,[20] variants of which have been put forward as accounts of human model-free RL in complex domains.[9] Consistent with Tsividis et al.'s[16] results, we also found that EMPA learned at a rate similar to humans (Figure S1A; $t(30) = 1.5, p = 0.13$, two-sample t test of fitted linear coefficients), while the DDQN learned significantly more slowly ($t(30) = 3.9, p = 0.0005$). Ablations to the intrinsic rewards, planner, and exploration components of EMPA revealed that intrinsic rewards are critical for this effect ($t(30) = 4.1$,

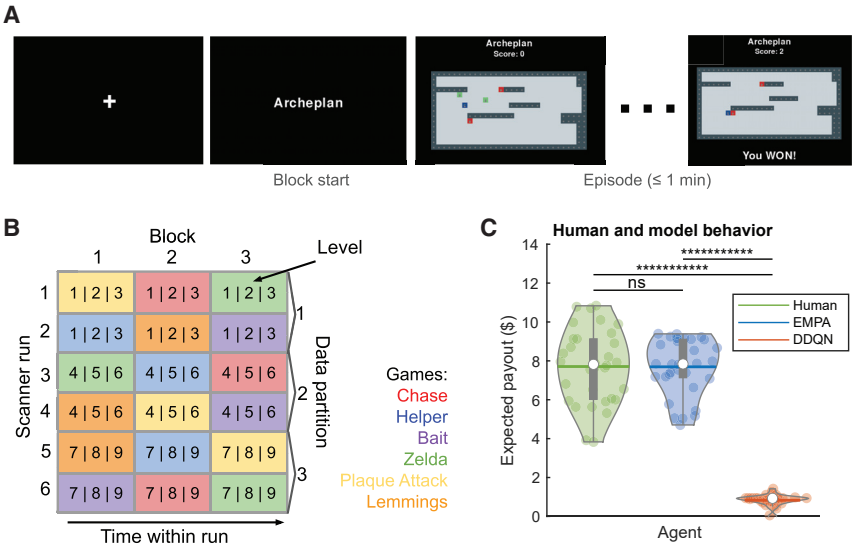

Figure 2. Study design and behavioral results

(A) Participants played different Atari-style games. The game names, colors, and textures shown were randomized for each participant and were unrelated to the game rules (but were consistent across different levels of the same game for each participant).

(B) Example scan session for single participant. Runs were paired into balanced data partitions, with blocks shuffled within each partition. See also Table S4.

(C) Behavioral results from participants (green) and generative play by EMPA (blue) and pretrained DDQN (red). Each colored dot represents a single real or simulated participant, respectively. White dots depict medians, box plots depict upper and lower quartiles, horizontal lines across kernel density estimates depict means. ns, not significant, **********$p < 10^{-10}$ (two-sided Wilcoxon rank-sum test). See also Figure S1.

$p = 0.0003$), while planner optimizations ($t(30) = 1.5, p = 0.15$) and theory-driven exploration ($t(30) = 1.0, p = 0.33$) may be less critical, at least in our design. All of our subsequent neural analyses are based exclusively on EMPA theory learning from human gameplay and thus do not rely on any of these components. For a thorough comparison between the behavior of humans, EMPA, different ablations of EMPA, and different deep RL models, refer to Tsividis et al.[16]

Theory representations in prefrontal cortex

The central component of theory-based RL is the theory that the agent continuously infers from experience. To identify brain regions representing the inferred theory, we replayed each participant's gameplay through EMPA and used the inferred theory sequences to predict the blood-oxygen-level-dependent (BOLD) signal in each voxel using a linear encoding model fit with Gaussian process (GP) regression (Figure 3A), a generalization of the more commonly used ridge regression.[21] In order to embed the symbolic theories in a vector space for the encoding model, we used holographic reduced representations[22] (HRRs), a method for encoding complex compositional structure in distributed form. Similarly to previous work,[9,23] we correlated the cross-validated predicted and actual BOLD time courses; we then Fisher-z-transformed the resulting Pearson correlation coefficients to compute a predictivity score for each voxel. Predictivity scores were aggregated across participants using t tests. The resulting group-level t-maps were thresholded at $p < 0.001$ and whole-brain cluster family-wise error (FWE) corrected at $\alpha = 0.05$. This revealed significant predictivity scores across a distributed bilateral network of regions (Figures 3B, S4A, and S4B; Table S1). In prefrontal cortex, we found bilateral clusters in IFG, as well as unilateral clusters in middle (MFG) and superior frontal gyrus (SFG) and the supplementary motor area. In posterior areas, we found a large bilateral cluster, starting from early visual regions in occipital cortex, extending into higher visual regions and then further into the ventral and dorsal streams, including FFG and middle temporal gyrus in temporal cortex, and inferior parietal gyrus and angular gyrus in parietal cortex.

We performed the same analysis using three control models[9]: DDQN agents pretrained on corresponding games to control for model-free RL representations (also used as a behavioral control in Tsividis et al.[16]), principal component analysis (PCA) to control for low-level visual features,[24,25] and a variational autoencoder (VAE) to account for high-level visual and state features.[26–28] Similar to Cross et al.,[9] we compared models using *a posteriori* bilateral anatomical regions of interest (ROIs) based on cross-referencing the t-maps from all models with the automated anatomical labeling atlas[29] (AAL3 atlas). We compared models separately in each ROI based on the fraction of voxels with a significant correlation ($\alpha = 0.05$) between predicted and actual BOLD signal. In prefrontal regions, EMPA largely outperformed all three control models (Figures 3C, S4A, and S4B), specifically in the triangular and opercular parts of IFG, as well as in MFG and SFG. Additionally, EMPA outperformed all three control models in middle temporal gyrus. A two-way ANOVA with model and ROI group as factors in a 4 × 4 design (Figure S4B) revealed a significant interaction between the effects of model and ROI group ($F(9, 496) = 4.5, p < 10^{-5}$, two-way ANOVA), with significant simple main effects both for model ($p < 10^{-8}$) and ROI group ($p < 10^{-8}$). This suggests that the effects of theory representation in those regions are not simply due to visual or model-free RL confounds.

In order to probe which aspects of the EMPA theory drive the encoding model results, we repeated the analysis using different components of the theory—objects, relations, and goals—but did not find any systematic differences (Figure S4C). We also repeated the theory analysis using simplified object embeddings in which every object has a single approach/avoid feature. We found that this improved predictivity across cortex (Figure S4H), pointing to a more parsimonious, action-related representation of objects than the one used by EMPA.[30]

In order to refute the possibility that the encoding model results are simply due to variance between games, we repeated

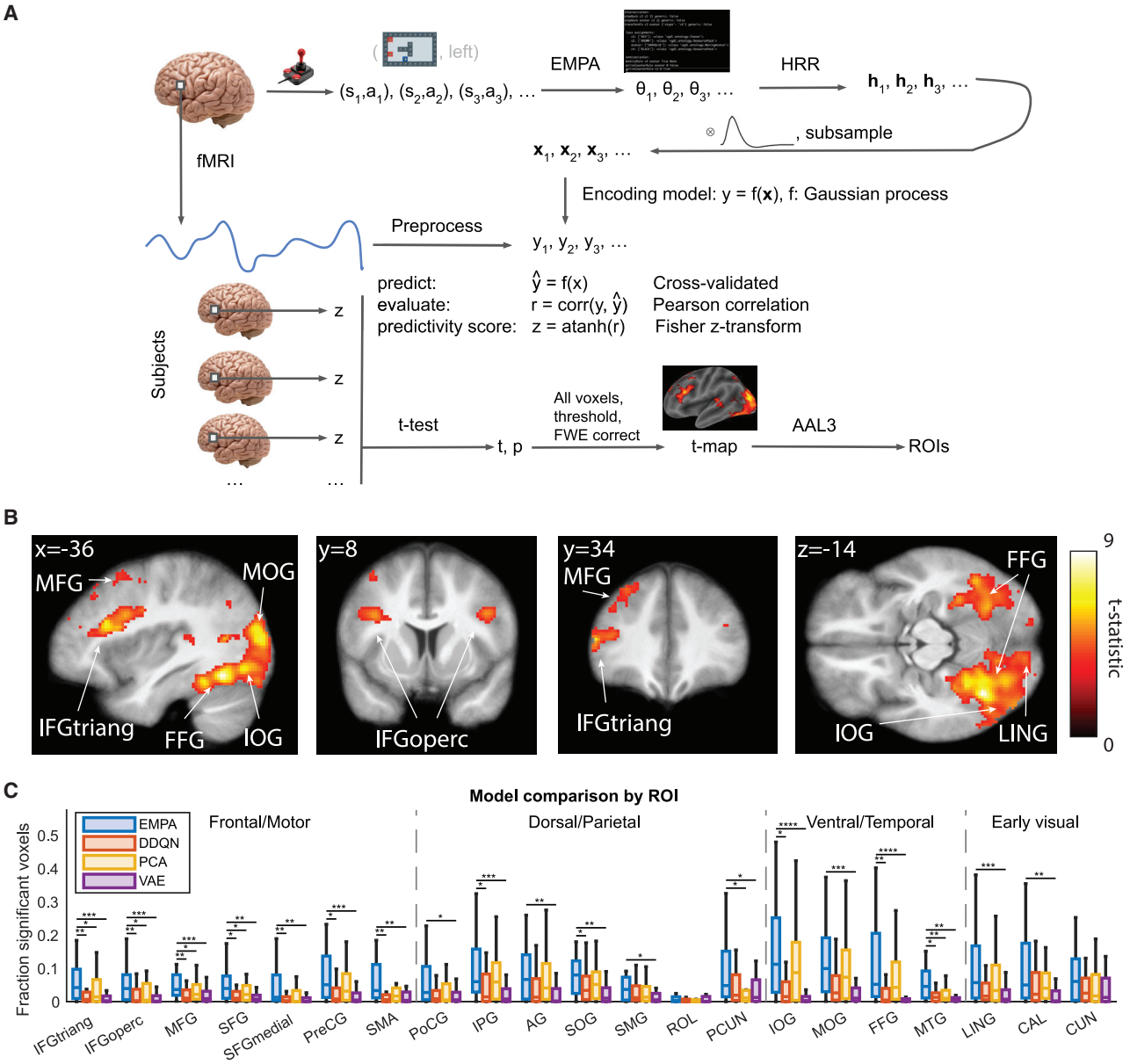

Figure 3. Theory representations map to regions in prefrontal cortex and ventral/dorsal streams

(A) Encoding model analysis pipeline. State-action sequences $((a_1, s_1), (s_2, a_2), (s_3, a_3), \ldots)$ from human gameplay were replayed through EMPA. Inferred theory sequences $(\theta_1, \theta_2, \theta_3, \ldots)$ were embedded in a vector space, convolved with the hemodynamic response function, and subsampled to get feature vectors $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots)$. Preprocessed BOLD signal from each voxel $(y_1, y_2, y_3, \ldots)$ was regressed onto feature vectors using GP regression. Resulting predictivity scores z were aggregated across participants using two-sided t tests. Resulting t-maps were thresholded at $p < 0.001$ and whole-brain cluster FWE corrected at $\alpha = 0.05$. Analogous analyses were performed with control models (DDQN, PCA, and VAE). See also Figures S1 and S2.

(B) Group-level t-maps from (A). ROIs are noted as IFGtriang, inferior frontal gyrus, triangular part; IFGoperc, inferior frontal gyrus, opercular part; MFG, middle frontal gyrus; SFG, superior frontal gyrus; PreCG, precentral gyrus; SMA, supplementary motor area; PoCG, postcentral gyrus; IPG, inferior parietal gyrus; AG, angular gyrus; SMG, supramarginal gyrus; ROL, rolandic operculum; PCUN, precuneus; IOG, inferior orbital gyrus; MOG, middle orbital gyrus; SOG, superior orbital gyrus, FFG, fusiform gyrus; MTG, middle temporal gyrus; LING, lingual gyrus; CAL, calcarine fissure; CUN, cuneus. See also Figure S3 and Table S1.

(C) Fraction of voxels with significant correlation ($\alpha = 0.05$) between predicted and actual BOLD in anatomical ROIs, aggregated across participants. Medians with boxes representing top and bottom participant quartiles and whiskers representing data range, excluding outliers (outliers plotted in Figure S4A and included in all statistical tests). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ (two-sided Wilcoxon signed rank tests). See also Figure S4.

the analysis for each game separately and averaged the results across games. We found that EMPA still outperformed all control models in prefrontal cortex (Figure S4D), indicating that these results are not merely due to game differences unrelated to theory learning, such as different sensory properties or sensorimotor contingencies. To investigate whether there are any systematic differences in theory encoding between games, we repeated this analysis for games that require more planning and games that require less planning. EMPA outperformed all control models in prefrontal cortex for games that require more planning (Figure S4E), but not for games that require less planning (Figure S4F). A direct comparison revealed stronger theory encoding in prefrontal cortex for games that require more planning compared with games that require less planning (Figure S4G).

