A Study on the Content, Format, and Implementation of Neural Representations That Underlie Flexible Human Cognition

Thesis by
Hristos Spiridonos Courellis

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
(Defended June 6, 2024)

Hristos Spiridonos Courellis
ORCID: 0000-0001-5963-679X

ACKNOWLEDGEMENTS

The work included in this thesis would not have been possible without the tireless efforts of many individuals whose names are not Hristos Spiridonos Courellis. There are many individuals who have contributed small pieces to this work in many ways over the years, but the most important are:

Ueli Rutishauser, who was my guide through the space of human cognition.
Ralph Adolphs, who gave a home to a graduate student he had never met.
Markus Meister and Richard Andersen, who pushed me to stand my intellectual ground.

Adam Mamelak, Chrystal Reed, and the rest of the Cedars clinical team, without whom not a single experiment would have been possible.

Gert Cauwenberghs, who gave me the opportunity to learn about computation.
Cory Miller, who gave me the opportunity to learn about the brain.

The members of the Adolphs and Rutishauser Labs, who have supported and challenged me to better myself and my science.

My friends at Caltech and at UCLA, who help banish the occasional monotony of scientific pursuit.

Samuel Nummela, who taught me to listen to neurons, and to doubt and be skeptical as only a primate electrophysiologist could.

My parents, Spiros Courellis and Panayiota Courelli, whose gamble at a better life created the opportunity for me here, and whose tireless efforts and unwillingness to give up on me ensured that I would get here.

And my sisters, Asimina, Vasiliki, and Nemea, without whom I would have slightly more money and slightly less stress, but whose companionship makes life worth living.

ABSTRACT

Humans are the most capable cognitive generalists to walk the earth. They have a remarkable capacity for flexibility reallocating cognitive resources to rapidly acquire and execute an effectively infinite number of tasks. By utilizing the opportunity to record single-neuron activity in the frontal and temporal lobes of awake, behaving neurosurgical patients, we aim to elucidate the principles by which task representations are organized at the neural-circuit level to give rise to flexible cognition and behavior.

Our research program consists of four inter-related projects, each of which seeks to clarify the content, format, and single-neuron implementation of the representations that underlie different aspects of cognition and behavior that are uniquely human. In the first project, we demonstrate that the emergence of disentangled task representations in the hippocampus correlates with the ability of an individual to discover and perform inference on the state of latent context variables in their environment. In the second project, we describe differences in the temporal stability of instructed task representations in the hippocampus and medial frontal cortex, and show that they rely on persistent activity of single-neurons that lasts for 1-2 orders of magnitude longer than is typically studied in working-memory tasks. In the third project, we study the neural mechanisms of task-switching costs, and show that the state of medial frontal cortical context-representing neurons immediately following instructions is predictive of switching cost. In the fourth project, we evaluate the extent to which frontal cortical task representations inherit the compositional structure of natural language, and attempt to predict the neural representation of novel tasks as patients perform zero-shot generalization in a large task space.

Together, these projects constitute a first step in understanding the neural computations that underlie cognitive processing used by humans to solve complex, multi-task environments.

PUBLISHED CONTENT AND CONTRIBUTIONS

Courellis, H. S. et al. (2023). “Abstract representations emerge in human hippocampal neurons during inference behavior.” *bioRxiv* 2023.11.10.566490.

H.S.C. participated in data collection, analysis, and writing of the manuscript.

TABLE OF CONTENTS

Chapter 1

Motivation

Flexible neurons support flexible behavior.

Why did nature bother to wire up so much telencephalon with such an intricate architecture? Why are we, as humans and animals, not simply an amalgamation of delay lines and lookup tables that allow for the appropriate action to be deployed slightly later than the onset of a given external stimulus? After all, we could just feel hypothalamic, act basal ganglionic, and observe superior tectonic could we not? Why go through the painstaking effort of wiring up so much brain to compute in such exotic and multifaceted ways? The trouble here (I suspect) is that the world changes, and rather quickly at that. The parts of our nervous system that are close to the sensory or motor periphery have wonderfully short time constants, implemented through cellular hardware, that allow them to receive the sensory world and control motor effectors with millisecond-precision; a fact that reflects the intrinsic timescales on which our reciprocal control of the world exists at the lowest level. However, as our distance from the periphery increases, measured in number of synaptic jumps, we find in general that the timescale of self-similarity of neural activity increases considerably. If we were to compare the time constant of the autocorrelation of spike trains discharged from a retinal ganglion cell ($\tau_{ms} = 5$)[1] and a neuron deep in the primate frontal cortex ($\tau_{ms} = 350$)[2], we find that their ratio is of the order of 10^2. The former is calibrated to process changes in visual scenes that are known to occur on the timescale of milliseconds. Thus, the existence of the former coupled with the presence of the latter suggests the need for cellular hardware in the brain that responds to changes in the world occurring on timescales of the order of ~1s (or longer, as we will see later on). Incidentally, the fact that many higher cognitive behaviors exhibited by humans (e.g. abstraction, inference, and instructed task switching to name a random few) occur on these timescales, coupled with the fact that lesions to parts of the brain with these long-τ neurons lead to deficits in the aforementioned behaviors[3], tantalizes the prospect of studying such neurons to better understand the neural computations that underlie flexible human cognition.

Apart from the matter of intrinsic timescale, there is the matter of dynamic reconfiguration of activity to accommodate different computations. Neurons that lie close to the sensory or motor periphery appear, by all accounts, to retain a certain amount of stability in their responses to different kinds of external stimuli or the production of different kinds of actions respectively. This is likely because the computations that these neurons perform are intrinsically tied to the physical properties of the end-organs to which they are synapsed. The receptive field and tuning properties of a bipolar cell in the retina are largely determined by the small number of photoreceptors to which it is synapsed, their physical location on the retina, and the opsins in those photoreceptors that determine the wavelength of light to which the photoreceptors are sensitive. The response properties that the bipolar cell inherits from these photoreceptors are thus largely immutable (modulo some non-canonical modulation that we will not consider here) as a function of the demands of the organism. That is to say, this bipolar cell will not suddenly begin responding to light arriving at a different part of the retina, or at a different frequency in the visible spectrum, even if that signal were crucial for

the animal's survival. Correspondingly, an α-motor neuron in the spinal cord that is responsible for contraction of a specific muscle fiber can modulate its gain to a certain degree, but it will not change its tuning to cause the recruitment of different fibers (at least not on the timescale of seconds). Thus, to a rough, first-order approximation, and at the risk of upsetting those who work on peripheral sensory and motor plasticity, I contend that the peripheral nervous system at interface with end-organs acts as a largely static array of sensors that accept high-dimensional signals, and acts as a largely static array of effectors that generate low(er) dimensional output signals. If we take this to be true: that both the input and output of the system are pinned in their computations, then how is it possible that humans can adapt to and overcome such a wide range of challenges so rapidly and systematically? Clearly, somewhere between the input and the output, there must exist some neural substrate that exhibits dynamic change in its computation, in its response to external stimuli and internal states, to support the massive cognitive behavioral repertoire that humans possess. We have now squeezed the object of study between the sensory input and the motor output (the most generous bound possible), but the observation that proximity to the boundary leads to an increase in stability further suggests that the further from the periphery we are, the more flexible and task-dependent neuronal responses might be. After all, without an end-organ to tie one down, one might be free to explore neural state space. The reality of the matter is that being electrochemically coupled to ~10^4 synaptic partners also creates neural response inertia (i.e. reduces dimensionality), but we'll also ignore this for now. Thus, we have some preliminary logic on which to ground the idea that looking deep within the human brain, far from the sensory periphery, can lead us to identify the neurons that are involved in generating the interesting cognitive behaviors we'd like to study.

Let's consider the counterfactual for a moment. If the world were standing still, and the generative factors that give rise to our environment remained immutable, then there would be no need to flexibly adapt one's behavior on any timescale. One could simply learn the optimal policy for behavior once, independently of the training time, and apply that policy comfortably with an infinite time horizon. However, the timescale upon which our environment seems to change in a way that is relevant for the decisions of an animal and its subsequent survival can frequently be measured in units of seconds or minutes; considerably faster than infinity. Furthermore, an accurate characterization of the state of the environment frequently depends on taking a guess, or making an inference, about that which is not directly available in the sensory input. Perhaps the relevant information was encountered at some point in the past, either explicitly or implicitly, and the information yielded by those past observations needs to be persistently represented in some form so that it can be used by the organism to acquire a tasty snack, avoid a predator, appropriately answer an email, or any one of a myriad of tasks that befell organisms in the proximal and distal past.

Adapting rapidly to a changing environment involves two practically interrelated, but technically dissociable behavioral processes that humans exhibit, and whose neural computational underpinnings will be studied herein: switching and learning. The process of switching involves an animal systematically altering behavior according to a different, but already known, set of state-action contingencies. Note: here "state" is a general term that refers to the current state of the environment, including both overt variables sampled through the senses and latent variables whose values are inferred, and the internal state of the animal. Typically, an animal might be prompted to exhibit a switch in behavior as a result of some

external signal or perturbation. In this case, we assume that the animal doing the switching does not necessarily need to re-learn state-action contingencies as it changes its behavior. One can imagine various situations where an animal alternates between two well-learned sets of behaviors, such as a squirrel alternating between foraging for food and scanning its environment for predators, or a graduate student alternating between writing a dissertation and browsing social media. In either case, the same peripheral sensors and effectors are rapidly reconfigured and brought to bear on generally similar environments in radically different ways. Extensive reconfiguration of neural resources within the brain is likely needed to support such rapid and flexible cognitive alternation. Now, let us consider the situation where the animal must also learn new state-action contingencies in its environment.

Animals learn about their environment in many different ways that can generally be binned into one of three different categories: trial-and-error (or experiential) learning, observational learning, and instructed learning. While most animals can be said to exhibit some form of experiential learning, this kind of learning is incredibly slow and inefficient, particularly when the state-action space an animal must explore is high dimensional and the animal does not have strong priors to constrain its search through the space such that it must exhaustively sample an exponentially growing space[4]. Observational learning, though interesting in its own right and potentially involving many computationally complex processes including social inference and the representation of self and other, will not be considered here. The final learning category, instructed learning, is unique to humans as far as we know. As a learning mechanism, it is incredibly efficient since exploration of the state-action space collapses from exponentially costly to solvable in constant time given exact specifications provided in a code, such as natural language, that is mutually comprehensible by the instructor and the student.

Suffice it to say that, when a transmitter of information can appear and specify states to identify and actions to perform with an arbitrary degree of time cost and complexity, particularly when language is involved, the degree of cognitive, and thus behavioral flexibility that must be exhibited is immense. Even the morphologically simple instruction requesting that the receiver retrieves a set of items: "Go get me X_1, X_2, and X_3." supports an upper bound N^3 potential sets of instructions, where N is the number of nouns in the English language. Webster's dictionary reported 470,000 word entries in 1993 and if we estimate that even 1% of these are nouns referencing-valid, retrievable objects (the true fraction is almost assuredly higher), the space of 3-item retrieval instructions contains more than 10^{11} possible requests, many of which could be comprehended and fulfilled by a human immediately. Even statements to the tune of "Go get an MD, a PhD, and neurosurgical board certification." can be comprehended and executed, though with considerable effort and ongoing sleep deprivation on the part of the receiver. Furthermore, the obvious exponential scaling of the space with N, coupled with the clear ability of humans to retrieve increasingly large sets of arbitrary items, suggests that the high degree of behavioral flexibility is supported by internal computational machinery that is more complex than a simple lookup table as proposed earlier, which would exhibit very poor scaling properties if naively implemented.

The point of all these examples is to emphasize that within animals, and within humans in particular, some physical hardware must be present that implements computational processes that allow for sensation and action to be steered in a very deliberate and efficient way to handle the dynamic complexities of the world. We have strong logical

priors to suggest that such processes are implemented in the brain as described in the previous paragraphs. However, we also have strong tangible experimental evidence. Much of the clinical and basic neuroscientific and psychological research that has been performed over the last century or so has, in some way, sought to contribute to the corpus of knowledge that documents and accounts for the neurological underpinnings of complex human cognition. The worlds of neurology and neurosurgery have been particularly productive and contributory in this regard since patients with both focal and generalized lesions, generated either (naturally) pathologically or iatrogenically, will present with very stark cognitive deficits that have been historically described as deficits in attention, complex planning, emotional regulation, episodic memory formation, etc…[5] While these terms are historically loaded and do not necessarily map onto individual groups of neurons or neural circuits, the causal impingement on regions deep in the frontal, temporal, and parietal lobes of the brain leading to profound deficits in cognition that leave much of sensory and motor processing intact provides evidence to suggest that neurons in these regions, though not demonstrably sufficient, are at least necessarily involved in the computations that generate the interesting suite of higher cognitive behaviors in humans that we would like to study.

However, when it comes to understanding the neural computation performed in the human brain in service of these flexible behavior at the level of spiking neural activity, we continue to be profoundly ignorant; a state of affairs that I assuredly will not be able to change by the end of this thesis. In fairness to those who have come before me, and for whom I now work, our field, that of systematic basic science for understanding computing in the human brain at the single-neuron level, has only existed for approximately 20 years. High throughput single-neuron recordings in the human brain were simply not performed until the turn of the millenium[6,7], and to this day, we continue to perform experiments that allow us to record 1-3 neurons at a time intra-operatively from the brain of an awake behaving patient[8]. In fact, I am about to go participate in my last such recording during my PhD as I sit here, writing this text. Though the process of recording neurons with individual tungsten electrodes has served animal neurophysiologists well in the past, such an approach, as it is limited by ethical and practical constraints in the human brain, is likely not the way forward for understanding how networks of neurons in the human brain compute in service of behavior. A detailed understanding of the neural computations that underly much of human cognitive flexibility remains thus elusive. There exist many features of flexible cognition which are uniquely and prominently instantiated in the human. With the advent of widespread single-neuron recordings from within the brains of awake, behaving neurosurgical patients, the opportunity now exists to obviate model systems for the human brain, and to study the neural computations underlying flexible behavior in the most capable cognitive generalist present on earth. In the following sections, I will discuss in greater detail specific aspects of human cognition that will be the subject of study in this thesis, and I will further motivate the experimental and computational tools through which that study has transpired.