Theory update signals in inferior frontal gyrus, occipital gyri, and fusiform gyrus

After identifying regions representing the inferred theory, we next sought to identify brain regions involved in theory inference. Based on our previous work,[31] we reasoned that such regions might show greater activity during theory updating, reflecting the temporary increase in computational demands. Because theory updates are triggered by surprising events which violate theoretical predictions, such an increase in neural activity could also be interpreted as a kind of theory prediction error. We used a general linear model (GLM) with impulse regressors at theory update events—frames at which EMPA switched from one most likely theory to another based on the participant's gameplay (Figures 4A and S5; Table S2). The group-level contrast for theory updating (Figures 4B and S7A, Table S3; thresholded at $p < 0.001$ and whole-brain cluster FWE corrected at $\alpha = 0.05$) revealed a distributed bilateral network of regions that largely overlapped with the regions from our theory representation analysis. Most notably, in prefrontal cortex, we found bilateral clusters in IFG, in addition to unilateral clusters in SFG, orbital frontal cortex, and the supplementary motor area. We also found a large bilateral posterior cluster covering early and late visual regions in occipital cortex, extending into angular gyrus and precuneus in the dorsal stream, and extending into FFG in the ventral stream.

To ensure enough power for this analysis, the game levels in our experiment were specifically designed to elicit learning throughout the entire session (Figure S5B; see experimental design). Nevertheless, the frequency of theory update events tended to decrease over the course of the session (Figure S5A: all games; $\tau_b = -0.21$, $n = 540$, $z = -7.34$, $p < 10^{-12}$, two-tailed Mann-Kendall test; $p < 10^{-8}$ for individual games, except for Avoid George, $p = 0.7$). This led us to hypothesize that the neural theory update effect might differ between earlier levels, when there is more theory learning, and later levels, when there is less theory learning (Figure S5). To investigate this hypothesis, we repeated this analysis separately for each data partition (Figures S7B–S7D). We found that the theory update effect qualitatively diminished over time, with fewer and smaller clusters surviving cluster FWE correction in later partitions. However, a direct contrast between the first data partition (Figure S7B) and the third data partition (Figure S7D) showed that this difference is not significant (no voxels survived cluster FWE correction), suggesting that EMPA is able to consistently capture theory updating throughout the entire session.

To control for potential confounds, we included a number of nuisance regressors in the GLM for events of non-interest, including visual changes, key presses, and game events relevant for theory updating (Table S2). A follow-up analysis using anatomical ROIs from the theory updating contrast for the entire session revealed that some nuisance regressors also show a significant effect (Figure S6). To directly compare the neural responses to different event types, we generated peri-event time histograms (PETHs) from the baseline-adjusted BOLD signal following theory updates and other control events (Figures 4C and 4D) in bilateral anatomical ROIs with a significant theory update effect (Figure S6). Notice that this is not a confirmatory analysis but rather a complementary analysis that (1) verifies whether the effect in those regions is driven by a positive BOLD response to theory updates rather than some combination of theory updates and nuisance regressors and (2) verifies whether the BOLD response to theory updates in those regions is stronger than the BOLD response to control events. We found that, in contrast to other control events, the increase in BOLD signal was larger and more sustained after theory updates in IFG, all three occipital gyri, and FFG (two-sided t tests in Figure 4C, paired t tests in Figure 4D). These results suggest that these regions respond specifically to theory updating, pointing to their potential involvement in computing theory-prediction errors—discrepancies between the perceived world state and the predicted world state based on the theory—or in performing the theory update computation in response to such errors. It is also noteworthy that these regions also appear in the theory representation brain maps (Figure 3B), with IFG specifically representing the learned theory (Figure 3C).

Separate update signals for different theory components

The EMPA theory consists of three components: a set of object types and their physical and/or intentional properties (because they could be other agents), a set of relations between objects describing the outcomes of object-object interactions, and a set of goals that the agent pursues. For tractability, EMPA factors theory inference into separate inference processes for objects, relations, and goals.[16] However, the theory update GLM described above does not distinguish between updates for separate theory components. Rather, theory update events occur when either objects, relations, or goals are updated (Figure 5A, top). When we repeated the PETH analysis described above for individual theory component updates, we found that some regions respond differentially to different component updates (Figures S7E and S7F). This led us to hypothesize that the brain might factor theory learning similarly to EMPA.

To investigate this hypothesis, we fit a GLM in which theory updating was split into three separate regressors for object, relation, and goal updates (Figure 5A, bottom). We additionally fit three control GLMs, each with a single component update (Figure 5A, middle). We compared GLMs using random effects Bayesian model selection[32] in the ROIs showing a significant

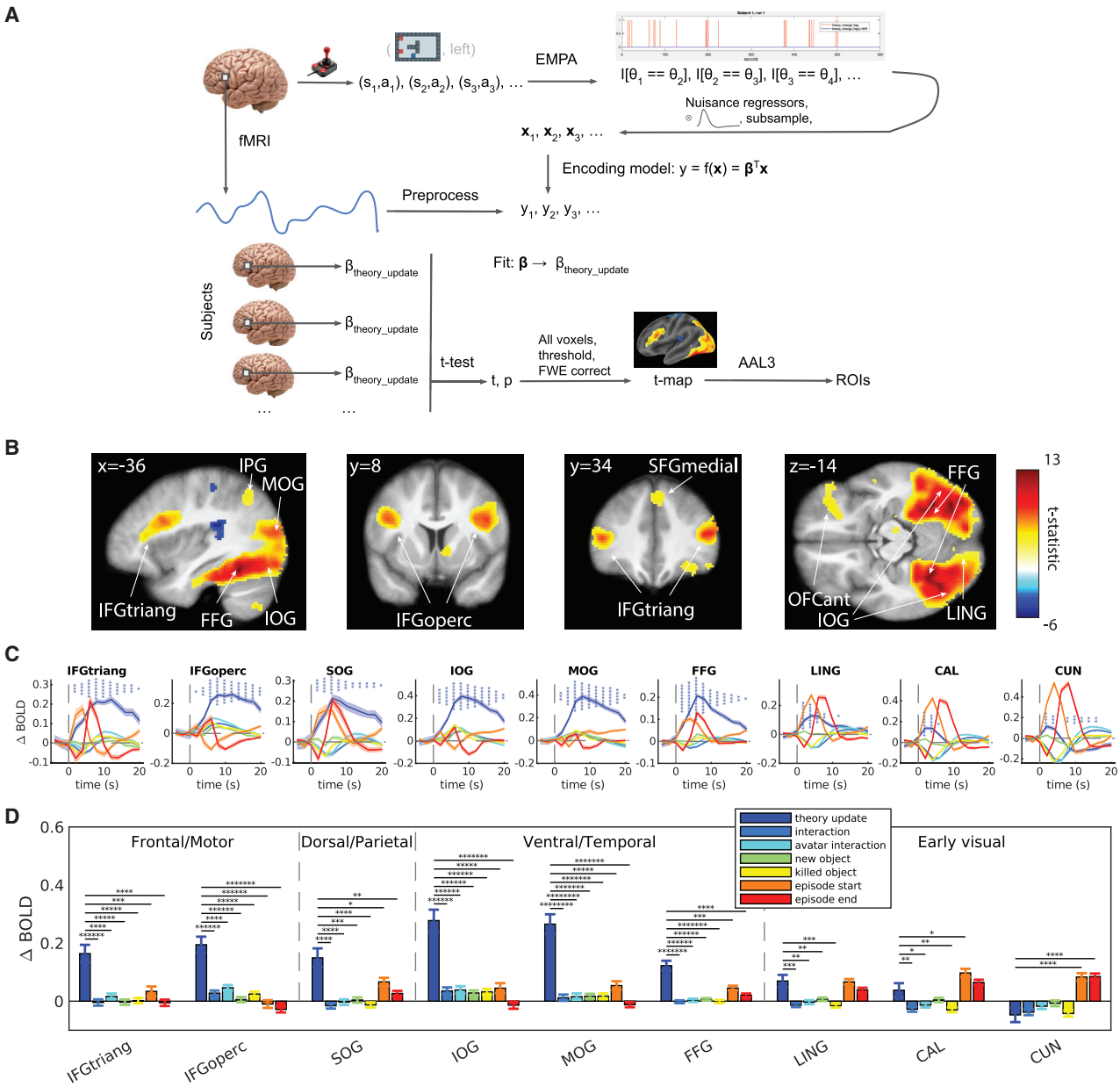

Figure 4. Theory learning signals in prefrontal cortex and ventral/dorsal streams

(A) GLM analysis pipeline. Similarly to Figure 3A, frame-by-frame state-action sequences $((a_1,s_1),(s_2,a_2),(s_3,a_3),\ldots)$ from human gameplay were replayed through EMPA. Corresponding theory update sequences $(I[\theta_1 \equiv \theta_2], I[\theta_2 \equiv \theta_3], I[\theta_3 \equiv \theta_4],\ldots)$ from EMPA were entered as regressors in a GLM. Resulting theory update beta estimates ($\beta_{\text{theory_update}}$) for individual voxels were aggregated across participants using two-sided t tests. Resulting t-maps were thresholded at $p < 0.001$ and whole-brain cluster FWE corrected at $\alpha = 0.05$. See also Figure S5 and Table S2.

(B) Group-level t-maps from GLM analysis in (A). ROIs noted as OFCant, anterior orbital gyrus; SFGmedial, superior frontal gyrus, medial, and the rest as in Figure 3B. See also Figures S6 and S7 and Table S3.

(C) Peri-event time histograms showing the average change in BOLD signal following theory updates and different control events in ROIs with significant $\beta_{\text{theory_update}}$. Colored fringes depict error bars (SEM) across participants. Stars indicate significance for theory updates for each time point. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$, *****$p < 0.00001$, ******$p < 10^{-6}$ (two-sided t tests).

(D) Change in BOLD signal from (C) averaged over 20 s following corresponding event. Error bars depict SEM across participants. Significance notation as in (C) (paired t tests).

BOLD increase in response to all three individual component updates (Figures S7E and S7F). We found that the GLM with separate component updates best explains the BOLD signal in IFG and all three occipital gyri (Figure 5B; Table 1). This suggests that, similarly to EMPA, the brain also performs a factored theory update.

Table 1. GLM comparison results

AAL3 region	GLM PXPs				
	Theory updates	Object updates	Relation updates	Goal updates	Object, relation, goal updates
IFG pars triangularis	<0.0001	<0.0001	<0.0001	<0.0001	0.9998
IFG pars opercularis	0.1717	0.1717	0.1586	0.1649	0.3328
Superior occipital gyrus	0.0004	<0.0001	<0.0001	<0.0001	0.99953
Inferior occipital gyrus	<0.0001	<0.0001	<0.0001	<0.0001	0.99998
Middle occipital gyrus	<0.0001	<0.0001	<0.0001	<0.0001	0.99996
Fusiform gyrus	0.7140	0.0097	0.0004	0.0004	0.2756

PXP, protected exceedance probability; IFG, inferior frontal gyrus.