Our scientific and clinical advantage

When trying to make sense of a complex, dynamical system with many degrees of freedom, one might wonder what the "best" level of description might be for trying to understand the inner workings of that system. The normative account of levels here might depend on one's

objectives. For example, in the realm of thermodynamics, one might rely on analyses of probability distributions over microstates using tools from statistical mechanics to develop understanding at the fundamental level at which matter is organized. However, if one's objective is to build a steam engine, then designing and building from first principles seems rather abusive. Instead, one might begin by leveraging the ideal gas law or some relaxation thereof, to design the basic specifications according to realistically tolerable temperatures, pressures, and the energy to be extracted. These equations, which provide a meso-scale level of description for a thermodynamic system, were initially described phenomenologically by synthesizing empirical gas laws that relate pressure, temperature, volume, etc.., and though they might not exactly mechanistically specify the exact state of molecules in the studied material at any point in time, the high level abstraction these equations provide is useful for understanding how thermodynamic systems behave at spatial scales that humans directly encounter with their natural senses, and is useful for generating physically realizable solutions to problems humans encounter in the world. It is possible that this correct level of description for useful analysis and control of neural networks is neural state space analysis, where the activity of each individual neuron in a recorded population defines an axis in a state space, the state of the neural population is defined and represented by points in this state space, and information about the external environment (e.g. a task) is read out from this state space using linear decoders, approximating the perspective of a downstream neuron. This general approach has been productive for furthering understanding about computations performed by networks of neurons in the brains of animals in recent history[9].

However, such lines of analysis are only productive if one has the capability to record from a large number of neurons within the brain of an awake, behaving human. Historically, such access to the human brain simply did not exist. The last 20 years have seen the development and widespread deployment of chronically-implantable electrodes within the brains of neurosurgical patients being treated for pharmacologically intractable epilepsy. These Behnke-Fried electrodes are comprised of a clinical stereo-EEG electrode with a hollow core that allows for the passage of high-impedance microelectrodes through the clinical electrode and into the distal brain tissue[6]. Much literature is available on the implantation and recording procedures induced by Behnke-Fried electrodes, and so I will not dwell on the matter much here. In brief, these electrodes allow for the simultaneous recording of: sEEG signal through the low-impedance clinical macroelectrodes located along the electrode shaft, and local field potentials and single-unit spiking activity through the high-impedance microelectrodes (microwires) protruding from the distal end of the electrode shaft. It should be noted here that, for recording units, each Behnke-Fried electrode provides 8 microwires that splay into the brain parenchyma and away from each other, yielding a maximum of 8 channels of 1-D voltage recordings that are processed and spike sorted independently. How, then, is it possible to achieve wide-spread, high-throughput recording of unit activity with so few channels? The answer lies in the clinical demands of the Phase II epilepsy patients treated at Cedars-Sinai Medical Center (CSMC). Through a longstanding collaboration with CSMC, we at Caltech have the unique privilege to work alongside a clinical neurological and neurosurgical team that is incredibly thorough and systematic in their recording from the brains of epilepsy patients. These same patients are kind and generous enough to be willing to volunteer their time and their brains such that we might deepen our understanding of how the human brain computes in service of behavior. Our hope

is that, through this work, we can eventually create enough useful, robust knowledge that allows for the explanation and prediction of neural activity in the brain such that that knowledge might be then leveraged in turn to develop treatments and therapies for individuals who have lost or are in the process of losing the higher cognitive functions we aim to study through our research.

Given the relatively short history of our field, and the near-infinite number of behaviors that humans can generate, we have no shortage of directions in which to steer and find ourselves at the frontier of what is known about how the brain implements and manipulates representations at the level of populations of single neurons to generate flexible cognitive behaviors. Thus, we will proceed by proposing four different, but interrelated projects, each of which explores a different aspect of the human ability to rapidly adapt ones behavior in complex environments with changing task rules.

In the first project, we demonstrate that the emergence of disentangled task representations in the hippocampus correlates with the ability of an individual to discover and perform inference on the state of latent context variables in their environment. In the second project, we describe differences in the temporal stability of instructed task representations in the hippocampus and medial frontal cortex, and show that they rely on persistent activity of single-neurons that lasts for 1-2 orders of magnitude longer than is typically studied in working-memory tasks. In the third project, we study the neural mechanisms of task-switching costs, and show that the state of medial frontal cortical context-representing neurons immediately following instructions is predictive of switching cost. In the fourth project, we evaluate the extent to which frontal cortical task representations inherit the compositional structure of natural language, and attempt to predict the neural representation of novel tasks as patients perform zero-shot generalization in a large task space.

Bibliography:

1.Berry, M. J., Warland, D. K. & Meister, M. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences* **94**, 5411–5416 (1997).

2.Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661–1663 (2014).

3.Fuster, J. M. The Prefrontal Cortex—An Update: Time Is of the Essence. *Neuron* **30**, 319–333 (2001).

4.Bellman, R. & Kalaba, R. A Mathematical Theory of Adaptive Control Processes. *Proceedings of the National Academy of Sciences of the United States of America* **45**, 1288–1290 (1959).

5.Fuster, J. M. The prefrontal cortex in the neurology clinic. *Handb Clin Neurol* **163**, 3–15 (2019).

6.Fried, I. *et al.* Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. *J Neurosurg* **91**, 697–705 (1999).

7.Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* **3**, 946–953 (2000).

8.Mosher, C. P., Mamelak, A. N., Malekmohammadi, M., Pouratian, N. & Rutishauser, U. Distinct roles of dorsal and ventral subthalamic neurons in action selection and cancellation. *Neuron* **109**, 869-881.e6 (2021).

9.Chung, S. & Abbott, L. F. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology* **70**, 137–144 (2021).

Chapter 2

Abstract representations emerge in human hippocampal neurons during inference behavior

Abstract:
Humans have the remarkable cognitive capacity to rapidly adapt to changing environments. Central to this capacity is the ability to form high-level, abstract representations that take advantage of regularities in the world to support generalization[1]. However, little is known about how these representations are encoded in populations of neurons, how they emerge through learning, and how they relate to behavior[2,3]. Here, we characterized the representational geometry of populations of neurons (single-units) recorded in the hippocampus, amygdala, medial frontal cortex, and ventral temporal cortex of neurosurgical patients performing an inferential reasoning task. We find that only the neural representations formed in the hippocampus simultaneously encode multiple task variables in an abstract, or disentangled, format. This representational geometry is uniquely observed after patients learn to perform inference, and consists of disentangled directly observable and discovered latent task variables. Learning to perform inference by trial and error or through verbal instructions led to the formation of hippocampal representations with similar geometric properties. The observed relation between representational format and inference behavior suggests that abstract/disentangled representational geometries are important for complex cognition.

Introduction:

Humans have a remarkable capacity to make inferences about hidden states that describe their environment[3–5] and use this information to adjust their behavior. One core cognitive function that enables us to perform inference is the construction of abstract representations of the environment[5–7]. Abstraction is a process through which relevant shared structure in the environment is compressed and summarized, while superfluous details are discarded or represented so that they do not interfere with the relevant ones[8,9]. This process often leads to the discovery of latent variables that parsimoniously describe the environment. By performing inference on the value of these variables, frequently from partial information, the appropriate actions for a given context can rapidly be deployed[5,10], thereby generalizing from past experience to novel situations. For example, a latent variable specifying being in a left- or right-driving nation can be used by a pedestrian to infer which way to look for oncoming traffic when crossing a road, even in the absence of a sensory cue such as traffic moving in that pedestrian's field of view, and when crossing roads they have never before encountered, for example in the countryside after visiting only cities. Through abstraction, the common, underlying structure of the world is represented in a way that facilitates adaptive behavior.

What would be the signature of an abstract neural representation that enables this kind of adaptive behavior? The simplest form of abstraction is one in which all the irrelevant information is discarded. For example, when the representation of pedestrian crossings in a left-driving nation is a unique pattern of neural activity that is always the same regardless of other sensory details (e.g. whether it is in an urban or rural area). A distinct pattern of activity represents all crossings in a right-driving nation (see Fig. 2.1a). This type of invariant, clustered representation is dissociated from specific instances, matching the way abstraction is defined in everyday language (see e.g. dictionaries like Webster). This type of abstract representation has also been proposed in neuroscience and studied in fMRI experiments by measuring clustering[11]. However, this kind of invariance is rarely observed in the brain, and this definition of abstract representation is too restrictive, failing to capture and explain much of the geometry of neural representations.

For this reason, a more general geometric definition of what an abstract representation is has recently been proposed[12]. For example, the disentangled geometry shown in Fig. 2.1b encodes two variables that characterize each crossing: the first one says whether the crossing is in a left or right-driving nation, while the second one expresses whether it is in a city or countryside. The two variables are represented along orthogonal directions. This non-trivial geometrical arrangement entails the existence of two subspaces in which the representation of each encoded variable is invariant (i.e. it does not depend on the value of the other variable). Indeed, when projecting along the green axis, we recover the geometry of Fig. 2.1a, for which the information about city or countryside is discarded. The advantage of this representation is that in the original space, that information is not lost, and actually, when projecting along the red axis, it is the only information that is represented, making the representation of city/countryside invariant with respect to the nation.

The reason why we care about this type of invariance is that it has important computational properties: it allows a simple linear readout to generalize to novel situations. For example, imagine we train a linear classifier to respond with "look left" in a right-driving nation and "look right" in a left-driving nation. However, we train it only on urban pedestrian crossings. Thanks to the representational geometry, this classifier would also work in rural areas, which are new to the classifier (we assume that the geometry has been learned from other experiences in urban and rural areas). As this out-of-distribution form of generalization is the property of the geometries in Fig. 2.1a and 1b that we consider important, we will use it as the defining characteristic of an abstract representation: a representation of a particular variable is abstract if a linear decoder trained to report

the value of that variable can generalize to novel conditions. The novel conditions are defined by the values of other variables. Representations with these properties have been observed in monkeys[12–14], in rodents[15,16] and in artificial neural networks[11,16,17]. Are these abstract representations also observed in the human brain? How do they form? Do they matter for behavior? At the level of neuronal activity, the answers to such questions have remained elusive. Prior research has implicated the hippocampus in the implementation of abstraction and inference-related computations, both through neuroimaging in humans during tasks that require abstraction and generalization[10,17–19], and through neurophysiology in rodents and non-human primates engaged in tasks with abstract spatial and non-spatial components[12,20–24]. To date, relatively few studies have explored the role that the geometry of task variable representations plays in shaping computation in the human brain at the single neuron level[25,26], and no study has, to our knowledge, reported the emergence and manipulation of this geometry on the short timescales that would be required for rapid learning in humans.

We recorded the activity of populations of neurons in the brains of awake, behaving epilepsy patients to study the emergence of abstract representations. Patients performed a reversal learning task with two latent contexts, each requiring different responses to the same stimuli. We find that as patients learned to perform inference on the latent context, an abstract representation of context emerged in the hippocampus. Importantly, the emergence of the abstract context variable was correlated with an individual's ability to rapidly perform inference on the state of the latent variable context and was absent during error trials. Furthermore, we found that this abstract hippocampal context representation could emerge in two ways: by learning through experience and through verbal instructions informing patients about the latent structure of the task.

Results:

Humans perform inference in a context dependent task.

Patients viewed a sequence of images and indicated for each whether they thought that the associated action was a "left" or "right" button press on a button box (Fig. 2.1c). Subjects discovered from the feedback provided after each response what the correct response is for a given image. There were two possible fixed mappings ("Stimulus-Response-Outcome/SRO maps", see Fig. 2.1e) between each of the four stimuli, the associated correct response (Left/Right button press), and monetary reward given for a correct response (25¢ or 5¢). Which of the two fixed mappings should be used depended on which context a given trial was in (Fig. 2.1d,e, Context 1 or 2). The two contexts alternated every 15-32 trials. Context was a latent variable that had to be inferred by subjects because no information was provided on the screen on which context was presently active or whether it had changed. Critically, the two stimulus-response maps are systematically related: all stimulus-response pairings are inverted between the two contexts (see Fig. 2.1e). With this design, an individual performing inference can detect a change in latent context after receiving feedback that their response was incorrect in a single trial and immediately update their stimulus-response associations for the remaining stimuli even though they have not yet been encountered in the new context. We refer to the trials in which a given stimulus is encountered for the first time following a covert context switch as inference trials and to the remaining trials as non-inference trials.

Patients (17 total, see Table 1) completed 42 sessions (180-320 trials/session, 10-16 blocks/session) of the task, typically in pairs of two back-to-back sessions on the same recording day (mean = 2.4 sessions per day, min = 2, max = 4, see Table 1). Novel stimuli were used in every session, thus requiring patients to re-learn the SRO maps through trial and error at the start of every session. Of the 42 sessions, 6 were excluded from analysis due to at-chance performance in non-

inference trials (Binomial Test, $p > 0.05$). Performance on non-inference trials was well above chance for the remaining 36 sessions (Fig. 2.E1a,b). Each of the 36 included sessions was classified as either a "Inference Present" or "Inference Absent" session depending on whether the patient performed significantly above chance on the first of the three possible inference trials occurring after context switches (Fig. 2.1f, timepoint 2).

Our task is designed such that by performing inference, patients can respond correctly the first time they see an image in a new context following the initial error trial (seen as significantly below chance performance in Fig. 2.1f at timepoint 1). This can be achieved by patients flexibly updating the currently active SRO map immediately after encountering an error, thereby allowing them to perform accurately for the remaining three stimuli that had not yet been seen in the new context. We took accuracy on the first of these three opportunities (the first inference trial) following a context switch as the behavioral signature of successful behavioral inference (timepoint 2 in Fig. 2.1f, Binomial Test, $p < 0.05$). Block-wise estimates of task performance for inference absent (Fig. 2.E1e) and inference present (Fig. 2.E1f) sessions reveal that during inference absent sessions, patients exhibit poor inference performance after every context switch throughout the task, although the performance at the end of every block is high. In contrast, during inference present sessions, inference performance rapidly rose over the first few blocks and remained high throughout the duration of the session (Fig. 2.E1f). Note that within a given session, the two latent contexts had identical stimuli, responses, and outcomes; the only difference was which stimulus was associated with which response and outcome. Correspondingly, subject-level accuracies (Fig. 2.E1c) and reaction times (Fig. 2.E1d) for the two contexts (arbitrarily labeled 1 and 2 across sessions) were not significantly different, indicating that there was no systematic performance bias for one of the two contexts.

When first performing the task, patients were told that they needed to learn arbitrary stimulus-response associations that would change over time, but were not informed about the latent contexts and their related structure. Sessions were recorded in back-to-back pairs (Session One/Two, Fig. 2.E1i inset) with verbal instructions (see Methods) detailing the latent structure of the task provided during the inter-session period (mean length of break = 241 s, range 102-524s), which was considerably shorter than the sessions themselves (mean = 1154 s, range 898-1900s). Importantly, the session following the verbal instructions required re-learning the SRO maps for new stimuli. We considered whether patients could discover the latent task structure before receiving instructions, and if not, whether verbal instructions successfully shaped behavior. Patients were split into three groups: A "post-instruction inference" group, which is composed of patients who did not perform inference during the first session and who did perform inference during the second session (5 patients, 10 sessions, Fig. 2.E1g); An "inference not exhibited" group, which were patients who did not perform inference during both Session One and Two (4 patients, 8 sessions, Fig. 2.E1h); and a "pre-instruction inference" group, which were patients who exhibited inference behavior during both Session One and Two (3 patients, 6 sessions, Fig. 2.E1i). Only patients who performed accurately in non-inference trials in both Session One and Two were included in one of these three groups (Fig. 2.E1g-i, "last"; 5 patients excluded, see Table 1). For each subject, no sessions after the first two were considered for this analysis. Thus, patients exhibited a variety of inference behaviors. Below, we contrast the neural representations between these groups of patients to examine how instructions shape neural representations on short timescales.

Electrophysiology and analysis approach

Neural data recorded over the 36 (of 42) included sessions yielded 2694 (of 3124) well isolated single-units, henceforth neurons, distributed across the hippocampus (HPC, 494 neurons), amygdala (AMY, 889 neurons), pre-supplementary motor area (preSMA, 269 neurons), dorsal anterior cingulate cortex (dACC, 310 neurons), ventromedial prefrontal cortex (vmPFC, 463 neurons), and ventral temporal cortex (VTC, 269 neurons) (Fig. 2.1g,k). Only well isolated neurons as assessed by spike sorting quality metrics were included (see methods). Action potentials discharged by these neurons were counted during two 1s long trial epochs: during the baseline period (base, -1s to 0s prior to stimulus onset), and during the stimulus period (stim, 0.2s to 1.2s after stimulus onset). For the stimulus period, since patients would sometimes respond before 1.2s (reaction time = 1.08 ± 0.04s over sessions), we determined that 75.15% of all spikes occurred before a response was provided across all recorded neurons, indicating that analyses performed with these spike counts predominantly, but not exclusively, reflect pre-decision processing.

Single neuron responses during the two analysis periods were heterogeneous. During the stimulus period, some neurons exhibited selectivity to one or several of the four variables stimulus identity, response, (predicted) outcome, and context (Fig. 2.1h-j and Fig. 2.E1j show example neurons tuned to response and context). Other neurons were modulated by combinations of these variables (Fig. 2.1j, example neuron tuned to conjunction of stimulus and context). Across all brain areas, 54% of units (1447/2694) were tuned to task variables, with 26% of units (706/2694) exhibiting only interaction effects, 17% (449/2694) exhibiting only main effects, and 11% (292/2694) exhibiting both when fitting a 3-Way ANOVA for Response, Context, and Outcome (Fig. 2.1l, RCO column, chance = 135/2694 units, factor significance at $p < 0.05$, Fig. 2.E1t shows each brain area separately). When neurons were separated into those recorded from inference absent and inference present sessions, 5-15% of neurons were significantly tuned for each of the main and interaction effects of the 3-Way ANOVA, with significant reductions in the proportion of neurons tuned to Outcome, Response x Context, and Response x Context x Outcome ($p_O = 0.0007$, $p_{RxC} = 0.0395$, $p_{RxCxO} = 0.0048$, two sample z-test) in inference present sessions compared to inference absent sessions (Fig. 2.E1q). Similar analyses conducted on a separate 2-Way ANOVA for Stimulus Identity and Context (Fig. 2.1l, SC column, Fig. 2.E1r, $p_S = 0.0165$, two sample z-test comparing inference absent with inference present sessions, Fig. 2.E1u shows each brain area separately), and for Stimulus Identity and Response (Fig. 2.1l, SR column, Fig. 2.E1s, $p_S = 0.0287$), revealed a significant decrease in the fraction of neurons tuned to Stimulus Identity in inference present compared to inference absent sessions, again with no significant changes in the proportion of neurons coding for Context, Response, or interactions. These findings indicate diverse tuning to many task variables simultaneously across all brain regions.

Measures of Neural Population Geometry

Given the heterogenous nature of the response pattern at the single neuron level (also see Fig. 2.E1k-p,t-u), we adopted a population-level approach performed on neural pseudopopulations constructed by pooling neurons across patients and sessions (see Methods). This approach allows us to assess which variables are encoded in distributed neural activity patterns, considering also the correlations of the neural responses across multiple conditions. Most importantly, it enables us to examine how these variables are represented and, in particular, to study the geometry of neural representations, which we use to define abstract representations (see the Introduction). In our task, the geometry of a representation is defined by the arrangement of the eight points in the activity space that represent the experimental conditions. Low dimensional disentangled geometries (e.g., when the eight points define a cube) would be abstract because they confer the ability to cross-

generalize to a linear readout. For example, consider a simplified situation with three neurons (the axes) and two stimuli in two contexts (Fig. 2.2a-c). Imagine that the 4 points (2 per context) are arranged on a relatively low dimensional square (the maximal dimensionality for 4 points is 3), with the context encoded along one side and stimulus along one of the two orthogonal sides (Fig. 2.2a). Then, a linear decoder for stimulus (A vs B), trained only on context 1 conditions, can readily generalize to the context 2 (Fig. 2.2b). This ability to generalize is due to the particular arrangement of the points, which make the stimulus coding direction in the two contexts parallel to each other (Fig. 2.2c). Moreover, context and stimulus are represented in orthogonal subspaces, and hence, they are called disentangled variables[27,28]. In the square example, the ability of a linear decoder to generalize across conditions (cross-condition generalization or CCGP) also applies to the variable context (i.e. a context decoder trained on stimulus A conditions will generalize to stimulus B conditions). As discussed in the introduction, we use CCGP as the defining characteristic of an abstract representation of a variable.

Notice that if the 4 points of the example are at random locations in the activity space defining a tetrahedron, the representation is "unstructured" and does not have any of the generalization properties described. On the other hand, these high dimensional representations allow a linear readout to separate (or shatter) the points in any arbitrary way, and hence confer to the readout the flexibility to implement any possible task. We refer to the number of ways the points can be separated into two groups by a linear decoder (dichotomies) as shattering dimensionality[12,29]. Recorded neural representations can have both the generalization properties of the abstract representations and the flexibility of the high dimensional representations[12].

We compared the representational geometry between inference absent and inference present sessions for neural pseudopopulations of all recorded neurons in each brain area. Analyses were performed on all variables defined by "balanced dichotomies", which are constructed by splitting the 8 task conditions into two groups of 4 conditions (Fig. 2.1b, E2, and 2d). To perform the analysis in an unbiased manner, we did not consider only the variables defining the task but all the variables that correspond to the 35 possible dichotomies of the 8 conditions (see Fig. 2.2d and E2 for an illustration of the dichotomies that correspond to specific task variables, Table 2; we refer to these as named dichotomies). The dichotomies corresponding to latent context, behaviorally relevant stimulus grouping (stim pair), and parity, which measures the degree of non-linear interactions of variables in the neural population. These variables are the most important for interpreting subsequent results, and are shown in Fig. 2.2d.

For each of the dichotomies, we computed the decoding accuracy, which tells us whether the corresponding variable is encoded, and the CCGP, which indicates whether the representation of that variable is disentangled from other variables. Both decoding accuracy and CCGP are reported in a cross-validated manner by training and testing decoders on single trials. We complemented this single-trial analysis with a third metric called the Parallelism Score. The Parallelism Score measures the cosine similarity of the coding directions of a specific variable. The coding directions are estimated using the average activity for each condition. A high Parallelism Score indicates that the variable is represented in an abstract format. The Parallelism Score is a direct geometrical measure that focuses on the structure of the representation (the CCGP also depends on the noise and its shape).

Hippocampal neural geometry correlates with inference behavior.

We first examined the decodability of each balanced dichotomy in different brain areas for sessions where inference was present and sessions where inference was absent. Following stimulus onset, the hippocampal neural population exhibited a significant increase in average decodability across all

balanced dichotomies in inference present sessions relative to inference absent sessions, defined as an increase in shattering dimensionality (Fig. 2e, inference absent vs present, 0.57 vs 0.62, $p_{RS} = 2.7x10^{-3}$, Ranksum over dichotomies). Latent context (Fig. 2.2e, red, inference absent vs. present $p_{RS} = 2.9x10^{-27}, p_{Absent} = 0.12$, $p_{Present} = 5.1x10^{-5}$; p_{Absent} and $p_{Present}$ are significance tests vs. chance and p_{RS} is a pairwise comparison between inference absent and inference present sessions) and stim pair (Fig. 2.2e, purple, inference absent vs. present, $p_{RS} = 5.0x10^{-27}$, $p_{Absent} = 0.015, p_{Present} = 7.9x10^{-7}$) emerged as the most strongly decodable named dichotomies in the inference present sessions. The stim pair dichotomy corresponds to the grouping of stimulus identities that elicit the same response (D&B vs. A&C), a relationship that remains the same across both contexts (Fig. 2.2d). This difference in representation between inference absent and inference present sessions is unique to HPC, as no other recorded region exhibited a significant change in decodability of the variable context (Fig. 2.2g, red). Rather, in vmPFC, stim pair (Fig. 2.2g, purple, inference absent vs. present, $p_{RS} = 6.9x10^{-20}$, $p_{Absent} = 0.21, p_{Present} = 0.0097$) and in preSMA, response (Fig. 2.2g, green, inference absent vs. present, $p_{RS} = 5.2x10^{-13}, p_{Absent} = 0.091, p_{Present} = 0.0057$) increased significantly in decodability in inference present compared to inference absent sessions, as expected (see discussion). The stim pair variable was also significantly decodable in AMY both during inference absent and present sessions, with no significant difference between the two (Fig. 2.2g, purple, inference absent vs. present, $p_{RS} = 0.88, p_{Absent} = 0.0016, p_{Present} = 0.0018$).

The expressiveness of a neural representation can be quantified by the decodability of dichotomies that probe for non-linear interactions of variables in the population. The parity dichotomy (Fig. 2.2d, orange) is only decodable if variables are encoded with a high degree of non-linear interactions in a neural population (see methods). We observed that in the hippocampus but not in other brain areas, parity decodability increased significantly in inference present relative to the inference absent sessions (Fig. 2.2e, orange, inference absent vs. present, $p_{RS} = 1.5x10^{-21}, p_{Absent} = 0.27, p_{Present} = 0.0055$) . Generalizing this finding, dividing different dichotomies into increasing levels of "difficulty", with more difficult dichotomies requiring stronger non-linear interactions of task variables, reveals that average decoding accuracy is highest for the most difficult dichotomies in the hippocampus (Fig. 2.E5). Together, these findings suggest that non-linearities in the hippocampal population response in the inference present relative to the inference absent sessions led to an increase in the number of dichotomies that could be decoded by a linear decoder. Notably, only the hippocampus exhibited significant parity decodability (Fig. 2.2g, orange, all $p > 0.05$), a significant increase in shattering dimensionality (Fig. 2.2g, all $p_{RS} > 0.05$), and the emergence of multiple, simultaneously decodable dichotomies between inference absent and inference present sessions.

We next examined the format of the decodable named dichotomies (context, stim pair, parity). During the stim period, CCGP (Fig. 2.2f, E3d) was significantly elevated for both the context (Fig. 2.2f, red, inference absent vs. present, $p_{RS} = 2.0x10^{-28}$, $p_{Absent} = 0.51, p_{Present} = 0.02$) and stim pair (Fig. 2.2f, purple, inference absent vs. present, $p_{RS} = 2.0x10^{-2}$, $p_{Absent} = 0.17, p_{Present} = 0.0011$) variables in inference present but not in inference absent sessions. In addition, the Parallelism Score was significantly larger than expected by chance for the variables context and stim pair in the inference present but not inference absent sessions (Fig. 2.E3g, red, $p_{Absent} = 0.55, p_{Present} = 1.4x10^{-15}$ *and* Fig. 2.E3g, purple, $p_{Absent} = 0.17, p_{Present} = 1.7x10^{-8}$). Elevated CCGP was also observed for the stim pair variable in vmPFC during inference present and not inference absent sessions (Fig. 2.E3a, purple, $p_{Absent} = 0.45, p_{Present} = 0.014$), the response variable in preSMA (Fig. 2.E3a, green, $p_{Absent} = 0.050, p_{Present} = 0.0010$), and the

stim pair variable in AMY during both inference absent and inference present sessions (Fig. 2.E3a, purple, $p_{Absent} = 0.050, p_{Present} = 0.039$). Taken together, the increased CCGP and Parallelism Score values in inference present relative to inference absent sessions indicate that the context and stim pair variables are both simultaneously represented in an abstract format in the hippocampus in sessions where inference behavior was observed. These two variables were not represented in an abstract format when inference is absent. While other task variables were also represented in an abstract format in other brain regions, only the hippocampus simultaneously represented these two variables in an abstract format (Fig. 2.E3a,b). These two disentangled variables are thus represented in approximately orthogonal subspaces.