Theory representations activated during updating

The overlap (Figure 6A) between the brain regions representing the theory (Figure 3) and the brain regions responding to theory updating (Figure 4) was somewhat surprising. *A priori*, these regions do not necessarily have to be the same: one analysis looks for regions consistently representing the theory, without any increase in activity around change points, while the other analysis looks for regions with increased activity at theory change points, without regard for the content of the theory itself. Indeed, we found no significant correlation between theory embeddings and theory updates (Figures S8A and S8B) derived from EMPA. This led us to hypothesize that the two computations

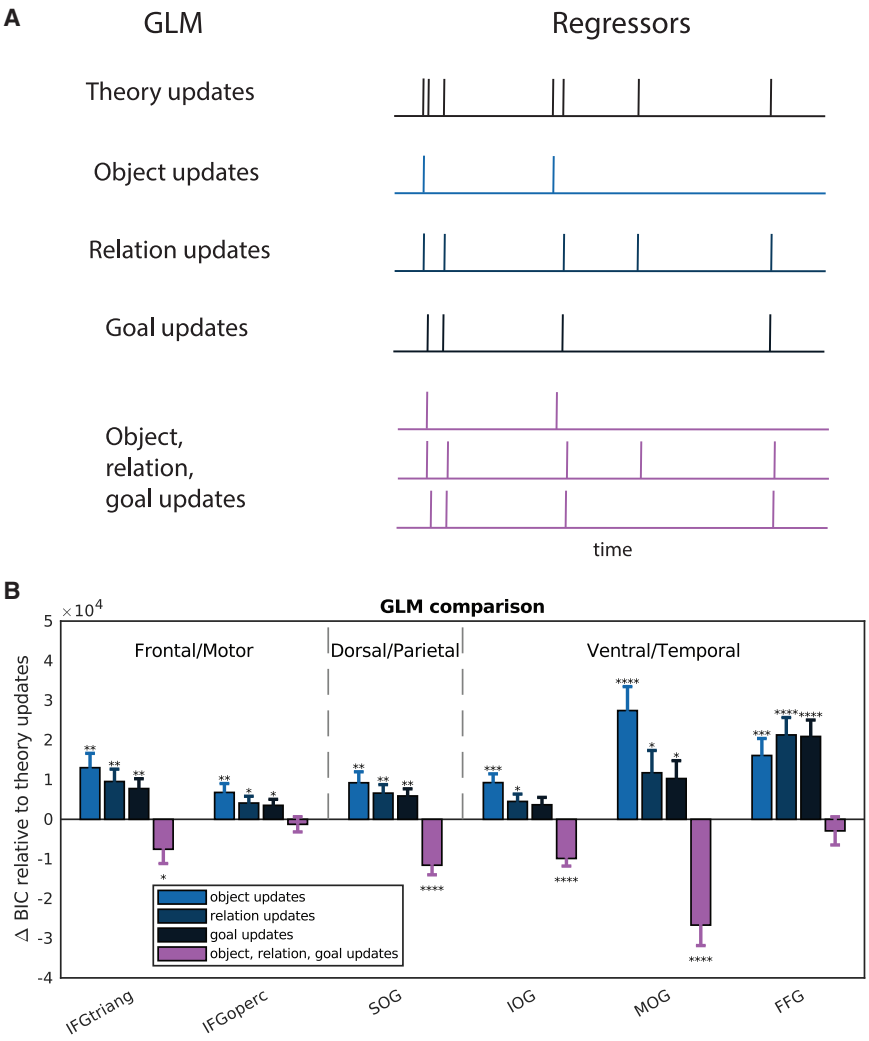

Figure 5. Separate update signals for different theory components

(A) Illustration of GLMs with impulse regressors for unified theory updates (top GLM; same as in Figure 4), single component updates (middle three GLMs), and separate updates for all three components (bottom GLM).

(B) GLM comparison in ROIs showing a significant increase in BOLD signal for all three theory components (Figures S7E and S7F). ROIs noted as in Figure 3B. Bars denote GLM BICs relative to theory update GLM BIC. Error bars denote SEM across participants. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ (two-sided t tests). BIC, Bayesian information criterion. See also Figure S7.

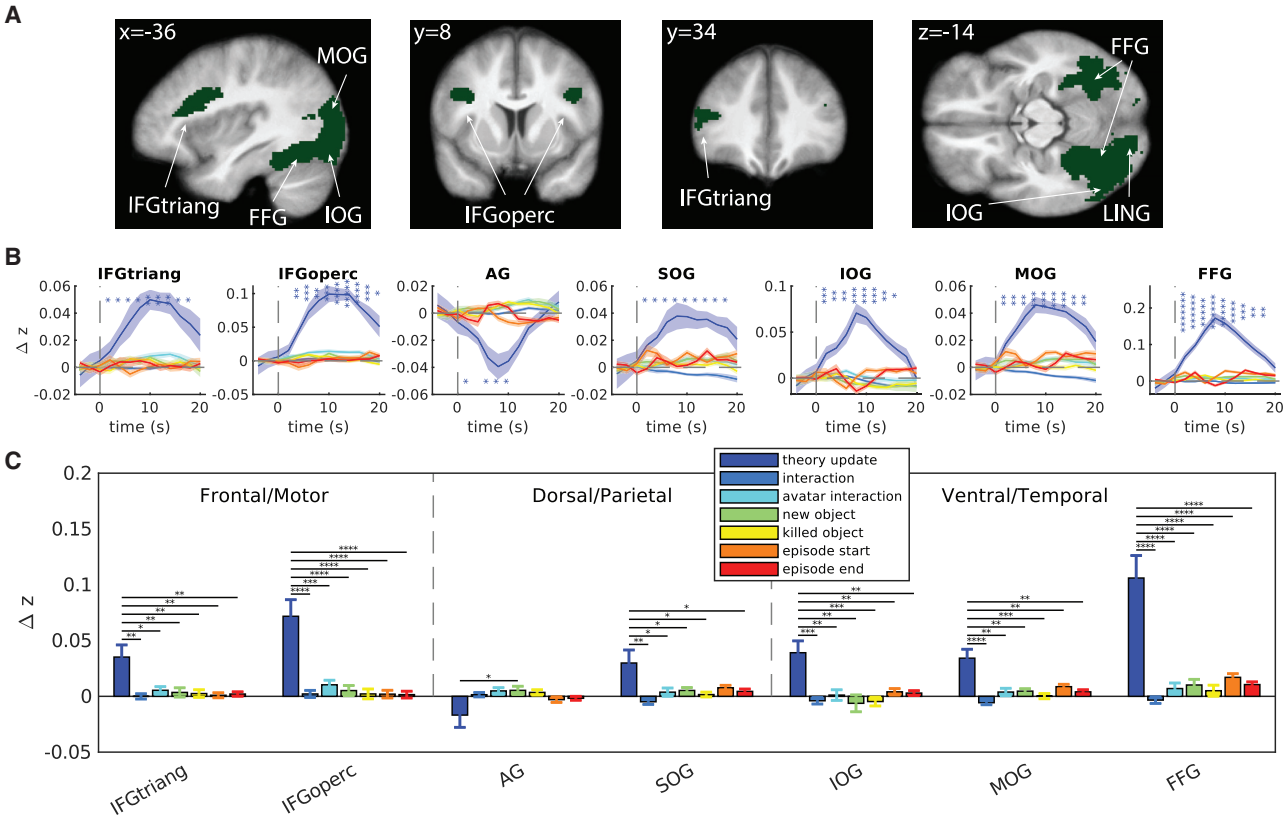

Figure 6. Theory representations activated during updating

(A) Overlap between significant clusters for theory representations (Figure 3B) and theory updating (Figure 4B). ROIs noted as in Figure 3B.

(B) Peri-event time histograms showing the average change in predictivity score following theory updates and different control events in the overlapping ROIs. Notation as in Figure 4C. Note that, in contrast to Figure 4C, the y axis is Δz, which quantifies how well an encoding model based on theory representations can predict instantaneous patterns of brain activity after a theory update, compared with before the update. See also Figure S8.

(C) Change in predictivity score from (B) averaged over 20 s following corresponding event. Notation as in Figure 4D.

are related in the brain. Specifically, we conjectured that theory representations are preferentially activated during theory updating, akin to being "loaded" into working memory for the necessary update.

To investigate this hypothesis, we plotted PETHs of the baseline-adjusted predictivity time course from the encoding model (Figure 3A) following theory updates and other control events in the ROIs from the overlap. This shows, at each time point after the event, how well the pattern of BOLD activity can be predicted based on the inferred theory, compared with immediately before the event. We found a significant sustained increase in predictivity after theory updates in IFG (triangular and opercular parts), all three occipital gyri (inferior, middle, superior), and FFG (Figure 6B; two-sided t tests). Furthermore, the magnitude of this increase was significantly greater for theory updates compared with other events (Figure 6C; paired t tests), suggesting that theory representations are activated in these regions specifically during theory updating. Additionally, among a set of *a priori* ROIs thought to be involved in the relational and semantic representations,[33–35] we found a significant effect in parahippocampal cortex (Figures S8C and S8D).

To investigate whether this effect varies between individual theory components, we repeated this analysis for separate component updates using the corresponding encoding models fit for objects, relations, or goals only. We found that most regions did not show a significant difference (Figures S8E and S8F), with the exception of FFG in which object representations were activated after object updates more strongly compared with relation and goal representations during their respective updates ($p < 0.001$, Bonferroni corrected), suggesting a specific role for FFG in object updating.

Effective connectivity during theory updating is consistent with predictive coding

Having identified brain regions involved in theory representation (Figure 3), theory updating (Figures 4 and 5), and the dynamic interplay between these processes (Figure 6), we finally sought to characterize the pattern of information flow between these regions. Using a beta series GLM,[36] we extracted estimates of instantaneous neural activity during theory update events from ROIs that showed a significant effect in the previous analyses. We additionally extracted estimates from visual and motor ROIs in order to include potential inputs and outputs to and

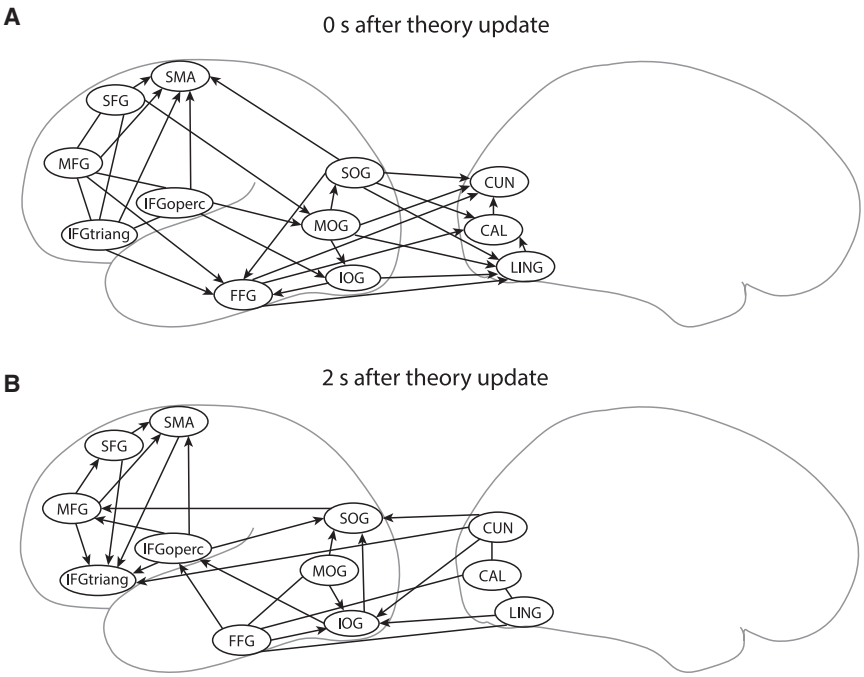

Figure 7. Effective connectivity during theory updating is consistent with predictive coding

(A) Best-fitting effective connectivity pattern based on neural responses to theory update events estimated using beta series GLM. ROIs noted as in Figure 3B.

(B) Same results using neural responses 2 s after theory update events.

from theory-coding and theory-updating regions. We entered the resulting estimates into the independent multiple-sample greedy equivalence search (IMaGES) algorithm[36,37] from the TETRAD software package for causal modeling,[38] which greedily searches the space of effective connectivity patterns for the one that best fits the data. Our hypothesis was that, during theory updating, information would flow in a bottom-up fashion, from early visual regions through theory-updating regions in occipital and temporal cortex to theory-coding regions in prefrontal cortex, where the updated theory is putatively stored.

To the contrary, we found the opposite pattern, with information flowing in a top-down fashion from prefrontal theory-coding regions to theory-updating regions in occipital and temporal cortex to early visual regions (Figure 7A). When we repeated the same analysis, except using neural activity 2 s after theory updates, we found a bottom-up pattern consistent with our prior expectations (Figure 7B). These findings are consistent with a predictive coding interpretation: information about the brain's internal model of the world (in our case, the theory) is flowing top-down from higher areas in prefrontal cortex, shaping sensory predictions in lower visual areas; when an inconsistency between predictions and observations is detected, this results in a theory prediction error that triggers a theory update, reversing the flow of information so that the new sensory data can be used to update the theory in the higher regions.