We also conducted a parallel analysis during the pre-stimulus baseline (Fig. 2.2, inset), analyzing the geometry of persistent representations of the previous trial. We found that context alone was encoded in an abstract format in the hippocampus only in sessions in which subjects could perform inference (Fig. 2.2h,i; Supplement S.1). This finding indicates that the hippocampal context representation was persistently maintained in an abstract format across trial epochs.

Lastly, we examined whether the geometry of the context representation was preserved such that context decoding could generalize across different inference present sessions. To do so, we aligned the geometries in the two sessions to each other in neural state space using a subset of task conditions and then examined whether decoding context generalized from one session to the other on held-out conditions (see methods). This analysis revealed high context parallelism between random subsets of different inference sessions during both the baseline and stimulus periods (Fig. 2.E3z, aa). Such cross-session context parallelism was not found when performing the same analysis for the inference absent sessions (Fig. 2.E3ab, ac). This indicates that the geometry of the hippocampal context representation generalizes across inference sessions.

The changes in hippocampal neural geometry are summarized in Fig. 2.2j, which shows a 3D MDS plot of the hippocampal neural data in inference absent (left) and present (right) sessions, with hypothetical linear boundaries (black lines) showing the separating hyperplanes for context and stim-pair, the two disentangled variables.

Hippocampal representation of context is absent in error trials

We next asked whether the presence of context as an abstract variable was associated with trial-level performance. To examine this question, we compared the decodability and geometry of all dichotomies between correct and incorrect (error) trials in sessions where patients exhibited inference. The first trial of every block was excluded from this analysis due to being necessarily incorrect by design (see Fig. 2.1d, trial 1). Contrasting correct with error trials, we found that shattering dimensionality was significantly higher in correct trials in the stimulus period (Fig. 2.E3e, inference present vs. inference present (error), 0.59 vs 0.54, $p_{RS} = 0.0048$), as was decodability of the parity dichotomy (Fig. 2.E3e, inference present vs. inference present (error), orange, $p_{RS} = 0.029$). Furthermore, the dichotomies context and stimulus pairing were significantly more decodable in correct compared to incorrect trials (Fig. 2.E3e, inference present vs. inference present (error), red, $p_{RS} = 1.2x10^{-20}$, purple, $p_{RS} = 1.0x10^{-9}$). Furthermore, Parallelism Score for context but not other variables was significantly elevated in correct trials but not in incorrect trials (Fig. 2.E3f, red, inference present vs. inference present (error), $p_{Present} = 1.1x10^{-16}, p_{Present\ (error)} = 0.083$) in inference present sessions. Similarly, the baseline representation of context also showed this effect, being decodable during correct trials but not during incorrect trials (Fig. 2.E3i, red, inference present vs. inference present (error), $p_{RS} = 3.5x10^{-1}$, $p_{Present} = 0.012, p_{Present\ (error)} = 0.47$). The baseline shattering dimensionality did

not significantly differ between correct and error trials (0.52 vs 0.51, $p_{RS} = 0.31$). Context was present in an abstract format only in correct trials based on the Parallelism Score for context being significantly larger than chance during correct but not incorrect trials (Fig. 2.E3j, red, inference present correct vs. error $p_{Present} = 0, p_{Present\ (error)} = 0.94$). The lack of decodability and parallelism of context in the baseline immediately prior to an incorrect trial indicates that the geometry of the representation at baseline is correlated with correct behavior in the upcoming trial. Together, these findings demonstrate that both the content and format of the hippocampal neural representation are correlated with behavior on a trial-by-trial basis. This effect is present during the stim period in incorrect trials, where shattering dimensionality, context decodability, and context Parallelism Score significantly decreased. This effect is also present during the baseline period, where a reduction in the decodability and parallelism of the context variable is associated with an error in an upcoming trial.

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

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

Second, to assess whether our results were influenced by pathology, we repeated our analysis after excluding hippocampal neurons that were located within clinically confirmed seizure onset zones (Supplement 2.S.3). We found no quantitative changes in our results, suggesting that hippocampal pathology did not influence our results.

Lastly, we examined whether our results were sensitive to behavioral accuracy in non-inference trials (Fig. 2.1d, trial indicated as 'last'). We repeated our analysis in a subset of inference absent and inference present sessions that were chosen such that non-inference trial performance was matched (Supplement 2.S.4). This control analysis revealed no qualitative changes in our results, suggesting that differences in non-inference trial performance cannot explain our results.

Abstraction of stimulus coding across contexts uniquely increases in the hippocampus

Individual hippocampal neurons in humans prominently encode the identity of visual stimuli[30]. Visually tuned neurons, whose firing rate is strongly modulated by the identity of presented images, are an example of such encoding[31]. We therefore next asked how the variable context, which we show above is encoded in the hippocampus, interacts with stimulus identity. As the four visual stimuli do not share any apparent structure, we do not expect to observe any interesting structured geometry when all the stimuli are studied together. For this reason, we studied the geometry of pairs of stimuli (e.g. stimulus A vs B) in the two contexts. We focused on HPC and the ventral temporal cortex (VTC). VTC neurons were strongly modulated by stimulus identity (see below)[32,33], but context was not decodable at the level of the balanced dichotomy analysis (baseline period: Fig. 2.E3u, red; compare with Fig. 2.2h; stimulus period: Fig. 2.E3v, red, compare with Fig. 2.2e).

First, we verified that neurons in both areas encoded the identity of the four stimuli presented. This was the case in both hippocampus and VTC: 109/494 (22%) of neurons in hippocampus (Fig.

2.3a, 2.E6g,h show examples) and 195/269 (73%) of neurons in VTC (Fig. 2.3d, 2.E6i,j show examples) were significantly modulated by stimulus identity following stimulus onset (1x4 ANOVA, $p < 0.05$). Similarly to hippocampus, at the population level, VTC neurons encoded stimulus identity-related balanced dichotomies in an abstract format (Fig. 2.E3u-y, purple, brown, pink, $p_{Absent/Present} < 10^{-10}$). Furthermore, error trial analysis revealed that stimulus-related dichotomies were also decodable during errors in VTC (Fig. 2.E3y, purple, brown, pink, $p_{Present\,(error)} < 10^{-10}$). This finding contrasts with the hippocampus (compare to Fig. 2.E3e, stim pair AC vs. BD dichotomy) and is consistent with the idea that neurons in VTC veridically represented the stimulus viewed on the screen by the patient during both correct and error trials.

We next conducted a geometric stimulus-pair analysis to study the interaction of stimulus identity and context coding in the same neural population. The stimulus-pair analysis was designed to detect the presence of simultaneous abstract coding of stimulus identity across contexts and abstract coding of context across stimuli (see Fig. 2.E6a-f for illustration).

The average stimulus decoding accuracy across all stimulus pairs in the hippocampus did not differ significantly between inference absent and inference present sessions (0.73 vs. 0.76; Fig. 2.E6m, $p_{RS} = 0.13$, RankSum over stimulus pairs), indicating that the decodability of stimulus information was not different when patients could perform inference vs. when they could not. In contrast, the geometry of the stimulus representation became disentangled from context: both the stimulus CCGP (Fig. 2.3b, $p_{RS} = 0.041$) and stimulus Parallelism Score (Fig. 2.3c, $p_{RS} = 0.040$) were significantly increased in inference present compared to inference absent sessions. This means that a decoder trained to differentiate between stimulus A and B in one context generalized better to the other context in inference present compared to inference absent sessions (and vice-versa). This finding suggests that the stimulus responses reorganized with respect to the emerging context variable. Note that context was not decodable in inference absent sessions as a balanced dichotomy (Fig. 2.2e, red). Nevertheless, stimulus decoders did not generalize well across the two contexts in inference absent sessions. This result indicates that context did modulate stimulus representations in the hippocampus, but in a way that was entangled with stimulus identity in inference absent sessions (see below). This effect was specific to the hippocampus: in VTC, the neural population geometry was unchanged, as indicated by no significant differences in stimulus decodability (Fig. 2.E6n, $p_{RS} = 0.15$), stimulus CCGP (Fig. 2.3e, $p_{RS} = 0.15$) and stimulus Parallelism Score (Fig. 2.3f, $p_{RS} = 0.39$) between inference absent and inference present sessions. In VTC, CCGP was high even in the inference absent session, indicating that shifts in context did not modulate stimulus identity representations like in the hippocampus.These analyses demonstrate that the representational geometry for stimulus identity in hippocampus becomes significantly more structured across contexts in inference present sessions compared to inference absent sessions in a manner that reflects an abstract format.

We next turned our attention to the generalization of the context code across stimuli. The presence of abstract coding for one variable (stimulus identity) does not necessarily imply that the other variable is also present in an abstract format, though we do have evidence that this is the case in hippocampus from the CCGP and Parallelism Score analysis over balanced dichotomies (Fig. 2.2f, E4g). Context decoding analysis conducted over stimuli (e.g. considering only trials with stimuli A&B shown, decode context) in hippocampus revealed that context was decodable for many stimuli both during inference absent and inference present sessions, without a significant difference between the two (0.63 vs. 0.67; Fig. 2.E7a, $p_{RS} = 0.065$). However, despite being decodable, context encoding during inference absent sessions was not in an abstract format for stimulus pairs as indicated by low context CCGP (Fig. 2.3g, $p_{Absent} > 0.10$ for all stim pairs) and low context

Parallelism Score (Fig. 2.3h, $p_{Absent} > 0.17$ for all stim pairs except for AB, where $p_{Absent} = 0.033$) values that were not significantly greater than chance. In contrast, in inference present sessions, both context CCGP (Fig. 2.3g, $p_{RS} = 0.012$) and context Parallelism Score (Fig. 2.3h, $p_{RS} = 0.015$) increased significantly relative to the inference absent group. This finding indicates that context emerged as an abstract variable at the level of individual stimulus pairs in the hippocampus.

We next contrast these findings with VTC. While context was decodable from some stimulus pairs during inference absent and inference present sessions (Fig. 2.E7b, $p_{Absent} \in (0.013, 0.074)$, $p_{Present} \in (0.020, 0.081)$ for all stim pairs), there was no significant change in context decodability between inference absent to inference present sessions (Fig. 2.E7b, $p_{RS} = 0.18$). Rather, there was a significant decrease in context CCGP (Fig. 2.E7c, $p_{RS} = 0.026$) and no significant difference in context Parallelism Score (Fig. 2.E7d, $p_{RS} = 0.39$) from inference absent to inference present. Together, these findings indicate that in the hippocampus, the context variable in the inference present sessions is in an abstract format because context coding directions become aligned (i.e. parallel) across stimuli. For VTC, on the other hand, the lack of an abstract context representation across stimulus identities in both inference absent and inference present sessions suggests that context, though decodable for individual stimulus pairs, is not organized in an abstract format and does not meaningfully correlate with inference behavior.

In summary, these findings indicate that the emergence of context as an abstract variable in the hippocampus when patients can perform inference is coupled with the reorganization of stimulus representations so they are also more disentangled, thereby forming a jointly abstracted code for stimuli and context. This transformation of the representation is visible directly in the data when projecting the neural representations in 3 dimensions using Multidimensional Scaling (Fig. 2.3i, 2.E8, Supplementary Video 1). This reorganization occurs without the encoding of additional stimulus information since individual stimuli are equally decodable in the presence or absence of inference behavior. This change in geometry relies on the encoding of context in an abstract format and is unique to hippocampus. In contrast, we found no systematic reorganization of stimulus representations in VTC.

How neural population geometry changes are implemented in hippocampal neural activity

We next examined what aspects of neuronal activity changed in the hippocampus to give rise to the abstract neural representations that we observed (i.e. the representations with elevated CCGP). We considered the following non-mutually exclusive possibilities (Fig. 2.4a-d). (i) The distances between conditions in state space could increase (Fig. 2.4a vs. Fig. 2.4b), either as a result of increased firing rate of variable-coding neurons, an increase in the fraction of tuned neurons, or an increase in the depth of tuning of these neurons. (ii) The variance of the population response projected along the coding direction could decrease (Fig. 2.4c). (iii) Parallelism could increase due to increases in the consistency of firing rate modulation in response to one variable over values of another (Fig. 2.4d).

We first examined whether mean firing rates across all recorded neurons differed between inference absent and inference present sessions in the hippocampus. The firing rate across conditions decreased from 3.37±0.13 to 1.36±0.03 Hz, a 60% reduction on average during the stimulus period (Fig. 2.4e, $p_{RS} = 8.3x10^{-5}$). Firing rates were also reduced during the baseline period (3.29±0.09 to 1.38±0.02 Hz, 58% reduction, Fig. 2.E9q). This firing rate reduction was unique to the hippocampus , with every other recorded region exhibiting no significant differences or increases in firing rate between inference absent and inference present during the stimulus (Fig. 2.E9c) and

baseline periods (Fig. 2.E9r). Further analysis revealed a small number of inference absent sessions with high firing rate that biased the mean inference absent firing rate. Repeated geometric analysis after removing these sessions (Fig. 2.E9x) revealed a more modest 32% firing rate difference (1.67±0.05 to 1.13±0.03 Hz, Fig. 2.E9y). Excluding these sessions did not alter the geometric results (Fig. 2.E9z-ab).