DISCUSSION

A longstanding question in neuroscience is how the brain represents the structure of the environment in order to support efficient learning and flexible generalization. One possible answer from cognitive science is that the brain learns a rich, abstract, causal model grounded in core cognitive concepts such as objects, relations, and goals, which is used to simulate the outcomes of different courses of action during planning.[15–17,39] We found support for this kind of theory-based RL using fMRI data from human participants learning to play different Atari-style games. The theory inferred by a theory-based RL agent can explain variance in IFG and other prefrontal regions better than control models, suggesting that those regions encode theory-like representations above and beyond visual and model-free RL features. In an overlapping network of regions, including IFG, occipital gyri, and FFG, we found theory learning signals that could not be explained by visual events, motor actions, or theory-related nuisance variables, suggesting those regions play a role in theory inference. In a subset of those regions, we found evidence for separate learning signals for objects, relations, and goals, suggesting that the brain factors theory inference similarly to our theory-based RL agent. We additionally found that the striking overlap between theory-coding and theory-learning regions is not coincidental, with theory representations being activated following theory updates. Finally, we found that the effective connectivity pattern during theory updates is consistent with predictive coding,[18,19] with feedback connections conveying theory predictions and feedforward connections conveying theory prediction errors.

The idea that animals learn rich, structured representations of their environments dates back to Tolman's work on latent learning.[40,41] Tolman observed that rats were able to quickly find newly placed rewards in a maze after repeated unrewarded exposures to the maze, leading him to hypothesize that this flexible generalization is supported by "cognitive maps[42–44]"—internal models of the world which allow animals to mentally search through space and find efficient paths to goals. Neural evidence for cognitive maps was famously identified in the hippocampus,[45] where place cells appear to encode an animal's location in space. Subsequent studies found evidence that cognitive maps can represent nonspatial domains[46,47] and also appear in other parts of the brain,[48–50] such as ventral prefrontal cortex, which includes IFG, a region our study implicates in theory coding. Our study found some evidence of theory representations in parahippocampal cortex but not in the hippocampus. This is likely due to the fact that the theory on its own does not constitute a map per se, but rather a set of abstract relational rules that, when grounded in a particular world state (such as a

video game frame), can be used to predict future world states. We conjecture that the hippocampus and medial entorhinal cortex might be involved in such grounded representations, encoding a theory-based transition structure between concrete world states that directly supports planning, rather than the abstract theory itself.

Our findings resonate with recent studies that have used computational modeling to identify a more specific role for prefrontal cortex in representing and/or updating an internal causal model of the world.[51] In an fMRI study comparing model-based and model-free RL prediction errors, Gläscher et al.[10] reported state prediction error signals—discrepancies between the observed state and the state predicted by the brain's internal model, akin to theory prediction errors in our study—in similar prefrontal regions, particularly in bilateral IFG. Another fMRI study by Lee et al.[52] reported evidence of rapid, one-shot learning of causal associations encoded in ventrolateral prefrontal cortex, including the IFG. An fMRI study of causal structure learning from our lab[31] found causal structure learning signals in a distributed bilateral network of regions, including IFG, MFG, and SFG, regions in occipital cortex, and regions in the ventral stream such as FFG. In that study, we also reported evidence of beliefs about causal structure being activated in response to feedback in a frontoparietal network of regions, including IFG. Another study from our lab[53] also reported evidence of beliefs about causal structure being activated in IFG during belief updating.

A separate line of work has implicated similar prefrontal regions in relational reasoning.[54,55] Knowlton et al.[56] unified some of these findings using a role-based relational reasoning model (LISA), according to which prefrontal cortex encodes abstract relational rules as distributed role-filler bindings at increasing levels of abstraction, from objects to relations to propositions, somewhat reminiscent of our HRR theory code. In LISA, rules are rapidly updated via spike-timing dependent plasticity in the anterior prefrontal cortex and are activated in working memory by long-distance connections from semantic units in posterior cortex. This bears a striking resemblance to our proposal and suggests that theory-based RL could serve as a unifying lens for results from the neuroscience literature on model-based RL, causal inference, and relational reasoning. According to this view, these findings could be interpreted as signatures of the same theory inference machinery applied to different, narrower domains, with IFG serving as the key locus of theory computation/storage in the prefrontal cortex and posterior regions computing theory prediction errors for theory learning.

Video games have long served as microcosms in which to compare human and machine intelligence in naturalistic, complex environments.[2,14] Most closely related to our work is a recent study by Cross et al.[9] in which fMRI data from human participants playing Atari games was analyzed using a deep RL network (DQN), a powerful model-free RL algorithm. The authors found evidence of DQN representations across a distributed network of regions, most notably in the dorsal visual stream and posterior parietal cortex. Despite similar methodology, there are crucial differences between our studies. The most critical difference is that we are interested in how people learn to play these games—an aspect of human behavior that is particularly well-captured by theory-based RL compared with model-free deep RL—whereas Cross et al.[9] are interested in the sensorimotor transformations that support gameplay after learning has plateaued. This in turn dictates important design decisions that differ between the two studies. Most importantly, we focus on games in which—according to our prior work—people's behavior seems to be model-based and, in particular, seems to follow the predictions of theory-based RL, whereas Cross et al.[9] focus on games in which people's behavior follows the predictions of the model-free DQN. As a result, our study includes more games which are played over shorter timescales and have less visually distinct features, more complex rules, and levels designed to maximize learning. This could explain the relatively poor performance of our model-free RL control in matching human performance and brain activity.

However, our results are not mutually exclusive with those of Cross et al.[9] Multiple studies have shown that brains employ a mix of model-free and model-based RL strategies.[10–13] Indeed, the results from Cross et al.[9] point to the dorsal stream, posterior parietal cortex, and motor areas as being the loci of model-free sensorimotor transformations, whereas they report little evidence for model-free representations in prefrontal regions and, in particular, they do not report any results in IFG. In contrast, our results point to prefrontal cortex—and IFG in particular—as the locus of theory encoding, and to occipital and ventral stream regions as the loci of theory learning; at the same time, we find little evidence for theory-based representations in the dorsal stream, posterior parietal cortex, or motor cortex. Thus, the results from the two studies can be seen as complementary, pointing to hybrid architecture that includes both theory-based and model-free components. Although EMPA in its current form is purely model-based, it can straightforwardly be extended to include learned policy and/or value components to help guide the search toward promising action plans. In the field of artificial intelligence, such hybrid approaches have recently achieved remarkable success in learning to play board games[3,4] and video games,[5] suggesting that this could be a fruitful avenue for future neuroscience research.

Although our results and the results of Tsividis et al.[16] cannot be accounted for purely by relatively straightforward deep RL approaches like DDQN, they certainly do not rule out more sophisticated deep RL architectures. For example, deep model-based RL architectures equipped with planning and model learning modules have shown much faster learning and superior performance on Atari games.[5,57] Similar to EMPA, such approaches can learn a model of the environment from scratch and use it to plan efficiently. Alternatively, deep meta-RL approaches use a model-free RL algorithm to learn a model-based RL algorithm.[58,59] Such models could in principle learn theory-like representations or even an EMPA-like theory-based RL algorithm from scratch. However, even if such models were able to capture human behavior and brain activity, their opacity would still leave open the question of what humans are actually learning. In contrast, EMPA and theory-based RL more generally characterize the structure and content of inductive biases and algorithms from the beginning to the end of gameplay, which was the original goal of our work and which other models have not

been able to explain. Beyond that, theory-based RL could be seen as the outcome of another learning process—perhaps spread across both evolution and development—which could in principle be modeled by deep RL.[58,59] We leave this as the subject of future work.

Our model relies on the same theory inference machinery for all games. This is somewhat at odds with the finding that games which require less planning show weaker theory representations (Figures S4E–S4G). Indeed, there is no way for EMPA to "know" that a game requires less planning until it has already inferred a theory for it and played it for a while. One possible explanation is that the reactive nature of these games prompts an alternative mechanism for generating actions that relies less on the theory. Model-free RL offers one such mechanism,[9] which once again points to the possibility of a hybrid theory-based/model-free architecture, highlighted above, and could be investigated in future work.

Relatedly, our model predicts that theory representations should be persistently active as they continuously inform planning during gameplay. The finding that theory representations are activated preferentially during theory updates is somewhat at odds with this prediction (Figure 6). One possible explanation is that the increased BOLD activity during theory updates results in an increased signal-to-noise ratio, allowing the encoding model to achieve better predictivity. If that could fully account for the effect, we would expect to see transient changes in predictivity for other events that elicit an increased (albeit to a lesser extent) BOLD response in those regions (Figure 4C), something we did not observe (Figure 6B). An alternative explanation is that the theory is not stored as a persistent pattern of neural activity but is rather stored "silently,"[60] perhaps in the pattern of synaptic weights, and is only activated when updated by the theory inference circuitry.

Our effective connectivity analysis suggests that top-down information about the theory from prefrontal regions flows to occipital and ventral stream regions for predicting sensory inputs and that when a discrepancy occurs—a kind of theory prediction error—information flows the other way for updating the theory in prefrontal regions based on sensory input from occipital and ventral stream regions. This is broadly consistent with hierarchical predictive coding[18,19]: the idea that top-down (feedback) connections convey model predictions originating in higher cortical areas that shape neural activity in lower cortical areas, which in turn compute prediction errors that are conveyed to higher areas via bottom-up feedforward connections for model updating. Despite this affinity, there are important differences between our proposal and traditional predictive coding accounts. First, the predictive coding interpretation only pertains to information flow between regions representing the learned theory and regions computing theory prediction errors. Importantly, it does not account for the processes of learning, planning, and exploration, which are core aspects of theory-based RL. Second, predictive coding models are usually employed in narrow domains, often focusing on simple problems of low-level perception[18] or simple RL problems.[61] In contrast, EMPA and theory-based RL more broadly focus on solving richer and more structured problems. Our approach considers perception and inference in the context of a complete modeling, planning, and exploring agent; the models and plans generated by EMPA—and those generated by the brain—have more structure to them than those generated by standard predictive coding approaches. Finally, theory-based RL and predictive coding are frameworks at fundamentally different levels of description[62]: theory-based RL is a computational-level proposal of exploration, modeling, and planning based on Bayesian inference over intuitive theories (with EMPA being a particular algorithmic instantiation of it), whereas predictive coding is an implementation-level proposal of neural coding and dynamics of modeling and perception.[63] Viewed in this light, our results suggest that the general predictive coding framework could be a promising starting point for studying theory predictions, theory prediction errors, and theory updating at the neural level. Future work could formally relate EMPA to particular predictive coding formulations, which could provide a richer theoretical framework for understanding the interplay between top-down and bottom-up inferential processes in the brain, as well as the interplay between model learning, exploration, and planning, relative to current predictive coding models.

One puzzling aspect of our results is the prevalence of visual regions, which raises the possible concern that our analysis was not selective enough to exclude visual confounds. This concern is partly addressed by our control analyses. In our encoding model analysis (Figure 3), we found that EMPA consistently outperformed all of our control models in prefrontal regions, but not in other cortical areas; indeed, in most other regions, EMPA was no better than PCA, suggesting that the theory effects in those areas could be partly explained by visual features. The theory update GLM (Figure 4) included visual nuisance regressors that showed a stronger effect in some regions, particularly in early visual areas, suggesting that those regions play a role in visual processing that is not specific to theory updating. Accordingly, we excluded early visual areas from reporting and follow-up analyses. Theory learning effects in higher visual areas could be partly explained by our effective connectivity results: according to the predictive coding interpretation, it is precisely visual regions that ought to compute theory prediction errors—discrepancies between theory-based predictions and sensory observations—which in turn serve as the basis for updating the theory in prefrontal regions. It is also worth noting that previous work on causal structure learning[31] has also reported evidence for model updating in visual areas. Additionally, to some extent our experimental design already controls for visual confounds by having participants play the same level on repeat for 1 minute. If they do end up playing the same level for multiple episodes, most learning occurs during the first episode(s), with the other episodes serving as implicit controls with nearly identical visual inputs but little-to-no theory learning. This idea could be taken further by having participants watch a replay of their own gameplay immediately after the game or in a subsequent scan session. We leave this kind of control study as future work.

In summary, our results are consistent with a neural architecture of theory-based RL in which theory representations in IFG and other prefrontal regions are activated and updated in response to theory prediction errors computed in occipital and ventral stream regions, such as FFG, in a way consistent with hierarchical predictive coding. Additionally, we hope that our work

highlights the benefits of combining sophisticated, interpretable, end-to-end cognitive models such as EMPA with naturalistic experimental environments such as video games. By comparing the internal representations of such models with brain activity, researchers can begin to uncover how the brain learns and represents an internal model of the environment that supports adaptive behavior in complex, naturalistic tasks.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Experimental Design
  - fMRI Data Acquisition
  - fMRI Preprocessing
  - EMPA
  - DDQN
  - Generative play
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Human and model behavior
  - Encoding model analysis
  - Gaussian Process regression
  - Holographic reduced representations
  - Control models
  - GLM analyses
  - GLM comparison
  - Theory activation timecourse
  - Effective connectivity

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.01.023.