The firing rate reduction led to a decrease in the average distance between class centroids (separation) across all dichotomies in inference present sessions (5.77 ± 0.22 to 4.17 ± 0.07 Hz, $p_{RS} = 2.9x10^{-8}$, Fig. 2.4g). However, the centroid distance for a single dichotomy, the context dichotomy, increased from inference absent to inference present sessions (4.3 vs 5.0 Hz, $p_{Absent} = 0.87, p_{Present} = 0.076$, $p_{\Delta Dist} = 0.040$, Fig. 2.4g, h, E9h). In fact, context was the dichotomy with the largest change in distance in firing rate space when comparing the inference present and inference absent conditions (Fig. 2.4h). This isolated significant rise in context separability was not seen in any of the other recorded areas during the stimulus period (Fig. 2.E9a,b). Similarly, during the baseline period, the distance between context centroids decreased the least in the hippocampus (5.6 vs 5.0 Hz, $p_{Absent} = 0.68$, $p_{Present} = 0.0007$, $p_{\Delta Dist} = 0.027$, Fig. 2.4j,k) despite the significant decrease in distance over all dichotomies that was also observed here due to the firing rate reduction (5.85±0.08 to 4.25±0.04 Hz, $p_{RS} = 6.5x10^{-13}$, Fig. 2.4j).

Next, we assessed changes in the variability of the population response along the coding direction of each dichotomy. The variance along the coding direction of neuronal responses in the hippocampus decreased for all dichotomies in inference present when compared to inference absent sessions during both the stimulus period (2.51 ± 0.16 vs. 1.53 ± 0.06, $p_{RS} = 6.5x10^{-13}$, Fig. 2.4i) and the baseline period (2.49 ± 0.09 vs. 1.58 ± 0.02, $p_{RS} = 6.5x10^{-13}$, Fig. 2.E9k,l). However, this decrease could be a simple consequence of the reduction in firing rates under the assumption of Poisson statistics. We conducted a condition-wise Fano-factor analysis to assess whether the variance reduction was beyond that expected for the reduction in firing rates. This analysis revealed no significant differences in Fano factors between inference absent and inference present sessions during the stimulus period (1.39±0.22 vs 1.36±0.14 , $p_{RS} = 0.99$, Fig. 2.4f) and the baseline period (1.61±0.26 vs 1.45±0.11, $p_{RS} = 0.19$). Together, these two findings suggest that the decrease in variance along dichotomy coding directions is explained by the decreases in firing rate.

Though the increase in distances between dichotomy centroids for context appears to be a distributed, population-level phenomenon (see Fig. 2.E4a-e), we sought to determine if a signature of this increase could be detected in the tuning of individual neurons (Supplement 2.S.5). We found that the proportion of neurons exhibiting univariate context tuning increased from inference absent to inference present sessions, thus partially explaining the increased representational distance. However, the hippocampal population geometry did not exclusively rely on these neurons because after excluding all neurons with significant univariate coding, we found no qualitative change in the population geometry (Fig. 2.E4). Furthermore, the reduction in hippocampal firing rate from inference absent to present sessions did not bias any geometric measure determined through a firing rate distribution-matched control analysis (Fig. 2.E9s-w).

Finally, we examined the tuning of individual neurons to investigate what gave rise to the increases in parallelism for context across stimuli that we observed (see Fig. 2.E7e-h for examples). A stimulus-tuned neuron also modulated by context could do so consistently across all stimuli (e.g. firing rate increased for all stimuli), or inconsistently (e.g. firing rate increased for some stimuli and decreased for others). Neurons consistently modulated by context would increase context parallelism as their responses would be compatible with those of linear mixed selectivity neurons. Thus, we quantified the consistency in the direction with which stimulus representations were modulated

(firing rate increased or decreased) across contexts for stimulus-identity tuned neurons. Context modulation consistency is computed for each neuron, and can take on values between 0 and 4, with 0 indicating no consistency in modulation and 4 indicating all stimuli exhibit the same firing rate modulation direction between contexts (see Methods for details). We find a significant increase in the consistency of context modulation in the hippocampus from inference absent to inference present sessions (Fig. 2.4l, E6i, 1.8±0.2 vs 2.9±0.3, $p_{RS} = 0.0049$). This effect was specific to hippocampus : in VTC, this metric decreased significantly (Fig. 2.E6i, 2.6±0.3 vs 1.6±0.2, $p_{RS} = 0.0039$). These findings indicate that, for the hippocampus, context parallelism arises in part due to an increase in the consistency with which the firing rates of stimulus-tuned neurons are modulated by context.

The changes in neural state space responses for hippocampus are summarized in Fig. 2.4m, and feature aspects of our previous hypotheses: (i) condition averages for context increase in separation despite relaxing towards the origin (decrease in firing rate), (ii) are accompanied by decreases in variance along the coding direction, and (iii) neurons become increasingly consistent (parallel) in their modulation across stimulus and context dimensions. Together, these changes explain the implementation of the context coding dimension in the hippocampal representation and how it can emerge as a simultaneous, linearly encoded variable alongside stimulus identity.

Context representations outside of the Hippocampus.

The only other area of the brain that we examined in which we found a representation of latent context that correlated with inference behavior was in the dACC, but only during the baseline and not the stimulus period (Supplement 2.S.6). Interestingly, context emerged in the task representation through a different implementation strategy, namely an increase in firing rates rather than a decrease as in the hippocampus.

Verbal instruction induces the representation of context in an abstract format.

In all analyses discussed thus far, we compared sessions in which patients performed inference (inference present) with those in which patients did not (inference absent), without regard to how patients transitioned from inference absent to inference present. We provided verbal instructions detailing the latent task structure after Session One (Fig. 2.5, inset) to all patients, allowing us to examine whether instructions lead to changes in neural representations and behavior in an immediately following session (Session Two). As shown above, patients were divided into three types based on behavior: those who exhibited inference behavior in the very first session (pre-instruction inference), those who did so after being given verbal instructions (post-instruction inference), and those who did not perform inference even after being provided with verbal instructions (inference not-exhibited). We next compared the neural representation of context between these three groups of patients.

The instructions provided to these three groups were identical, and all patients acknowledged receipt of the instructions. All included patients responded with high accuracy on non-inference trials before and after being given instructions, indicating that they understood the task and learned the SRO maps. The principal difference between the post-instruction (Fig. 2.5a, 2.E10a,b) and inference not-exhibited (Fig. 2.5a, 2.E10h,i) groups is their ability to perform inference following the verbal instructions, with both groups performing the task accurately otherwise. The pre-instruction inference group, on the other hand, exhibited above-chance inference performance during both Session One and Two (Fig. 2.5a, 2.E10o,p).

In the post-instruction inference group, context was decodable in the HPC during the stimulus period on correct trials in the session following the verbal instructions (Fig. 2.5b, $p_{One} =$

$0.17, p_{Two} = 0.016, p_{RS} = 3.1x10^{-1}$). This representation of context was in an abstract format, as indicated by significant increases in both CCGP (Fig. 2.E10c; $p_{One} = 0.28, p_{Two} = 0.047, p_{RS} = 8.4x10^{-16}$) and Parallelism Score (Fig. 2.E10d; $p_{One} = 0.023, p_{Two} = 1.2x10^{-6}$). Successful performance in the task was associated with context being represented abstractly in HPC, as both the decodability (Fig. 2.5b, $p_{Two\ (error)} = 0.99$, Session Two correct vs error, $p_{RS} = 4.3x10^{-20}$) and Parallelism Score (Fig. 2.E10d, $p_{Two\ (error)} = 1.1x10^{-4}$) of context decreased significantly on error trials in Session Two. Context was also encoded in an abstract format during the baseline period in the same performance dependent manner as context in the stimulus period (Fig. 2.E10e-g). At the single neuron level, this effect can be appreciated by an increase in the proportion of neurons that are significantly linearly tuned to context ($p < 0.05$, one-way ANOVA for context) during both the stimulus (8% (6/75 neurons) vs. 18% (17/93 neurons), $p = 0.027$) and baseline (7% (5/75 neurons) vs. 16% (15/93 neurons), $p = 0.029$) periods in Session Two compared to Session One (Fig. 2.5f shows an example). Thus, the ability of post-instruction group patients to perform inference following instructions was associated with the rapid emergence of an abstract context variable in their hippocampus.

In contrast to the post-instruction inference group, in patients in the inference not-exhibited group, context was not encoded by HPC neurons during the stimulus (Fig. 2.5c, E10j,k, all $p_{One/Two} > 0.05$) or the baseline (Fig. 2.E10l-n all $p_{One/Two} > 0.05$) periods in Session Two. Furthermore, there was no significant change in tuning to context at the single-neuron level in the hippocampus of patients in this group both during the stimulus period (6% Session One vs 6% Session Two, $p = 0.41$) and the baseline period (8% Session One vs 5% Session Two, $p = 0.27$). These data indicate that receiving verbal instructions describing the latent context alone is insufficient to generate an abstract context representation in the hippocampus. Instead, context was only represented abstractly in the subset of subjects that productively applied the instructions to change their inference behavior.

For the pre-instruction inference patient group, context was already decodable during Session One (Fig. 2.5d, $p_{One} = 0.014$), and dropped slightly below significance ($p_{Two} = 0.17$) during Session Two, likely due to the small number of patients and neurons present in this analysis. The CCGP was not significant (Fig. 2.E10q, $p_{One} = 0.21, p_{Two} = 0.31$), but the Parallelism Score in both Sessions One and Two was significant and near the top of the dichotomy rank order in both cases (Fig. 2.E10r, $p_{One} = 1.5x10^{-9}, p_{Two} = 1.7x10^{-6}$). This finding suggests that the context variable these patients learned experientially during Session One (before the instructions) was in an abstract format as assessed by Parallelism Score, and that receiving the instructions did not significantly alter the behavior or the neural geometry of the context representation. A similar trend was observed with the baseline context representation for these patients (Fig. 2.E10s-u). Note that such discrepancies between CCGP and Parallelism Score are not unexpected since, when less data is available (fewer neurons, decreased firing rates), single trial measures (Decoding, CCGP) become less sensitive to representational structure than measures that operate on condition averages (Parallelism Score).

We also examined firing rate changes of hippocampal neurons separately for the post-instruction inference, pre-instruction inference, and inference not-exhibited groups (Fig. 2.E10v). This analysis revealed significant reductions in firing rate across all conditions from Session One to Session Two for the post-instruction inference patients alone (-0.39 ± 0.15 Hz, $p_{Post-Instructi} = 1.4x10^{-4}$), confirming that reductions in hippocampal firing rate in the same neurons recorded across adjacent sessions were associated with increases in inference performance. With a cell-by-cell comparison, 22/46 neurons (48%) show a significant decrease in firing rate with an average

reduction of 1.3 Hz. Hippocampal neurons from the inference not-exhibited group showed an increase in firing rate (0.10 ± 0.04 Hz, $p_{Not-Exhibite} = 1.2x10^{-4}$) and pre-instruction inference group firing rates did not significantly change (0.06 ± 0.08 Hz, $p_{Pre-Instructio} = 0.08$).

Lastly, we compared the geometry of the context representations formed by each of these patient groups using the Parallelism Score (balancing number of neurons, see methods). Parallelism Score for context increased significantly in the post-instruction inference group, from levels not different from chance during Session One ($p_{One,Post-inst} = 0.20$, Fig. 2.5e) to a level comparable to the pre-instruction inference group during Session Two ($p_{Two,Post-inst} = 0.0028, p_{Two,Pre-inst} = 0.0035$, Fig. 2.5e). The Parallelism Score in the pre-instruction inference group, on the other hand, did not change significantly and was already above chance in Session One. This finding suggests that hippocampal neurons in the pre-instruction inference group carried an abstract representation of context before receiving high-level instructions, and retained that geometry after receiving instructions. On the other hand, hippocampal neurons in the post-instruction inference group did not encode an abstract representation of context before receiving instructions. During Session Two, subjects in the post-instruction inference group could perform inference, and neurons in their hippocampus started to encode a task representation whose geometry resembled that of the pre-instruction group. This result indicates that a similar representational geometry can be constructed through either experience or instruction. Lastly, subjects in the inference not-exhibited group could not leverage the information provided in the instructions to perform inference, and accordingly, their hippocampi never encoded an abstract representation of context.

Discussion

The ability to perform inference in our task was associated with the hippocampus forming an abstract representation of the environment. This representation encoded stimulus identity and latent context in approximately orthogonal subspaces, was behaviorally relevant on the level of individual trials, and emerged with learning (Fig. 2.2, 2.3). To implement this representation, the context coding directions for different visual stimuli became more parallel, the distance between contexts in neural state space increased, and the overall variance in firing was reduced due to a reduction in mean firing rates (Fig. 2.4). This representation could emerge quickly, with some patients spontaneously learning the latent task structure during their first session and others exhibiting abstract representations within minutes of receiving verbal instructions explaining the task structure despite having no previous experience before the data shown here was recorded (Fig. 2.5). Abstract representations of context and stimulus identity following stimulus onset were only present in the hippocampus and not in the other brain areas we examined. Together, this data reveals that hippocampal population codes can be restructured by learning and verbal instructions within minutes to support inference in a new task.

How can a neural or biological network efficiently encode multiple variables simultaneously[12,34]? One solution is to encode variables in an abstract format so they can be re-used in novel situations to facilitate generalization and compositionality[27,35–39]. Here, we show that in the human brain, such a disentangled representation emerged as a function of learning to perform inference in our task. The format by which latent context and stimulus identity were represented was predictive of the ability to perform behavioral generalization that relies on contextual inference. Crucially, patients performed well on non-inference trials in all sessions included in the analysis, indicating that they understood the task and successfully learned the stimulus-response associations in both contexts. Therefore, the difference between the inference present and absent sessions was only in whether they performed inference following the covert context switch (Fig. 2.1f). For those

sessions where patients did not perform inference, there was no systematic relationship between context coding vectors across stimuli. For sessions where patients performed inference, there was alignment of the context coding direction across stimuli (making them parallel), indicating that the context variable had been disentangled from the stimulus identity variable in the hippocampi of these patients (Fig. 2.2j, 2.3i). As a result, the two variables became disentangled, thereby allowing for generalization. This representation was implemented by the hippocampus using a broadly distributed code as evidenced by the high context parallelism score (Fig. 2.E3f,g,j,n), and the lack of reliance on univariately tuned context neurons to generate the abstract context representation (Fig. 2.E4a-j, Supplement 2.S.2). Thus, the geometry we study here did not trivially arise from classically tuned neurons.

Inferential reasoning is thought to rely on cognitive maps, which have been observed in the hippocampus and other parts of the brain[20,40–44] . Cognitive maps are thought to underlie inferential reasoning in various complex cognitive and spatial domains[3,10,40,41,45,46]. However, little is known about how maps for cognitive spaces emerge at the cellular level in the human brain as a function of learning. Here, we show that a cognitive map that organizes stimulus identity and latent context in an ordered manner emerges in the hippocampus. The cognitive map emerges because task states in one context, indexed by stimulus identity, become systematically related to the corresponding task states in the other context through a dedicated context coding direction that is disentangled from stimulus identity (Fig. 2.3b,c,g-i). Furthermore, the relational codes between task states (stimuli) in each context are preserved across contexts.

Hippocampal cognitive maps observed in other studies are often different from those that we observed. Indeed, the encoded variables are observed to non-linearly interact, which is a signature of high dimensional representations. These representations are believed to be the result of a decorrelation of the neural representations (recoding) that is aimed at maximizing memory capacity[47–49]. This form of pre-processing leads to widely observed response properties, like those of place cells[50]. However, there is some evidence of hippocampal neurons that encode one task variable independently of others[16,22,51–56]. In these studies, no correspondence was shown between different representational geometries in the hippocampus and differences in behavior. Here, the task representations generated when patients cannot perform inference (but can still perform the task) are systematically different from the abstract hippocampal representations of context and stimulus identity that correlate with inference behavior[12]. Finally, it is important to stress that we also observed an increase in the shattering dimensionality, which has been in shown in other studies to be compatible with the low dimensionality of disentangled representations[12,16].

We found stimulus identity codes in brain regions other than the hippocampus, but these mostly lacked reorganization as a function of learning to perform inference. This code stability is particularly salient in the ventral temporal cortex, a region analogous to macaque IT cortex, in which neurons construct a high-level representation of visual stimuli[57–59]. Some studies conducting unit recordings in this general region in humans show that neurons exhibit strong tuning to stimulus identity[60]. We similarly find that VTC neurons encode visual stimulus identity (Fig. 2.3d-f, 2.E6n). However, these responses were not modulated by latent context in a systematic manner. As a result, despite being decodable for some individual stimulus pairs, context was not represented in an abstract format. Rather, in VTC, context was only weakly decodable for a subset of the stimuli, context decodability did not change between inference absent and inference present sessions (Fig. 2.E7b,c), and stimulus identity geometry was not reorganized relative to context in inference present sessions (Fig. 2.3e,f). Our study therefore shows that disentangled context-stimulus representations emerged in the hippocampus, but not in the upstream visually responsive region VTC.

Apart from the hippocampus, abstract representations also emerged in two other brain areas we studied: stim-pair and response representations emerging in the vmPFC and preSMA, respectively (Fig. 2.2g, 2.E3a). While interesting in their own right, these variables were the only encoded variables in each respective region, thus preventing us from studying the geometry of multiple simultaneously abstract variables in these two areas. The hippocampus was also not unique in its representation of context. A weaker representation of context was also found in the dACC, but only during the baseline period. This finding aligns with work implicating the dACC in the representation of task rules and task sets [61–66]. Following stimulus onset, however, dACC did not contain a representation of latent context (Fig. 2.2g). In contrast, previous studies in tasks with explicitly cued context switches[25,26,67,68] find that neurons in the medial frontal cortex (dACC and preSMA) are tuned to task context following stimulus onset. We hypothesize that this might be due to differences in task demands: context switches were uncued in our task and had to be inferred from outcomes. It remains an open question to examine whether cued vs. inferred context switches engage different mechanisms of switching between contexts and/or different context encoding schemes in the hippocampus.

The focus of our study was to examine how representations of context, stimulus identity, response, and predicted outcome change as a function of learning. In a prior study in macaques[12], the representation of the same variables in two very well trained animals was examined in HPC, dlPFC, and ACC in a similar task after the completion of training. Several notable differences exist between the two studies. First, context was encoded in an abstract format at baseline and was decodable after stimulus onset in all three brain areas examined in the macaques. In contrast, in humans, context is only strongly decodable in the HPC. We hypothesize that the wide-spread encoding of context in the macaque study was due to the extensive training the animals received before recordings commenced. In contrast, our patients had no prior task experience. It is possible that early on during learning, latent context representations are present only in the hippocampus and are propagated to the cortex (dACC) with extensive task experience. This hypothesis is supported by prominent direct and indirect projections from the hippocampus to dACC in primates[69–71], and flexible, context-dependent interactions between medial frontal cortical neurons and hippocampal outputs[25,72]. Second, the human hippocampus exhibited abstract stimulus representations, unlike the abstract response or "choice" representation in the macaques (in the interval during the presentation of the stimulus). Notably, the abstract stimulus pair and response dichotomies are constructed such that high CCGP for one will necessarily lead to below-chance CCGP for the other, which was indeed the case for both our study (high stim pair, low response) and the primate study (low stim pair, high response). One potential reason for these differences is a species difference: human HPC neurons are strongly modulated by the identity and semantic category of presented images[25,30,73–75], making it natural to organize representations of context relative to this existing representation. Similarly, representations of choices are not prominent in the human HPC[25]. Another potential reason is a difference in task construction: our task employed semantically identifiable images, whereas the prior experiment with macaques used fractals. Third, unlike macaque HPC, human HPC did not encode predicted outcome. We note that in our task, outcome prediction was not necessary to perform the task because context switches were signaled by the accuracy of the response (correct or incorrect), which was independent of predicted outcome received for a correct response. Furthermore, all possible task-states were uniquely indexable using stimulus identity and context, rendering outcome prediction representation unnecessary for unambiguously defining the current task state. Finally, another possibility is that the reward for the macaques (juice volume) was more motivationally salient than the small monetary reward (25¢ or 5¢) patients received. We hypothesize

that these reasons obviated the need for a predictive representation of outcome to complete the task in our patients. It remains an important question whether representations similar to those seen in macaques emerge in the other brain areas we examined following extensive training. Our data indicates that on short experiential timescales, the human hippocampus generates a representation that encodes the minimum set of variables required to solve the task.

In our study, verbal instructions resulted in changes in hippocampal task representations that correlated with behavioral changes. The emergence of this representation in the session immediately following the instructions in the post-instruction inference group is correlated with their newfound ability to perform inference and suggests that hippocampal representations can be modified on the timescale of minutes through verbal instructions (Fig. 2.5). This change in representation is qualitatively different from the standard neurophysiological approach of studying the emergence of a "learning set", wherein a low-dimensional representation of abstract task structure emerges slowly over days through trial-and-error learning [52,76,77]. Our finding of similar representational structure in the hippocampus in subjects who learned spontaneously and those who only learned after receiving verbal instructions suggests that both ways of learning can potentially lead to the same solution in terms of neural representations. In complex, high-dimensional environments, learning abstract representations through trial and error becomes exponentially costly (the curse of dimensionality), and instructions can be used to steer attention towards previously undiscovered latent structure that can be explicitly represented and utilized for behavior. The process of instruction-dependent restructuring of hippocampal representations is likely cortical-dependent, given the role of the cortex in language comprehension, but the exact mechanism by which this process occurs remains to be explored[36,78]. Our findings suggest that when high-level instructions successfully alter behavior, underlying neural representations can be rapidly modified to resemble one learned through experience. However, in our experiment, pre and post-instruction inference groups were mutually exclusive and we did not assign subjects to either group by design. Further experiments are needed to directly test how, if any, differences exist between experientially learned and instructed task representations.

Acknowledgements: We thank R. Adolphs for advice and support throughout all stages of the project and the members of the labs of R. Adolphs, U. Rutishauser, and M. Meister for discussion. We thank all subjects and their families for their participation and the staff and physisicans of the Cedars-Sinai and Toronto Western Epilepsy Monitoring Units for their support.
Funding: This work was supported by the BRAIN Initiative through the NIH Office of the Director (U01NS117839 to U.R.), NIMH (R01MH110831 to U.R.), the Caltech NIMH Conte Center (P50MH094258 to R.A. and U.R.), the Simons Foundation Collaboration on the Global Brain (to S.F., and U.R.), the Moonshot R&D JPMJMS2294 (to Kenji Matsumoto), and by a merit scholarship from the Josephine De Karman Fellowship Trust (to H.S.C.).
Author contributions: Conceptualization: J.M., U.R., C.D.S., S.F., Task design: J.M. and D.L.K., Data collection: J.M., H.S.C, and A.R.C., Data analysis: H.S.C. and J.M., Writing: H.S.C., U.R., and S.F., Clinical care and experiment facilitation: C.M.R. Surgeries: A.N.M. and T.A.V., Supervision: U.R. and S.F.
Competing interests: None.
Data availability statement: Data will be made available on public repositories such as OSF or DANDI upon acceptance, as we commonly do for our publications.

Code availability statement: Example code to reproduce the results will be made available on public repositories such as Github upon acceptance.

Methods:

Participants: The study participants were 17 adult patients who were implanted with depth electrodes for seizure monitoring as part of an evaluation for treatment for drug -resistant epilepsy (see Table 1). 14 were monitored at Cedars-Sinai Medical Center (CSMC) and the other 3 were monitored at Toronto Western Hospital (TWH). All patients provided informed consent and volunteered to participate in this study. All research protocols were approved by the institutional review boards of CSMC, TWH, and the California Institute of Technology.

Psychophysical Task and Behavior: Participants performed a serial reversal learning task. There were two possible static stimulus-response-outcome (SRO) maps, each of which was active in one of the two possible contexts. Context was latent and switches between context were uncued. Each recording session consisted of 280-320 trials grouped into 10-16 blocks of variable size (15-32 trials/block) with block transitions corresponding to a change in the latent context. Each trial consisted of a blank baseline screen, stimulus presentation, speeded response from the participant, followed by feedback after a brief delay (Fig. 2.1a). Responses were either "left" or "right" in every trial. In each session, stimuli were four unique images, each chosen from a different semantic category (human, macaque, fruit, car). If a patient performed multiple sessions, new images not seen before by the patient were chosen for each session. The task was implemented in MATLAB (The Mathworks, Inc., Natick, MA) using PsychToolbox-3[79]. Images were presented on a laptop positioned in front of the patient and subtended approximately 10 degrees of visual arc (300 px^2, 1024x768 screen resolution, 15.6 inch (40 cm) monitor, 50 cm viewing distance). Patients provided responses using a binary response box (RB-844, Cedrus Inc.).

Receipt of reward in a given trial was contingent on the accuracy of the response provided. In each trial, either a high or low reward (25¢ or 5¢) was given if the response was correct, and no reward (0¢) if incorrect. Whether a given trial resulted in high or low reward if the response was correct was determined by the fixed SRO map (see Fig. 2.1c). Stimulus-response associations were constructed such that two out of four images (randomly selected) were assigned one response and the other two images were assigned the other (e.g. human and fruit = left, macaque and car = right). Thus, in each context, each stimulus was uniquely specified by a combination of its correct response (left/right) and reward value (high/low). Crucially, the SRO maps of the two possible contexts were constructed so that they were the opposite of each other from the point of view of the associated response (Fig. 2.1c). To fully orthogonalize also associated reward, half of the reward values stayed the same and the others switched. This structured relationship of stimuli across contexts led to the full orthogonalization of the response, context, and reward variables (Fig. 2.1b-c). Crucially, the stimulus-response map inversion across contexts provided the opportunity for patients to perform inferential reasoning about the current state of the SRO map, and therefore the latent context.

Since rewards were provided deterministically, participants could switch context upon receiving a single error. Therefore, if patients performed inference, they should be able to respond correctly after receiving a single error. The behavioral signature of inferential reasoning was thus the accuracy in the trials that occurred immediately after the first error trial. Specifically, we took a participant's performance on the first instance of each of the three remaining stimuli in the new context is to measure a participants inference capabilities.

Patients completed multiple sessions of the task, in each of which new stimuli were chosen. After completion of the first session, the experimenter provided a standardized description of the latent contexts and SRO reversal to the patient (see below). These instructions were given regardless

of how well the patient performed in the immediately preceding session. After this brief interlude, the participants completed the task again with a novel set of four stimuli.

Instructions given to patients:
---------- Instruction set 1 (before first session) ----------
In this task, we will show you a series of images, 4 of them in total. Your objective is to learn the correct response for each image (either left or right). In the beginning, you will not know what the correct answer is, so take a guess. The correct answer for an image may occasionally change, so pay close attention. For every correct answer you will receive a reward of either 25 or 5 cents. For an incorrect answer you will receive 0 cents. This is real money that you will receive before you leave the hospital in the form of a gift card to your favorite place (ex. Starbucks). You will have the opportunity to take a break halfway through.
---------- Instruction set 2 (before second session) ----------
You may have noticed that some images have the same correct response and some images have the same reward. Even when the correct response changes, they usually change together. In this experiment, we are going to try a different strategy. Pay attention to which images go together (i.e. have the same correct response and similar reward). This should make it a lot easier to perform the task. To make the task a little more difficult, now the correct response for each image will change a little more frequently.

Behavioral Control: We administered a control version of the task identical to the 'first session' described above to n=49 participants recruited on Amazon Mechanical Turk (MTurk). We then used this data to calibrate the difficulty of the task. A majority (~75%) of the control subjects demonstrated proper inference performance, and the remaining 25% demonstrating slow updating of SROs after a context switch, consistent with a behavioral strategy where each stimulus is updated independently (see Fig. 2.E1a).

Electrophysiology: Electrode Placement and Recording: Extracellular electrophysiological recordings were conducted using microwires embedded within hybrid depth-electrodes (AdTech Medical Inc.). The patients we recruited for this study had electrodes implanted in at least the hippocampus, as well as in addition subsets of amygdala, dACC, pre-SMA, vmPFC, and VTC as determined by clinical needs (see Table 1). Implant locations were often bilateral but some patients only had unilateral implants as indicated by clinical needs. Broadband potentials (0.1Hz – 9kHz) were recorded continuously from every microwire at a sampling rate of 32kHz (ATLAS system, Neuralynx Inc.). All patients included in the study had well isolated single neuron(s) in at least one of the brain areas of interest.