ACKNOWLEDGMENTS

This research was supported by the Toyota Corporation, the Center for Brains, Minds, and Machines (CBMM), funded by NSF STC award CCF1231216, and the Multi-University Research Initiative Grant (ONR/DoD N00014-17-1-2961). This work involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program award number S10OD020039. We acknowledge the University of Minnesota Center for Magnetic Resonance Research for use of the multiband-echo-planar imaging (EPI) pulse sequences. We are grateful to Alicia Chen, Yichen Li, Zhenglong Zhou, Daphne Cornelisse, Jiajia Zhao, Chelsea Guglielmi, Dimitar Karev, and Jason Ma for their help with data collection and initial prototyping.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S.T., P.A.T., T.P., J.B.T., and S.J.G.; data curation, M.S.T.; formal analysis, M.S.T.; funding acquisition, S.J.G.; investigation methodology, M.S.T., P.A.T., T.P., J.B.T., and S.J.G.; project administration, M.S.T.; software, M.S.T., P.A.T., and S.J.G., supervision, S.J.G.; validation, M.S.T.; visualization, M.S.T.; writing – original draft, M.S.T., P.A.T., T.P., J.B.T., and S.J.G.; writing – review and editing; M.S.T., P.A.T., T.P., J.B.T., and S.J.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 14, 2022
Revised: November 6, 2022
Accepted: January 27, 2023
Published: March 9, 2023

REFERENCES

1. Sutton, R., and Barto, A. (2018). Reinforcement Learning: An Introduction, Second Edition (MIT Press). https://doi.org/10.1109/TNN.1998.712192.
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533. https://doi.org/10.1038/nature14236.
3. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. Nature *550*, 354–359. https://doi.org/10.1038/nature24270.
4. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science *362*, 1140–1144. https://doi.org/10.1126/science.aar6404.
5. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. Nature *588*, 604–609. https://doi.org/10.1038/s41586-020-03051-4.
6. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
7. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879. https://doi.org/10.1038/nature04766.
8. Niv, Y. (2009). Reinforcement learning in the brain. J. Math. Psychol. *53*, 139–154. https://doi.org/10.1016/j.jmp.2008.12.005.
9. Cross, L., Cockburn, J., Yue, Y., and O'Doherty, J.P. (2021). Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. Neuron *109*, 724–738.e7. https://doi.org/10.1016/j.neuron.2020.11.021.
10. Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595. https://doi.org/10.1016/j.neuron.2010.04.016.
11. Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027.
12. Lee, S.W., Shimojo, S., and O'Doherty, J.P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron *81*, 687–699. https://doi.org/10.1016/j.neuron.2013.11.028.
13. Kool, W., Cushman, F.A., and Gershman, S.J. (2018). Chapter 7. Competition and cooperation between multiple reinforcement learning systems. In Goal-Directed Decision Making, R. Morris, A. Bornstein, and A. Shenhav, eds. (Academic Press), pp. 153–178. https://doi.org/10.1016/B978-0-12-812098-9.00007-3.

14. Tsividis, P.A., Pouncy, T., Xu, J.L., Tenenbaum, J.B., and Gershman, S.J. (2017). Human learning in atari. 2017 AAAI Spring Symposium Series, Science of Intelligence: Computational Principles of Natural and Artificial Intelligence, Technical Report SS-17-07 (Association for the Advancement of Artificial Intelligence).

15. Pouncy, T., Tsividis, P., and Gershman, S.J. (2021). What is the model in model-based planning? Cogn. Sci. *45*, e12928. https://doi.org/10.1111/cogs.12928.

16. Tsividis, P.A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., Gershman, S.J., and Tenenbaum, J.B. (2021). Human-level reinforcement learning through theory-based modeling, exploration, and planning. arXiv. https://doi.org/10.48550/arXiv.2107.12544.

17. Pouncy, T., and Gershman, S.J. (2022). Inductive biases in theory-based reinforcement learning. Cogn. Psychol. *138*, 101509. https://doi.org/10.1016/j.cogpsych.2022.101509.

18. Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

19. Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836. https://doi.org/10.1098/rstb.2005.1622.

20. van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. Proceedings of the AAAI Conference on Artificial Intelligence *30*, 2094–2100. https://doi.org/10.1609/aaai.v30i1.10295.

21. Seeger, M. (2004). Gaussian processes for machine learning. Int. J. Neural Syst. *14*, 69–106. https://doi.org/10.1142/S0129065704001899.

22. Plate, T.A. (1995). Holographic reduced representations. IEEE Trans. Neural Netw. *6*, 623–641. https://doi.org/10.1109/72.377968.

23. Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., and Fedorenko, E. (2021). The neural architecture of language: integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. USA *118*. e2105646118. https://doi.org/10.1073/pnas.2105646118.

24. Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609. https://doi.org/10.1038/381607a0.

25. Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14. https://doi.org/10.1016/j.cell.2017.05.011.

26. Mohamed, S., and Jimenez Rezende, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. In NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds. (Curran Associates, Inc.), pp. 2125–2133.

27. Watter, M., Springenberg, J., Boedecker, J., and Riedmiller, M. (2015). Embed to control: a locally linear latent dynamics model for control from raw images. In Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds. (Curran Associates, Inc.).

28. Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. (2017). DARLA: improving zero-shot transfer in reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y.W. Teh, eds. (PMLR), pp. 1480–1490.

29. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., and Joliot, M. (2020). Automated anatomical labelling atlas 3. NeuroImage *206*, 116189. https://doi.org/10.1016/j.neuroimage.2019.116189.

30. Mahon, B.Z., Milleville, S.C., Negri, G.A., Rumiati, R.I., Caramazza, A., and Martin, A. (2007). Action-related properties shape object representations in the ventral stream. Neuron *55*, 507–520. https://doi.org/10.1016/j.neuron.2007.07.011.

31. Tomov, M.S., Dorfman, H.M., and Gershman, S.J. (2018). Neural computations underlying causal structure learning. J. Neurosci. *38*, 7143–7157. https://doi.org/10.1523/JNEUROSCI.3336-17.2018.

32. Rigoux, L., Stephan, K.E., Friston, K.J., and Daunizeau, J. (2014). Bayesian model selection for group studies — revisited. NeuroImage *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065.

33. Epstein, R.A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn. Sci. *12*, 388–396. https://doi.org/10.1016/j.tics.2008.07.004.

34. Bonner, M.F., and Price, A.R. (2013). Where is the anterior temporal lobe and what does it do? J. Neurosci. *33*, 4213–4215. https://doi.org/10.1523/JNEUROSCI.0041-13.2013.

35. Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat. Neurosci. *20*, 1643–1653. https://doi.org/10.1038/nn.4650.

36. Poldrack, R., Mumford, J., and Nichols, T. (2011). Handbook of Functional MRI Data Analysis (Cambridge University Press).

37. Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., and Glymour, C. (2010). Six problems for causal inference from fmri. NeuroImage *49*, 1545–1558. https://doi.org/10.1016/j.neuroimage.2009.08.065.

38. Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. (1998). The tetrad project: constraint based aids to causal model specification. Multivariate Behav. Res. *33*, 65–117. https://doi.org/10.1207/s15327906mbr3301_3.

39. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253. https://doi.org/10.1017/S0140525X16001837.

40. Tolman, E.C., and Honzik, C.H. (1930). Introduction and removal of reward, and maze performance in rats. Univ. Calif. Publ. Physiol. *4*, 257–275.

41. Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208. https://doi.org/10.1037/h0061626.

42. Kaplan, R., Schuck, N.W., and Doeller, C.F. (2017). The role of mental maps in decision-making. Trends Neurosci. *40*, 256–259. https://doi.org/10.1016/j.tins.2017.03.002.

43. Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. Neuron *100*, 490–509. https://doi.org/10.1016/j.neuron.2018.10.002.

44. Boorman, E.D., Sweigart, S.C., and Park, S.A. (2021). Cognitive maps and novel inferences: a flexibility hierarchy. Curr. Opin. Behav. Sci. *38*, 141–149. https://doi.org/10.1016/j.cobeha.2021.02.017.

45. O'Keefe, J., and Nadel, L. (1978). The Hippocampus as a Cognitive Map (Oxford University Press).

46. Schuck, N.W., Cai, M.B., Wilson, R.C., and Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. Neuron *91*, 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019.

47. Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science *352*, 1464–1468. https://doi.org/10.1126/science.aaf0941.

48. Walton, M.E., Behrens, T.E., Buckley, M.J., Rudebeck, P.H., and Rushworth, M.F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron *65*, 927–939. https://doi.org/10.1016/j.neuron.2010.02.027.

49. Rudebeck, P.H., and Murray, E.A. (2011). Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. J. Neurosci. *31*, 10569–10578. https://doi.org/10.1523/JNEUROSCI.0091-11.2011.

50. Jocham, G., Brodersen, K.H., Constantinescu, A.O., Kahn, M.C., Ianni, A.M., Walton, M.E., Rushworth, M.F., and Behrens, T.E. (2016). Reward-guided learning with and without causal attribution. Neuron *90*, 177–190. https://doi.org/10.1016/j.neuron.2016.02.018.

51. Donoso, M., Collins, A.G.E., and Koechlin, E. (2014). Human cognition. Foundations of human reasoning in the prefrontal cortex. Science *344*, 1481–1486. https://doi.org/10.1126/science.1252254.

52. Lee, S.W., O'Doherty, J.P., and Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. PLoS Biol. *13*, e1002137. https://doi.org/10.1371/journal.pbio.1002137.

53. Dorfman, H.M., Tomov, M.S., Cheung, B., Clarke, D., Gershman, S.J., and Hughes, B.L. (2021). Causal inference gates corticostriatal learning. J. Neurosci. *41*, 6892–6904. https://doi.org/10.1523/JNEUROSCI.2796-20.2021.

54. Waltz, J.A., Knowlton, B.J., Holyoak, K.J., Boone, K.B., Mishkin, F.S., de Menezes Santos, M., Thomas, C.R., and Miller, B.L. (1999). A system for relational reasoning in human prefrontal cortex. Psychol. Sci. *10*, 119–125. https://doi.org/10.1111/1467-9280.00118.

55. Krawczyk, D.C., Michelle McClelland, M., and Donovan, C.M. (2011). A hierarchy for relational reasoning in the prefrontal cortex. Cortex *47*, 588–597. https://doi.org/10.1016/j.cortex.2010.04.008.

56. Knowlton, B.J., Morrison, R.G., Hummel, J.E., and Holyoak, K.J. (2012). A neurocomputational system for relational reasoning. Trends Cogn. Sci. *16*, 373–381. https://doi.org/10.1016/j.tics.2012.06.002.

57. Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. (2021). Mastering atari games with limited data. In Advances in Neural Information Processing Systems, M. Syst, A. Ranzato, Y. Beygelzimer, P.S. Dauphin, J. Liang, and V. Wortman, eds. (Curran Associates, Inc.), pp. 25476–25488.

58. Duan, Y., Schulman, J., Chen, X., Bartlett, P.L., Sutskever, I., and Abbeel, P. (2016). $R1^2$: fast reinforcement learning via slow reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.1611.02779.

59. Wang, J.X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2016). Learning to reinforcement learn. arXiv. https://doi.org/10.48550/ARXIV.1611.05763.

60. Beukers, A.O., Buschman, T.J., Cohen, J.D., and Norman, K.A. (2021). Is activity silent working memory simply episodic memory? Trends Cogn. Sci. *25*, 284–293. https://doi.org/10.1016/j.tics.2021.01.003.

61. Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., and Dolan, R.J. (2013). The anatomy of choice: active inference and agency. Front. Hum. Neurosci. *7*, 598. https://doi.org/10.3389/fnhum.2013.00598.

62. Marr, D., and Poggio, T. (1976). From understanding computation to understanding neural circuitry. Artificial intelligence laboratory. A.I. Memo. http://hdl.handle.net/1721.1/5782.

63. Aitchison, L., and Lengyel, M. (2017). With or without you: predictive coding and bayesian inference in the brain. Curr. Opin. Neurobiol. *46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010.

64. Rosa, M.J., Bestmann, S., Harrison, L., and Penny, W. (2010). Bayesian model selection maps for group studies. NeuroImage *49*, 217–224. https://doi.org/10.1016/j.neuroimage.2009.08.051.