Electrode Localization: Electrode localization was conducted using a combination of pre-operative MRI and post-operative CT using standard alignment procedures as previously described[25,67]. Electrode locations were co-registered to the to the MNI152-aligned CIT168 probabilistic atlas[80] for standardized location reporting and visualization. Placement of electrodes in gray matter was confirmed through visual inspection of subject-specific CT/MRI alignment, and not through visualization on the atlas.

Spike Detection and Sorting: Raw electric potentials were filtered with a zero-phase lag filter with a 300Hz-3kHz passband. Spikes were detected and sorted using the OSort software package[81]. All

spike sorting outcomes were manually inspected and putative single-units were isolated and used in all subsequent analyses. We evaluated the quality of isolated neurons quantitatively using our standard set of metrics[73,82,83] including proportion of inter-spike interval violations < 3ms, signal-to-noise ratio of the waveform, projection distance between pairs of isolated clusters, and isolation distance of each cluster relative to all other detected spikes.

Selection of Neurons, Trials, and Analysis Periods: Activity of neurons was considered during two epochs throughout each trial: the baseline period (base), defined as -1s to 0s preceding stimulus onset on each trial, and the stimulus period (stim), defined as 0.2s to 1.2s following stimulus onset on each trial. Spikes were counted for every neuron on every trial during each of these two analysis periods. The resulting firing rate vectors were used for all encoding and decoding analyses. Tests of single-neuron selectivity were conducted using N-way ANOVAs with significance at $P < 0.05$, where N was either 2 for models of stim id (A, B, C, D) and context (1, 2), or 3 for models including outcome (High, Low), response (Left, Right), and context (1, 2). All variables were categorical, and all models were fit with all available interaction terms included.

Population analysis – decoding: Single-trial population decoding analysis was performed on pseudo-populations of neurons assembled across all neurons recorded across all patients. We pooled across sessions within each anatomically specified recording area as described previously[25,26]. We aggregated neurons across subjects into a pseudo-population that consists of all neurons recorded in a given brain area, which allows us to examine populations of several hundred neurons in humans despite inability to record this many neurons simultaneously. This analysis approach is possible because all subjects performed exactly the same task, so that conditions could be matched across all relevant variables for a given trial in the pseudo-population (For example, trial 1 might be context 1, correct response, stimulus A, response right, outcome high). The justification for using this approach is three-fold. First, independent population codes, in which the information that each neuron provides can be characterized by its own tuning curve, can be understood by recording one neuron at a time and aggregating them for analysis[84]. This is the type of code we are examining. Second, we seek to establish the content and structure of information that is reliably present in a given brain area across subjects. This can only be achieved by recording in many subjects. Third, in most instances, decoding from pseudo-populations yields the same results than from simultaneously recorded neurons[85,86]. Results between the two approaches can differ when noise correlations are considered, which can have complex effects on the geometry of the underlying representation[84]. Here, noise correlations are not the topic of interest. Noise correlations are present for the subgroups of neurons in the pseudo-population that were recorded simultaneously. To avoid potential effects of these remaining noise correlations, we removed them by randomly scrambling the order of trials for every neuron included in the pseudo-population (as we have described before[25,26]).“
Decoding was conducted using support vector machines (SVM) with a linear kernel and L2 regularization as implement in matlab’s fitcsvm function. No hyperparameter optimization was performed. All decoding accuracies are reported for decoding accuracy for individual trials. Decoding accuracy is estimated out-of-sample using 5-fold cross-validation unless otherwise specified (e.g. cross-condition generalization). Many of the decoding analyses in this work consist of grouping sets of distinct task conditions into classes, then training an SVM to discriminate between those two groups of conditions. Neurons included in the analysis were required to have at least K correct trials of every unique condition in order to be included in the analysis (K = 15 trials unless otherwise stated). To construct the pseudopopulation, we then randomly sampled K trials

from every unique condition and divided those trials into the groups required for the current decoding analysis for every neuron independently. Randomly sampling correct trials in this way allowed us to destroy noise-correlations that might create locally correlated sub-spaces from neurons recorded in the same area and session[25].

To account for the variance in decoding performance that arose from this random sub-sampling procedure, all reported decoding accuracies are the average resulting from 1000 iterations of sub-sampling and decoder evaluation. A similar trial balancing and sub-sampling procedure was conducted for all analyses that report decoding accuracy on incorrect trials, but with K = 1 trial/condition required as incorrect for the neuron to be included in analysis. Various other analyses conducted throughout this work, including representation geometry measures, centroid distances, and coding direction variances, all rely on this procedure of balanced correct and incorrect trial sub-sampling, and averaging over 1000 iterations of the computed metric to study the relationships between task conditions in an unbiased manner. All reported values have been computed with this approach unless otherwise stated.

Construction of Balanced Dichotomies: Our task has 8 possible states (Fig. 2.1b). We characterized how neurons represented this task space by assessing how a decoder could differentate between all possible "balanced dichotomies" of these 8 task conditions (Fig. 2.1b). The set of all possible balanced dichotomies is defined by all possible ways by which the 8 unique conditions can be split into two groups containing 4 of the conditions each (e.g. 4 points in context 1 vs 4 points in context 2 is the context dichotomy). There are 35 possible balanced dichotomies (nchoosek(8,4)/2). Some of the possible balanced dichotomies are easily interpretable because they correspond to variables that were manipulated in the task. We refer to these balanced dichotomies as the "named dichotomies", which are: context, response, outcome, stimulus pair (stim pair), and parity. These dichotomies are shown individually in Fig. 2.E2. The stim pair dichotomy corresponds to the grouping of stimuli for which the response is the same in either context (A&C vs. D&B; see Fig. 2.E2). The parity dichotomy is the balanced dichotomy with the maximal non-linear interaction between the task variables (Fig. 2.E2).

Defining decoding difficulty of dichotomies: We quantify the relative degree of non-linear variable interactions needed by a neural population to classify a given dichotomy using a difficulty metric that rates dichotomies that require proximal task conditions to be placed on opposite sides of the decision boundary as more difficult. Note that proximity of task conditions in task space here is defined with respect to the variables that were manipulated to construct the task space. The conditions corresponding to (Response L, Outcome Low, Context 1) and (Response L, Outcome Low, Context 2) are proximal since their task specifications differ by a single variable (hamming distance 1) whereas (Response L, Outcome Low, Context 1) and (Response R, Outcome High, Context 2) are distal since their task specifications differ by all three variables (hamming distance 3). With this perspective, we can systematically grade the degree of non-linearity required to decode a given dichotomy with high accuracy as a function of the number of adjacent task conditions that are on opposite sides of the classification boundary for that dichotomy. For a set of 8 conditions specified by 3 binary variables, this corresponds to the number of adjacent vertices on the cube defined by the variables that are in opposing classes (See Fig. 2.E5a). We define this number as the "difficulty" for a given dichotomy, and can compute it directly for every one of the 35 balanced dichotomies. The smallest realizable dichotomy difficulty is 4, and corresponds only to named dichotomies that align with the axis of one of the three binary variables used to specify the task space.

The largest realizable dichotomy is 12, and this corresponds to the parity dichotomy since the dichotomy difficulty (number of adjacent conditions with opposing class membership) is maximized in this dichotomy by definition. All remaining dichotomies lie between these two extremes in difficulty, and computing average decoding accuracy over dichotomies of increasing difficulty gives a sensitive readout of the degree of non-linear task variable interaction present in a neural population.

Geometric Analysis of Balanced Dichotomies: We used three measures to quantify the geometric structure of the neural representation[12]: shattering dimensionality, cross-condition generalization performance (CCGP), and parallelism score.

Shattering Dimensionality is defined as the average decoding accuracy across all balanced dichotomies. It is an index of the expressiveness of a representation, as representations with higher Shattering Dimensionality allow more dichotomies to be decoded. The content of a representation is assessed by considering which balanced dichotomies are individually decodable better than expected by chance.

CCGP assesses the extent to which training a decoder on one set of conditions generalized to decoding a separate set of conditions. Note that to compute CCGP, all trials from a set of conditions are held out from the training data, which is different from the "leave-one-out" type decoding used to estimate Shattering Dimensionality. The remaining held-in conditions are used to train the decoder, and performance is then evaluated on the held-out conditions (trial-by-trial performance). The CCGP for a given balanced dichotomy is the average over all possible 16 combinations of held-out conditions on either side of the dichotomy boundary. One of the 4 conditions on each side of the dichotomy are used for testing, whereas the remaining three on each side of the dichotomy are used for training. For each of the 16 possible train/test splits, the decoder is trained on all correct trials from the remaining six conditions, and performance is evaluated on the two held-out conditions.

Parallelism Score assesses how coding directions for one variable are related to each other across values of other variables in a decoder agnostic manner. The Parallelism Score is defined for every balanced dichotomy as the cosine of the angle between two coding vectors pointing from conditions in one class to conditions in the other for a given dichotomy. These vectors are computed by selecting four conditions (two on either side of the dichotomy), computing the normalized vector difference between the mean population response for each of the two pairs, then computing the cosine between said coding vectors.This procedure is repeated for all possible pairs of coding vectors, and the average over all cosines is reported. Since the correct way of "pairing" conditions on either side of the dichotomy is not known a-priori, we compute the cosine average for all possible configurations of pairing conditions on either side of the dichotomy, then report the Parallelism Score as the maximum average cosine value over configurations.

Null distribution for geometric measures: We used two approaches to construct null distributions for significance testing of the geometric measures Shattering Dimensionality, CCGP, and Parallelism Score.

For the Shattering Dimensionality and decoding accuracy of individual dichotomies, the null distribution was constructed by shuffling trial labels between the two classes on either side of each dichotomy prior to training and testing the decoder. After shuffling the order of the trial labels, the identical procedures for training and testing were employed. This way of constructing the null distribution destroys the information content of the neural population while preserving single-neuron properties such as mean firing rate and variance.

For the CCGP and Parallelism Score, we employed a geometric null distribution[12]. Prior to training, we randomly swapped the responses of pairs of neurons within a given condition. For example, for one task condition, all of neuron 1's responses are assigned to neuron 2 and all of neuron 2's responses are assigned to neuron 1, for another task condition, all of neuron 1's responses are assigned to neuron 3, etc…). This way of randomly shuffling entire condition responses leads to the situation where neural population response statistics by-condition are held constant, but the systematic cross-condition relationships that exist for a given neuron are destroyed. This way of shuffling creates a maximally high dimensional representation, thereby establishing a conservative null distribution for the geometric measures CCGP and Parallelism Score.

Neural Geometry Alignment Analysis: To answer the question of whether the geometry of a variable was common across different groups of sessions, we aligned representations between two neural state spaces. Each state space is formed by non-overlapping sets of neurons, and the two spaces are aligned using subsets of task conditions. A cross-session-group parallelism score was then computed by applying the same alignment to a pair of held-out conditions, one on either side of the current dichotomy boundary. Alignment and cross-group comparisons were performed in a space derived using dimensionality reduction (6 dimensions). For a given dichotomy, two groups of sessions with N and M neurons were aligned by applying SVD to the firing-rate normalized condition averages of all but two of the eight task conditions, one on either side of the dichotomy boundary. The top six singular vectors corresponding to the non-zero singular values from each session group were then used as projection matrices to embed the condition averages from each session group in a 6-dimensional space. Alignment between the two groups of sessions, in the 6-dimensional space, was then performed by computing the average coding vector crossing the dichotomy boundary for each session group, with the vector difference between these two coding vectors defining the "transformation" between the two embedding spaces. To compare whether coding directions generalize between the two groups of sessions, we then used the data from the two remaining held out conditions (in both session groups). We first projected these data points into the same 6-dimensional embedding spaces and computed the coding vectors between the two in each embedding space. We then applied the transformation vector to the coding vector in the first embedding space, thereby transforming it into the coordinate system of the second session groups. Within the second session group embedding space, we then computed the cosine similarity between the transformed coding vector from the first session group and the coding vector from the second session group to examine whether the two were parallel (if so, the coding vectors generalize). We repeated this procedure for each of the other three pairs of conditions being the held-out pair, thereby estimating the vector transformation of each pair of conditions independently. The average cosine similarity was then computed over the held-out pairs. All possible configurations of conditions aligned on either side of the dichotomy boundary are considered (24 in this case), and the maximum cosine similarity over configurations is returned as the parallelism score for that dichotomy (plotted as 'cross-half' in Fig. 2.E3z). As a control, we also computed the parallelism score for held-out conditions within the same embedding space without performing cross-session alignment (plotted as 'half-split' in Fig. 2.E3z). Note that the differences in both the average parallelism score and the null distribution when comparing within-session and across-session parallelism are expected behavior and arise from the increased expressive power of the cross-sesion approach due to fitting transformation vectors in a relatively low-dimensional (6D) space. This step is not performed for the within-session control since there is no need to align neural activity to its own embedding space.

Multi-Dimensional Scaling: Low-dimensional visualization of neural state spaces was achieved using multi-dimensional scaling (MDS) performed on matrices of condition-averaged neural responses. Pair-wise 33idscale33 distances between condition averages were initially computed in N-dimensional neural state space, where N is the number of neurons used to construct the space. Pairwise distances were then used to compute either a 2-dimensional or 3-dimensional representation of the condition averages using the "33idscale" method in Matlab. In figures where two different MDS plots are shown side-by-side, canonical correlation analysis was used to align the axes of the two dimensionally reduced neural state spaces. This approach was necessary since, in general, neural state spaces constructed with different sets of neurons were being compared. We note that we use MDS only to summarize and visualizing high-dimensional neural representations. All conclusions drawn are based on geometric measures computed in the original full neural state space.