65. Perez-Liebana, D., Samothrakis, S., Togelius, J., Schaul, T., Lucas, S.M., Couëtoux, A., Lee, J., Lim, C.U., and Thompson, T. (2016). The 2014 general video game playing competition. IEEE Trans. Comput. Intell. AI Games *8*, 229–243. https://doi.org/10.1109/TCIAIG.2015.2402393.

66. Schaul, T. (2013). A video game description language for model-based or interactive learning. In IEEE Trans. Comput. Intell. AI Games, pp. 1–8. https://doi.org/10.1109/CIG.2013.6633610.

67. Tomov, M.S., Truong, V.Q., Hundia, R.A., and Gershman, S.J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. Nat. Commun. *11*, 2371. https://doi.org/10.1038/s41467-020-15766-z.

68. van der Kouwe, A.J.W., Benner, T., Salat, D.H., and Fischl, B. (2008). Brain morphometry with multiecho MPRage. NeuroImage *40*, 559–569. https://doi.org/10.1016/j.neuroimage.2007.12.025.

69. Moeller, S., Yacoub, E., Olman, C.A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. Magn. Reson. Med. *63*, 1144–1153. https://doi.org/10.1002/mrm.22361.

70. Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Glasser, M.F., Miller, K.L., Ugurbil, K., and Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. PLoS One *5*, e15710. https://doi.org/10.1371/journal.pone.0015710.

71. Xu, J., Moeller, S., Auerbach, E.J., Strupp, J., Smith, S.M., Feinberg, D.A., Yacoub, E., and Uğurbil, K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3 T. NeuroImage *83*, 991–1001. https://doi.org/10.1016/j.neuroimage.2013.07.055.

72. Carey, S. (2000). The origin of concepts. J. Cogn. Dev. *1*, 37–41. https://doi.org/10.1207/S15327647JCD0101N_3.

73. Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fmri. NeuroImage *56*, 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073.

74. Güçlü, U., and van Gerven, M.A.J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. *35*, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015.

75. Driscoll, L.N., Pettit, N.L., Minderer, M., Chettih, S.N., and Harvey, C.D. (2017). Dynamic reorganization of neuronal activity patterns in parietal cortex. Cell *170*, 986–999.e16. https://doi.org/10.1016/j.cell.2017.07.021.

76. Gayler, R.W. (2004). Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. arXiv. https://doi.org/10.48550/arXiv.cs/0412059.

77. Franklin, N.T., Norman, K.A., Ranganath, C., Zacks, J.M., and Gershman, S.J. (2020). Structured event memory: a neuro-symbolic model of event cognition. Psychol. Rev. *127*, 327–361. https://doi.org/10.1037/rev0000177.

78. Spirtes, P. (2005). Graphical models, causal inference, and econometric models. J. Econ. Methodol. *12*, 3–34. https://doi.org/10.1080/1350178042000330887.

79. Meek, C. (1997). Graphical Models: Selecting causal and statistical models. Ph.D. thesis (Carnegie Mellon University).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
fMRI and behavioral data	this paper	https://doi.org/10.18112/openneuro.ds004323.v1.0.0
Software and algorithms		
MATLAB R2022b	MathWorks	https://www.mathworks.com/
SPM 12	Rosa et al.[64]	https://www.fil.ion.ucl.ac.uk/spm/software/spm12/
CCNL fMRI	Samuel Gershman	https://github.com/sjgershm/ccnl-fmri
EMPA, fMRI task, fMRI regressor generation	Tsividis et al.[16]; this paper	https://github.com/tsividis/vgdl/tree/refactor_fMRI_cannon
DDQN	van Hasselt et al.[20]; Tsividis et al.[16]; this paper	https://github.com/tomov/RC_RL/tree/fmri
data analysis code	this paper	https://github.com/tomov/VGDL-fMRI-Data-Analysis

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Momchil Tomov (mtomov@g.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- De-identified human behavioral and functional MRI data have been deposited at OpenNeuro and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- Task and analysis code is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Thirty-two healthy participants were recruited from the Cambridge, MA community: 15 female, 17 male, 19-36 years of age, mean age 24 ± 4 years, all right-handed and with normal or corrected-to-normal vision. The study was approved by the Harvard University Institutional Review Board and all participants gave informed consent. All participants were paid for their participation.

METHOD DETAILS

Experimental Design

Each participant played 6 different Atari-style games adapted from Tsividis et al.[16] over the course of 6 scanner runs in a single session (Figure 2B). Six games were played across 6 scanner runs. Each run consisted of 3 blocks. Each block consisted of 3 levels of a given game. Each level was played on repeat for 1 minute total: if the episode ended before 1 minute had elapsed, a new episode began on the same level. Nine levels were played in total for each game. Scanner runs were grouped in 3 data partitions for cross-validation. Game order was pseudo-randomized such that each data partition contained one block of each game, ensuring that games and levels were balanced across partitions.

For each participant, games were randomly assigned names that were unrelated to the game rules (Archeplan, Deception Eagle, Dreams of Origins, Giants of Solitude, Questtide, Fuseville, Prime Origin). At the beginning of each block, the game name was shown for 2 s (Figure 2A). During an episode, the game name and the current score were displayed at the top of the screen. At the end of an

episode, the outcome ("You WON!" or "You LOST!") was shown at the bottom of the screen and the final frame was frozen for 2 s. Timing was adjusted such that each level was played for one minute total. After one minute, the current episode was interrupted with a "End of level" outcome (to distinguish it from a win or loss) shown for 2 s, unless the participant was already on a win/loss screen. There was a 10-s fixation cross at the beginning and end of each run to account for scanner stabilization and the hemodynamic lag, respectively. Each run was 566 s in total.

Following Tsividis et al.,[16] in order to avoid biasing learning with semantic priors based on object appearance, all games were played in "color-only" mode: all objects were visualized as colored squares with symbols on them. Objects of the same kind had the same color and symbol, while objects of different kinds had different colors and symbols. Color and symbol assignments were randomized across games and participants. The game descriptions were inspired by and/or drawn from the General Video Game AI (GVGAI) competition[65] and expressed in the Video Game Description Language[66] (VGDL). Participants 1 through 11 played Chase, Helper, Bait, Zelda, Lemmings, Plaque Attack. Participants 12 through 32 played the same games, except for Plaque Attack which was replaced by Avoid George. These games are a subset of the games used by Tsividis et al.[16] and, in choosing them, we aimed to cover a large and heterogeneous space in order to demonstrate the flexibility of human gameplay behavior and the versatility of our model. All games were fully observable, i.e. no memory of past states is required to win. Each game had 5 actions: move left, move up, move down, move right, action key. The levels were designed to ensure continuous learning about the game rules. Specifically, different levels involved different object layouts and later levels occasionally introduced opportunities to learn about game rules that were not available in earlier levels. Game descriptions, winning strategies, and example screenshots are shown in Table S4. Level descriptions are available at https://github.com/tomov/RC_RL/tree/fmri/fmri_all_games.

Participants were told that they would be playing a sequence of Atari-style games with different rules and that they will have to learn the rules of each game from experience. The game and level order and timing was explained to them (Figure 2B), as well as that they would be playing all games in color-only mode and what that is. Specifically, they were told that the colors, symbols, and game names convey no information about the game rules, except that objects of the same kind look the same in a given game. They were also told that colors and symbols in one game convey no information about objects in another game. All participants were paid a base of $80 for their participation. Additionally, to incentivize learning, we paid participants a bonus based on performance. Specifically, for each participant, we randomly chose a level and paid them the maximum score they achieved (in dollars) at the end of any episode on that level, counting only episodes which they won. If they never won that level, the bonus was $0. This bonus scheme was explained to them in detail. They were also told that it is meant to encourage efficient learning and gameplay: they should aim to maximize the score and win each level within 1 minute.

In the scanner, participants played using a 5-finger button box, with each button corresponding to a game action (index finger = move left, middle finger = move up, ring finger = move down, pinky finger = move right, thumb = action key). Before entering the scanner, participants practiced by playing 3 levels (1 block) of a different game (Sokoban) on the laptop using a similar key setup. This game was not played in the scanner. Overall, the entire scan session took 2.5 hrs per participant, 1.5 hrs of which was spent in the scanner, 1 hr of which was spent on BOLD acquisition and gameplay.

fMRI Data Acquisition

We followed a similar protocol to our previous work.[67] All participants were scanned using a 3T Siemens Magnetom Prisma MRI scanner with the vendor 32-channel head coil (Siemens Healthcare, Erlangen, Germany) at the Harvard University Center for Brain Science Neuroimaging. A T1-weighted high-resolution multi-echo magnetization-prepared rapid-acquisition gradient echo (ME-MPRAGE) anatomical scan[68] of the whole brain was acquired for each participant prior to any functional scanning: 176 sagittal slices, voxel size = 1.0×1.0×1.0 mm, TR = 2530 ms, TE = 1.69–7.27 ms, TI = 1100 ms, flip angle = 7°, FOV = 256 mm. Functional images were acquired using a T2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition.[69–71] We collected 6 functional runs for each participant, each with 283 timepoints (Figure 2B). Scan parameters: 87 interleaved axial-oblique slices per whole-brain volume, voxel size = 1.7×1.7×1.7 mm, TR = 2000 ms, TE = 30 ms, flip angle = 80°, in-plane acceleration (GRAPPA) factor = 2, multi-band acceleration factor = 3, FOV = 211 mm. Functional slices were oriented to a 25° tilt towards coronal from AC-PC alignment. The SMS-EPI acquisitions used the CMRR-MB pulse sequence from the University of Minnesota.

All 32 participants were included in the analysis. Scanner runs with excessive motion (> 3 mm translation or > 3° rotation) were excluded from the analysis.

fMRI Preprocessing

Following our previous work,[67] we preprocessed functional images using the SPM12 MATLAB toolbox (Wellcome Department of Imaging Neuroscience, London, UK). Each functional scan was realigned to correct for small movements between scans, producing an aligned set of images and a mean image for each participant. The high-resolution T1-weighted ME-MPRAGE images were then co-registered to the mean realigned images and the gray matter was segmented out and normalized to the gray matter of a standard Montreal Neurological Institute (MNI) reference brain. The functional images were then normalized to the MNI template (resampled voxel size 2 mm isotropic), spatially smoothed with a 8-mm full-width at half-maximum (FWHM) Gaussian kernel, high-pass filtered at 1/ 128 Hz, and corrected for temporal autocorrelations using a first-order autoregressive model.

EMPA

A detailed technical description of EMPA can be found in Tsividis et al.,[16] which we summarize here. EMPA learns a model (or theory), θ, of each game expressed in VGDL.[66] VGDL breaks down the game rules into three different components corresponding to core aspects of human intuitive theories[39,72]: objects (sprites), relations (interactions) between objects, and goals.

A VGDL game description consists of a SpriteSet, θ_S, which specifies the type, appearance, and dynamic properties each object (e.g., "red objects chase the avatar at a speed of 3 squares per second"); an InteractionSet, θ_I, which specifies what happens when two objects interact (e.g., "when a red object collides with the avatar, the avatar dies"); and a TerminationSet, θ_T, which specifies the win/loss conditions of the game (e.g., "when the avatar dies, the game is lost with a score of 0"). A VGDL description thus procedurally defines a Markov Decision Process: the state at every timestep is described by the object instances and locations, the avatar's internal state, and any events due to collisions between pairs of objects; the transition function is defined by the SpriteSet, the InteractionSet, and the TerminationSet; and the reward function is defined by the InteractionSet and the TerminationSet.

EMPA learns the rules of each game by inferring a probability distribution over the space of possible VGDL theories, Θ, from experience using Bayesian inference:

$$p(\theta|s_{1:T}, a_{1:T-1}) \propto p(s_{1:T}|\theta, a_{1:T-1})p(\theta), \quad \text{(Equation 1)}$$

where $\theta = (\theta_S, \theta_I, \theta_T)$ is the inferred theory describing the game rules, T is the current timestep, $s_{1:T}$ is the history of observed states, $a_{1:T-1}$ is the history of avatar actions, and $p(\theta)$ is a minimum description length prior favoring simpler theories.

To choose actions, EMPA uses the maximum *a posteriori* theory, θ^*, together with a simulation-based planner that searches for action sequences that lead to rewarding outcomes under θ^*. Specifically, EMPA pursues exploitative goals that lead to winning (according to θ_T^*), as well as exploratory goals that reduce the uncertainty in θ (e.g., inducing an unobserved collision). Pursuit of these sparse goals is aided by subgoals, which represent partial progress towards goals (e.g., "3 blue objects remaining"), and goal gradients, which represent preferences for states that are spatially closer to achieving a subgoal (e.g., "the closest blue object is 3 squares away"). Planning is further aided by state pruning and re-planning based on prediction errors, as described in Tsividis et al.[16]

In our study, we used the same EMPA parameters and settings as those in Tsividis et al.[16] The code for EMPA will be available at https://github.com/tsividis/vgdl upon publication.

To investigate the contribution of different EMPA components to behavior, we additionally performed three ablations from Tsividis et al.[16] (Figure S1A):

- no intrinsic rewards – no subgoals or goal gradients, leaving the planner to rely only on the sparse environmental rewards,
- no iterative width – the planner cannot rely on the iterative width heuristic, which prunes states that are similar to already visited states and greatly ameliorates the combinatorial explosion associated with longer plans,
- ϵ-greedy – theory-driven exploration favoring novel interactions is replaced with ε-greedy exploration ($\epsilon = 0.1$).

DDQN

Following Tsividis et al.,[16] as a control model we trained a deep reinforcement learning network (DDQN) based on the public repository https://github.com/dxyang/DQN_pytorch with parameter settings $\alpha = 0.00025, \gamma = 0.999, \tau = 100$, experience-replay max = 50,000, batch size = 32, and image input recrop size = 64 × 64 × 3. The exploration parameter, ϵ, was annealed linearly from 1 to 0.1 using a decay rate of 200 steps. Following,[2] the DDQN had 3 convolutional layers (conv1: 32 filters with size = 8×8 and stride = 4; conv2: 64 filters with size = 4×4 and stride = 2; conv3: 64 filters with size = 3×3 and stride = 1), followed by a fully connected layer (linear1: 512 units), followed by the output layer (linear2: 5 units). Each convolutional layer was followed by batch normalization and linear rectification (ReLU). ReLU units also followed the fully connected layer. The input was a 64×64×3 scaled game frame with 3 color channels (RGB). To ensure a fair comparison with EMPA, we pretrained a separate DDQN for each game using a VGDL environment for 100 epochs of 250,000 steps. Levels were alternated across epochs to ensure exposure to all levels. Specifically, in each epoch, the DDQN was trained on a given level for one or more episodes, restarting the level if it was won or lost. During epoch 1, we trained on level 1, during epoch 2, we trained on level 2, and so on, starting over from level 1 after level 9. We used the same pretrained DDQNs for both the behavioral and the neural analyses.

The code for the DDQN is available at https://github.com/tomov/RC_RL.

Generative play

To compare human performance with EMPA and DDQN performance, we valuated the models on the same games and levels as the human participants. We simulated each participant with EMPA by having a separate EMPA instance play all levels of each game generatively, in order. As with human participants, each level was played for 1200 frames (60 sec at 20 Hz), restarting the level if won or lost before that. Similarly to humans, performance was evaluated based on the expected bonus payout, namely the maximum per-level winning score, averaged across all levels and games. We simulated 32 participants independently, each simulation corresponding to a single human participant. We similarly simulated 32 participants with the pretrained DDQNs. Note that, unlike the DDQNs, EMPA does not require pretraining.

QUANTIFICATION AND STATISTICAL ANALYSIS

Human and model behavior

We compared human and model generative performance using two-sided Wilcoxon rank sum tests based on actual (for participants) and simulated (for models) expected bonus payouts (Figure 2C). To compare human and model learning, we fitted a second-degree polynomial (no intercept) to the average learning curve (Figure S1A) and compared the resulting linear coefficients for humans and models using two-sample t-tests.

Encoding model analysis

To compare EMPA theories to brain activations, we used an encoding model[9,73,74] that maps EMPA theory embeddings to BOLD signal (Figure 3A). For each participant, we first replayed the sequence of states, actions, and rewards from their gameplay in the scanner through EMPA, using a separate EMPA instance for each game. This produced an EMPA theory for each frame, corresponding to the theory that EMPA would have inferred at that timepoint if it had observed the same sequence of events as the participant (Figure S1D). We embedded each theory in a vector space using holographic reduced representations (HRRs; see below), resulting in a sequence of HRR embeddings. To account for the stochasticity inherent in HRRs, we independently generated 100 such sequences, each with a different random HRR initialization. Each sequence was convolved with the canonical hemodynamic response function from SPM and subsampled at the scanner frequency (TR = 2 s, or 0.5 Hz).

For each voxel, we predicted the BOLD signal with Gaussian process (GP) regression (see below) using cross-validation across the 3 data partitions (Figure 2B). We quantified accuracy by correlating the predicted with the actual BOLD signal for each partition, averaging the resulting Pearson correlation coefficients across partitions, and Fisher z-transforming the result to obtain a single predictivity score z for that voxel. To aggregate across participants, we performed a two-sided t-test against 0 across participants for each voxel, producing a group-level statistical map (t-map). Following our previous work,[67] we thresholded single voxels at $p < 0.001$ and applied cluster family-wise error (FWE) correction at significance level $\alpha = 0.05$. We visualized the corrected t-maps using the bspmview toolbox in MATLAB.

Anatomical regions of interest (ROIs) were extracted by cross-referencing the peak voxels in each cluster (up to 3 peaks peaks per cluster, minimum 20 voxels apart) with the automated anatomical labeling atlas[29] (AAL3 atlas). Confirmatory ROI analyses were performed using bilateral anatomical ROIs from all models (see Control models below). In a given ROI, for each participant we computed the fraction of significant voxels as the number of voxels with a significant Pearson correlation at the $\alpha = 0.05$ significance level, divided by the total number of voxels in the ROI. We compared models in each ROI using Wilcoxon signed rank tests across participants. To aggregate ROIs into ROI groups (Figures S4C and S4D), we simply merged ROIs from a given cortical region into a single "macro-ROI" and performed the same analysis.

We similarly applied GP regression with our control models.

For the within-games model comparison (Figures S4D–S4G), we repeated this analysis separately for each game, only using the BOLD signal from TRs corresponding to that game. To aggregate across games, we averaged the fraction of significant voxels across games for each participant.

To look for differences between games (Figures S4E–S4G), we designated games that involve reasoning sequentially over multiple kinds of interactions (e.g., picking up a key to unlock a door to reach a goal, as in Bait; pushing an object into another object to destroy it, as in Helper; destroying objects so that other agents can reach a goal, as in Lemmings) as requiring more planning, and the rest as requiring less planning.

Gaussian Process regression

For the encoding model we used Gaussian process (GP) regression,[21] a nonparametric method for predicting values of unseen data points based on similarity with observed data points. Ridge regression – a more commonly used encoding model[9,75] – can be derived as a special case of GP regression. However, unlike ridge regression, GP regression avoids the need to fit weights to individual HRR components (which by design are random) and allows for straightforwardly accounting for the randomness of HRRs.

To justify the use of GP regression, first consider the standard general linear model (GLM) formulation:

$$y = f(\theta) + \epsilon, \quad \text{(Equation 2)}$$

$$f(\theta) = \varphi(\theta)^\top \mathbf{w} = \mathbf{x}^\top \mathbf{w}, \quad \text{(Equation 3)}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2), \quad \text{(Equation 4)}$$

where y is the neural signal at a given time point, θ is the EMPA theory, $\varphi(\theta) = \mathbf{x}$ is the HRR embedding of θ, $\mathbf{w}$ are the component weights (often referred to as beta coefficients), and ϵ is Gaussian noise with zero mean and variance σ_n^2. Such GLMs are routinely used to fit brain data and the resulting weights $\mathbf{w}$ – often fit using maximum likelihood estimation – are used to interpret whether a given feature is represented in brain activity.

High-dimensional feature spaces pose a challenge to this approach, as the weights might be underconstrained. One way around this is to impose a prior distribution on the weights:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_w}), \quad \text{(Equation 5)}$$

where $\mathbf{\Sigma_w}$ is the prior weight covariance matrix. The maximum *a posteriori* solution to this Bayesian linear regression problem is equivalent to ridge regression, where a regularization term that constrains the weights arises naturally from the weight prior.

The challenge with applying ridge regression is that HRR embeddings are random, which 1) renders the weights meaningless, and 2) necessitates averaging over that randomness. These issues can both be addressed by GP regression. First, the predicted neural signal $\hat{y}_*$ for a theory θ_* can be directly computed in closed form from the training data $\boldsymbol{\theta}, \mathbf{y}$,[21] bypassing the need to compute the weights:

$$\hat{y}_* | \theta_*, \boldsymbol{\theta}, \mathbf{y} = \frac{1}{\sigma_n^2} \mathbf{x}_*^\top A^{-1} \mathbf{X} \mathbf{y}, \quad \text{(Equation 6)}$$

$$A = \sigma^{-2} \mathbf{X}\mathbf{X}^\top + \mathbf{\Sigma_w}^{-1}, \quad \text{(Equation 7)}$$

where $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \ldots]^\top$ and $\mathbf{y} = [y_1, y_2, y_3, \ldots]^\top$ are the training theory and neural activation sequences, respectively, θ_* and y_* are the held-out theory and neural activation, respectively, $\mathbf{x}_* = \varphi(\theta_*)$ is the HRR embedding of the held-out theory, and r$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots] = [\varphi(\theta_1), \varphi(\theta_2), \varphi(\theta_3), \ldots]$ is the training data design matrix (Figure S2A). Note that we are only using the posterior means and omitting the variances for ease of exposition.

This can be further rearranged by applying the "kernel trick[21]", resulting in GP regression:

$$\hat{y}_* | \theta_*, \boldsymbol{\theta}, \boldsymbol{y} = \mathbf{k}_*^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y}, \quad \text{(Equation 8)}$$

$$\mathbf{k}_* = \mathbf{X}^\top \mathbf{\Sigma_w} \mathbf{x}_*, \quad \text{(Equation 9)}$$

$$K = \mathbf{X}^\top \mathbf{\Sigma_w} \mathbf{X}. \quad \text{(Equation 10)}$$

Here, the covariance matrix (or kernel) K quantifies the similarity between every pair of theories in the training data (Figure S2B), and the covariance vector $\mathbf{k}_*$ quantifies the similarity between every training theory and the held-out theory. In our case, they were computed based on the HRR design matrix $\mathbf{X}$, but in principle we could use a similarity metric that does not rely on explicitly computed features. We used $\mathbf{\Sigma_w} = I$, so our similarity metric for each pair of theories was effectively the dot product of their HRR embeddings. Intuitively, Equation 8 says that the predicted held-out neural activation is the average of the training neural activations, weighted by the similarity between the corresponding training theories and the held-out theory.

Finally, we can account for the randomness of HRRs by marginalizing over different HRR embedding functions φ resulting from different HRR initializations:

$$\hat{y}_* | \theta_*, \boldsymbol{\theta}, \mathbf{y} = \int_\varphi \mathbf{k}_*^\top \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y} p(\varphi) d\varphi. \quad \text{(Equation 11)}$$

From the central limit theorem and the stochasticity of HRRs, the resulting distributions of K and $\mathbf{k}_*$ are approximately Gaussian, so we chose to simplify further by approximating them using Dirac delta functions around their means, $\overline{K} = \mathbb{E}_\varphi(K)$ and $\overline{\mathbf{k}}_* = \mathbb{E}_\varphi(\mathbf{k}_*)$, yielding the final GP formulation that we used:

$$\hat{y}_* | \theta_*, \boldsymbol{\theta}, \mathbf{y} \approx \overline{\mathbf{k}}_*^\top \left(\overline{K} + \sigma_n^2 I\right)^{-1} \mathbf{y}. \quad \text{(Equation 12)}$$

We used a sampling approximation for $\overline{K}$ and $\overline{\mathbf{k}_*}$ by averaging over the kernels for 100 different HRR initializations. In practice, during cross-validation, we had a set of held-out data points $\boldsymbol{\theta}_*, \mathbf{y}_*$ (rather than a single data point) with a corresponding covariance matrix K_* between the training and held-out data points. So for each HRR initialization we computed a single kernel for all three data partitions, averaged the kernels across different HRR initializations, and then selected submatrices of the average kernel to get $\overline{K}$ and $\overline{K}_*$ accordingly for each cross-validation fold.

Our initial results indicated that our analysis is confounded by game identity: it produces nearly identical results to those of a simple model were the feature vector $\mathbf{x}$ is a 6-dimensional one-hot vector representing the game currently being played (Figure S3). To address this, we regressed out game identity from the BOLD signal and the model:

$$\mathbf{y}' = R\mathbf{y}, \quad \text{(Equation 13)}$$

$$K' = RKR^{\top}, \quad \text{(Equation 14)}$$

where $R = I - \mathbf{X}_g\mathbf{X}_g^{\dagger}$ is the residual forming matrix for the game identity encoding model defined by the design matrix $\mathbf{X}_g$ († denotes the Moore-Penrose pseudoinverse). This is equivalent to using the residuals from a game identity GLM fit to the BOLD signal.

Holographic reduced representations

We embedded EMPA theories in a vector space using holographic reduced representations[22] (HRRs), a kind of vector symbolic architecture[76] that can represent compositional structure in distributed form. HRRs were originally proposed as a model of associative memory and have since been used for modeling structured memories of past events.[77] HRRs use circular convolution (⊛) to associate pairs of items, represented by vectors, and addition to create bags of associations. The resulting vectors can be further associated or grouped together to represent higher-order compositions. Individual items could be extracted from the resulting vector using circular correlation, although we do not take advantage of this in our work.

An EMPA theory consists of three components – objects (SpriteSet), relations (InteractionSet), and goals (TerminationSet) – that we embed separately and then combine into a single vector (Figure S1B).

The SpriteSet is a set of object (sprite) classes, each consisting of a set of properties with given values (e.g., type=Missile, color=blue, speed=slow). The base vectors corresponding to properties (e.g., type) and their values (e.g., Missile) are drawn from isotropic D-dimensional Gaussian distributions $\mathcal{N}(\mathbf{0}, \sigma_h^2 I)$, where $\sigma_h = 1/\sqrt{D}$, bound together using circular convolution, and added to produce the vector for the corresponding sprite class (e.g., c3 = type ⊛ Missile + color ⊛ blue + speed ⊛ slow). The vectors for different sprite classes are scaled to unit length and added together to produce the SpriteSet vector, which is also normalized to unit length.

The InteractionSet is a set of relations (interactions) between pairs of sprite classes, each describing the outcome of an interaction. Each interaction has three key properties: an agent sprite class, a patient sprite class, and an interaction type describing the outcome of the interaction (e.g., killObject). In addition, there may be other optional properties (e.g., scoreIncrement). As with the SpriteSet, the base vectors for properties and their values are drawn from D-dimensional isotropic Gaussian distributions, with the exception of values for agent and patient vectors which are the SpriteSet vectors for the corresponding sprite classes. The property and value vectors are bound together and added to produce the interaction vector (e.g., i3 = patient ⊛ c0 + agent ⊛ c3 + interaction ⊛ killObject). The vectors for different interactions are normalized to unit length and added together to produce the InteractionSet vector, which is also normalized to unit length.

The TerminationSet is a set of exploitative goals (termination conditions) and exploratory goals. Each termination condition has a type (e.g., counter), a sprite class, an outcome (e.g., loss), as well as any additional properties (e.g., count). Exploratory goals have two sprite classes whose interaction is yet unobserved, as well as other optional properties. As with the InteractionSet, the base vectors for properties and their values are drawn from D-dimensional isotropic Gaussian distributions, with the exception of spite classes whose vectors are the corresponding SpriteSet vectors. The property and value vectors are bound together and added to produce the goal vector (e.g., t0 = type ⊛ counter + sprite ⊛ c0 + outcome ⊛ loss). The values for different goals are normalized to unit length and added together to produce the TerminationSet vector, which is also normalized to unit length.

The resulting SpriteSet, InteractionSet, and TerminationSet vectors are finally added to produce the theory vector, which is also normalized to unit length. Following Plate,[22] we chose the dimension D of the vectors as:

$$D = 3.16(k - 0.25)\ln\frac{m}{q^3} \approx 348 \quad \text{(Equation 15)}$$

Where $k = 10$ is the number of stored vectors, $m = 10$ is the vocabulary size, and $q = 0.05$ is the probability of retrieval error.

For an intuitive example of HRRs, see Figure S1C. Notice that while individual HRR features are meaningless by design (Figure S1C, second panel), the similarity between HRR embeddings reflects their semantic similarity (Figure S1C, third panel), which is in turn captured by the GP kernel (Figure S1C, fourth panel) and exploited by GP regression for prediction. For a real example of HRRs for an actual participant and how they evolve over time, see Figure S2.

For the encoding model analysis using simplified object embeddings (Figure S4H), each sprite class had a single property whose value could be either approach, avoid, or neutral, based on whether the avatar ought to approach, avoid, or not be concerned with sprites of that class, respectively. The approach sprites were (for each game): scared (Chase); box1, box2, box3 (Helper); mushroom, key, goal, box (Bait); goal, lemming (Lemmings); deadMolarInf, deadMolarSup, hotdog, burger (Plaque Attack); annoyed (Avoid George); key, goal (Zelda). The avoid sprites were: angry (Chase); hole (Bait); hole (Lemmings); annoyed (Avoid George); monsterQuick, monsterNormal, monsterSlow (Zelda). All other sprites were neutral. The rest of the theory embedding was generated as described above.

Note that we are not making a strong commitment to HRRs as a neural code. Specifically, we are not testing the hypothesis that the brain encodes theories in a form similar to HRRs; rather, we are using HRRs as a cognitively plausible theory embedding to construct the theory similarity kernel K for our encoding model. We leave the question of theory coding as the topic of future work.

Control models

We performed a similar encoding model analysis with 3 control models:

- DDQN, to account for model-free RL representations,
- PCA, to account for low-level visual representations,
- VAE, to account for higher-level state representations.

Deep RL networks (DQNs) have achieved human-level performance on Atari games[2] and have been put forward as an account of human model-free RL in complex domains.[9] Following Tsividis et al.,[16] we used a double DQN (DDQN), which is a version of the original DDQN with improved convergence properties.[20] We ran the sequence of frames (scaled to 64 × 64 × 3), actions, and rewards from each participant through DDQNs pre-trained for the corresponding games, as described above. Following Cross et al.,[9] we performed PCA on the resulting activations for each layer separately (except the output layer, which has only 6 units). This balances the number of features for different layers and facilitates comparisons between DDQNs across games. For each frame, we concatenated the top 100 principal components from all layers into a single 406-dimensional feature vector. The resulting feature vector sequences were fed through the same analysis pipeline as the EMPA theory embeddings (Figure 3A).

Principal component analysis (PCA) has been used to explain brain activity in the visual pathway[24,25] and has been utilized as a control model for human RL in Atari games.[9] We first extracted principal components from 430,000 randomly chosen frames (scaled to 64 × 64 × 3) across all participants. We used the incremental PCA algorithm from the sklearn Python library with a batch size of 10,000. We then projected the frame sequence from each participant's gameplay on to the top 100 principal components and fed the resulting feature vectors through the same analysis pipeline as the EMPA theory embeddings.

Variational autoencoders (VAEs) extract a latent representation of an input space by learning to compress and then reconstruct the input data using a deep neural network.[26–28] VAEs have also been used as a control model for human RL in Atari games.[9] We used an open-source VAE implementation (https://medium.com/dataseries/variational-autoencoder-with-pytorch-2d359cbf027b). The encoder had 3 convolutional layers (conv1: 8 filters with size = 3×3 and stride = 2; conv2: 16 filters with size = 3× 3 and stride = 2; conv3: 32 filters with size = 3×3 and stride = 2), followed by 2 fully connected layers (linear1: 128 units, linear2: 128), followed by the bottleneck layer (latent: 128 units). The decoder had a correspondingly inverted architecture, with 2 fully connected layers followed by 3 convolution transpose layers. The VAE was trained by maximizing the evidence lower bound (ELBO) on the marginal likelihood of the training data. As with PCA, we trained on 430,000 random frames across all participants. We used batch size = 256 and trained for 1,000 epochs using the Adam optimizer with learning rate = 0.001 and weight decay = 10^{-5}. We then ran the frame sequence from each participant's gameplay through the VAE and used the bottleneck activations as the feature vectors which were fed through the same analysis pipeline as the EMPA theory embeddings.

GLM analyses

To look for brain regions sensitive to theory updates, we employed a standard GLM approach using SPM12 (Figure 4A). We created a GLM with impulse regressors at time points when the theory inferred from EMPA changed ($\theta_t \neq \theta_{t-1}$). We also included nuisance regressors for visual and motor confounds, variables relevant for theory updating, as well as motion estimates derived from realignment and run-specific intercepts (Table S3). All regressors were convolved with the canonical hemodynamic response function. As in our previous work,[67] group-level statistical maps were thresholded at $p < 0.001$ and cluster FWE corrected at $\alpha = 0.05$.

As with the encoding model, ROIs were extracted by cross-referencing the peak voxels from the group-level t-map (up to 3 peaks peaks per cluster, minimum 20 voxels apart) with the AAL3 atlas.[29] For our confirmatory analysis, we used all anatomical ROIs with an average beta coefficient for theory updating which was significantly different from zero across participants (Figure S6). We generated PETHs for a given participant and ROI by taking the 20-s (10 TRs) BOLD timecourse following every theory update event, averaged across all voxels in the ROI, and subtracting a baseline BOLD signal averaged over the preceding 4 s (2 TRs) to obtain the change in BOLD signal in response to theory updating. The resulting traces were averaged across theory update events and aggregated across participants to obtain the final PETHs (Figure 4C). The same analysis was performed for the control events. To directly compare the change in BOLD signal in response to different kinds of events (Figure 4D), we averaged the BOLD timecourse within the 20-s window following each event before averaging across event instances and aggregating across participants.

To check if theory updating exhibits a monotonic trend over time, we performed a two-tailed Mann-Kendall test using the theory update histogram averaged across all games (Figure S5A: All Games), downsampled to 1 Hz for ease of computation. We also performed the same test using theory update histograms for individual games (Figure S5A, panels 1-7).

GLM comparison

To identify regions which are sensitive to different update events, we constructed 4 additional GLMs analogous to the theory update GLM: 3 GLMs for single component updates (objects, relations, and goals, respectively) and a single GLM with separate regressors for all three component updates (Figure 5A). Following our previous work,[31,53] we compared GLMs using random effects Bayesian model selection,[32] a standard method for comparing models in fMRI studies.[64] We approximated the log model evidence as LME = -0.5 * BIC, where BIC is the Bayesian information criterion based on the maximum likelihood estimate of the GLM parameters. This quantifies how well the GLM fits the BOLD signal in the ROI for a given participant (penalizing for model complexity), with

lower values indicating a better fit. We report the protected exceedance probability (PXP), which is the posterior probability that a given model is most prevalent in the population (Table 1).

Theory activation timecourse

We generated the overlap between our theory representation and theory updating t-maps by taking the voxels that were significant in both group-level t-maps (Figure 6A). To generate PETHs with predictivity scores (Figure 6B) for a given ROI and participant, we first obtained a predictivity timecourse by computing the Fisher z-transformed Pearson correlation between the predicted and actual pattern of BOLD activity across voxels at each TR. We then proceeded in a similar fashion to the BOLD PETHs described above: the 20-s predictivity timecourses following theory updates were baseline-subtracted (average of preceding 4 s), averaged across theory update events, and aggregated across participants to obtain the PETHs. The same analysis was performed for the control events. As with the BOLD PETHs, to directly compare the change in predictivity score in response to different kinds of events (Figure 6C), we averaged the predictivity timecourse within the 20-s window following each event before averaging across event instances and aggregating across participants. When performing this analysis for separate theory component updates (Figure S8), we used predictivity scores from encoding models fit separately for objects, relations, and goals, respectively.

Effective connectivity

Following our previous work,[53] we investigated the pattern of effective connectivity between brain regions during theory updating using structural equation modeling.[36,37,78] We constructed a beta series GLM with separate impulse regressors for individual theory update events. Since the BOLD signal is highly autocorrelated, which violates the structural equation modeling assumptions, we only included events that are at least 10 s apart, using a rolling window starting from the first theory update event in each run. The resulting beta coefficients are estimates of the instantaneous neural activity at each theory update event. For each ROI, we averaged the estimates across voxels. We searched the space of connectivity patterns using the IMaGES (independent multiple-sample greedy equivalence search) algorithm[36,37] from the TETRAD software package for causal modeling.[38] IMaGES is a version of greedy equivalence search[79] (GES), which starts with an empty causal graph and greedily adds edges that improve the fit to the data according to the BIC. IMaGES extends GES to multiple datasets (e.g., multiple fMRI participants) by averaging the BICs across datasets. To find the effective connectivity pattern 2 s after theory updating, we performed the same analysis except with all theory updates shifted back by 2 s.