Analysis of Incorrect Trials: For determining decoding accuracy for trials in which subjects provided an incorrect response ("error trials"), decoders were trained and evaluated out of sample on all correct trials in inference absent and inference present sessions (denoted as "inference absent" and "inference present" trials respectively). The accuracy of the decoder was then evaluated on the left out error trials in the inference present sessions (denoted as "inference present (error)" trials) that were balanced by task condition. Neurons from sessions without at least one incorrect trial for each of the 8 conditions were excluded. We did not estimate CCGP separately for correct and incorrect trials. The Parallelism Score was estimated using only correct trials for inference present and inference absent. For inference present (error), parallelism was computed using one coding vector (difference between two conditions) from correct trials and one coding vector from incorrect trials. All other aspects of the Parallelism Score calculation remained as described earlier. The very first trial after a context switch was excluded from analysis (it was incorrect but by design, as the subject cannot know when a context switch occurred).

Stimulus Identity Geometry Analysis (Fig. 2.3): We repeated the geometric analysis described above for subsets of trials to examine specifically how the two variables context and stimulus interact with each other. To do so, we considered each possible pair of stimuli (AB, AC, AD, BC, BD, CD) separately. For each stimulus pair, we then examine the ability to decode and the structure of the underlying representation for two variables: stimulus identity (see Table 3) and context (see Table 4).

For stimulus identity, what is decoded is whether the stimulus identity is the first or second possible identity in each pair (i.e. "A vs. B" for the AB pair). Stimulus CCGP (Fig. 2.3b,e) is calculated by training a decoder to decide "A vs. B" in context 1 and testing the decoder in context 2 and vice-versa (the CCGP is the average between these two decoders). Stimulus Parallelism Score (Fig. 2.3c,f) is the angle between the two coding vectors "A vs. B" in context 1 and 2.

For context, decoding accuracy is estimated by training two decoders to decide "Context 1 vs. Context 2" for each of the two stimuli in a stimulus pair. The reported decoding accuracy is the average between these two decoders (Fig. 2.E7a,b). For example, for the stimulus pair AB, one such decoder each is trained for all "A" trials and all "B" trials. Context CCGP (Fig. 2.3g, 2.E7c) is calculated by training a decoder to differentiate between Context 1 and 2 based on the trials in the first identity of the pair, and tested in the second pair and vice-versa. The reported Context CCGP value for a given stimulus pair is the average between the two. Similarly, context Parallelism Score (Fig. 2.3h, 2.E7d) is the angle between the two coding vectors Context 1 vs. Context 2 estimated separately for the first and second stimulus in a pair.

Distance/Variance Analysis (Fig. 2.4): We computed a series of metrics to quantify aspects of the population response that changed between inference absent and inference present sessions. We used (i) the firing rate, (ii) distance in neural state space between classes for balanced dichotomies and stimulus dichotomies (dichotomy distance), (iii) variance of neural spiking projected along the coding directions for those dichotomies (coding direction variance), and (iv) the condition-wise fano factor.

Firing rate (Fig. 2.4e) was the mean firing rate averaged across all neurons during the stimulus period, reported separately for correct trials of every unique task condition. Values reported during the baseline (Fig. 2.E9q,r) are computed with an identical procedure using firing rates from before 1s prior to stimulus onset.

Dichotomy distance (Fig. 2.4g,h,j,k) was defined as the Euclidean distance in neural state space between the centroids of the two classes on either side of the decision boundary for that dichotomy. Centroids were computed by constructing the average response vector for each class using a balanced number of correct trials from every condition included in each class through a resampling procedure (described below). Null distributions reported for dichotomy distances are geometric null distributions.

Coding direction variance (Fig. 2.4i) was computed for a given balanced dichotomy by projecting individual held-out trials onto the coding vector of the decoder trained to differentiate between the two groups of the balanced dichotomy being evaluated. The coding direction was estimated by training a linear decoder on all trials except eight (one from each condition either side of the dichotomy). The vector of weights estimated by the decoder (one for each neuron) was normalized to unit magnitude to estimate the coding vector. The projection of the left out trial onto this coding vector was than calculated using the dot product. This process was repeated 1000 times, generating a distribution of single trial projections onto the coding vector for each dichotomy. The variance of the distribution of 1000 projected data point was then computed and reported as the variance for a given balanced dichotomy (Fig. 2.4i).

The condition-wise Fano factor (Fig. 2.4f) was computed separately for each neuron. We used all correct trials for a given balanced dichotomy to estimate the mean firing rate and standard deviation and then took the ratio between the two to calculate the Fano factor for each neuron. Reported fano factors are the average of all fano factors across all neurons from that area/behavioral condition. Fano factors are computed by-condition since grouping trials across conditions could lead to task variable coding (signal) contaminating the fano-factor measurement, which should ideally only reflect trial-by-trial variation around the mean for approximately poisson-distributed firing rates.

The context-modulation consistency (Fig. 2.4l) was also computed separately for each neuron. Context modulation consistency is the tendency for a neuron's firing rate to shift consistently (increase or decrease) to encode context across stimuli. For each neuron, it was computed by deteriming the sign of the difference (+/-) between the mean firing rate for a given stimulus between the two contexts, and summing the number of stimuli that exhibit the same modulation (either increase or decrease) across the two contexts. This consistency can take on values between 0 (increase in firing rate to encode context for half of the stimuli, decrease in firing rate for the other half) and 4 (either increase or decrease in firing rate for all four stimuli).

Bootstrap Re-sampled Estimation of Measures and Null Distributions: All the measures described in the preceding sections were estimated using a trial and neuron-based re-sampling wmethod. This resampling strategy was used to assure that every measure reported is comparable

between a set of conditions by assuring that the same number of neurons and data points are used to train and test classifiers. Metrics were re-computed 1000 times with resampling and all null distributions were computed with 1000 iterations of shuffling and re-computing. Plotted boundaries of null distributions correspond to the 5^{th} and 95^{th} percentiles as sestimated from the 1000 repetitions. A single iteration of the re-sampling estimation procedure proceeds as follows. For all analyses that involved a comparison of a metric between two behavioral conditions (inference absent vs. inference present or Session One vs Session Two), the same number of neurons was included in both conditions by on a region by region basis. For a neuron to be included, at least 15 correct trials for each of the 8 unique task conditions had to exist (120 correct trials total). Across patients, the number of correct trials per condition varied: min = 10.9 ± 1.3 trials/condition, mean = 25.0 ± 0.6 trials/condition, max = 39.6 ± 1.2 trials/condition (mean ± s.e.m.). After identifying the neurons that met this inclusion criteria, an equal number were randomly sampled from both behavioral conditions. The number of considered neurons was set to the number of neurons available in the smallest group.

When constructing feature matrices for decoding, 15 trials were randomly selected from each unique condition that was included in the given analysis. Trial order was shuffled independently for every neuron within condition to destroy potential noise correlations between neurons that were simultaneously recorded. For decoding and Shattering Dimensionality, out-of-sample accuracy was estimated with 5-fold cross validation. For generalization analyses (CCGP), all trials were used in training since performance is evaluated on entirely held-out conditions. For vector-based measures (dichotomy distance, variance, Parallelism Score), all trials in relevant conditions were used to compute condition centroids. In the case of variance estimation, all trials except one on either side of the dichotomy boundary were used to learn the coding axis, then the held-out trials were projected onto the coding axis. As previously stated, these procedures were repeated 1000 times with independent random seeds to ensure independent random sampling of neurons and trials across iterations.

Figures:

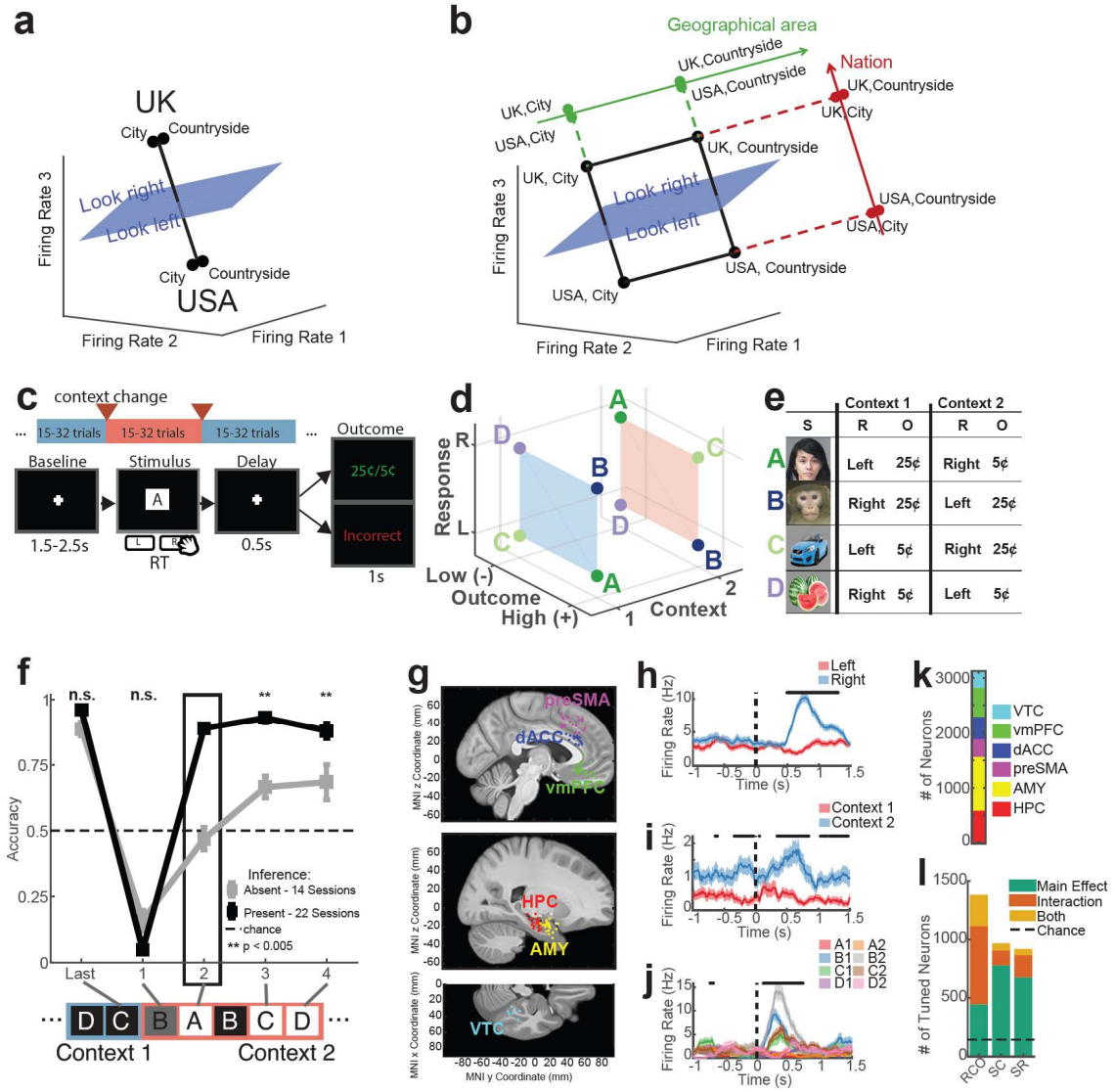

Figure 2.1. Task, behavior, recording locations, and single-neuron tuning. (a-b) Illustration of two possible definitions of abstraction. **(a)** Abstraction defined as clustering. Only nation, but not geographical area, is preserved. **(b)** Abstraction defined as generalization. Both geographical area and nation is preserved orthogonally to each other, facilitating left-right looking generalization (blue plane) without discarding geographic area information. **(c)** The task consisted of variable-length blocks (15-32 trials) that alternated between two latent contexts (red and blue). Context changes (red arrows) were covert. Trials consisted of a pre-stimulus baseline followed by stimulus presentation during which patients executed the associated response (left or right button press) in a speeded manner. After button press, the stimulus was replaced with a fixation cross, followed by the outcome (either high/low reward or incorrect) was presented after a fixed 0.5s delay. **(d)** Illustration of the task structure. Each stimulus (A-D) is associated with a single correct response and results in either a high or low reward if the correct response is given. All stimulus-response relationships are inverted between context 1 (blue) and 2 (orange). This visualization is reflective of the disentangled structure of the task variables, and does not necessarily reflect how neurons

will organize their responses in neural state-space to each of these conditions. **(e)** Example images (left) associated with the stimuli A-D. Note: stimuli A and B are masked due to copyright. These associations are randomized for every session. **(f)** Task performance split by whether inference was present or absent on the first inference trial following context switches in a given session. Sessions where inference performance was significantly above chance (22 sessions, $p < 0.05$, Binomial Test on inference trial 1) were deemed "inference present" (blue), and those where inference performance was not above chance (14 sessions, $p > 0.05$, Binomial Test on inference trial 1) were considered "inference absent" (red). Plot shows performance on the last trial before the context switch, the first trial after the context switch, and for the remaining three inference trials averaged over all trials in each session (mean ± s.e.m. across sessions). Dashed line marks chance. Black box indicates inference trial 1. **(g)** Electrode locations. Each dot corresponds to a single microwire-bundle. Locations are shown on the same hemisphere (right) for visualization purposes only. Shown are pre-Supplementary Motor Area (preSMA, purple), dorsal Anterior Cingulate Cortex (dACC, blue), ventromedial Prefrontal Cortex (vmPFC, green), Hippocampus (HPC, red), Amygdala (AMY, yellow), and Ventral Temporal Cortex (VTC, teal). **(h-j)** PSTH of three example neurons that encode response **(h)**, context **(i)**, and mixtures of stimulus id and context **(j)**. Stimulus onset occurs at time 0. Black points above PSTH indicate times where 1-way ANOVA over the plotted task variables was significant ($p < 0.05$). **(k)** Number of single units recorded across all brain areas (3124 neurons recorded in total). **(l)** Number of single units across all brain areas exhibiting significant Main effects or interaction effects (n-way ANOVA with interactions, $p < 0.05$, see methods) to at least one of the principal task variables (R = Response, C = Context, O = Outcome, S = Stimulus ID) or to combinations of variables. A unit is linearly tuned if it has at least one significant main effect, and non-linearly tuned if it has at least one significant interaction term in the ANOVA model.

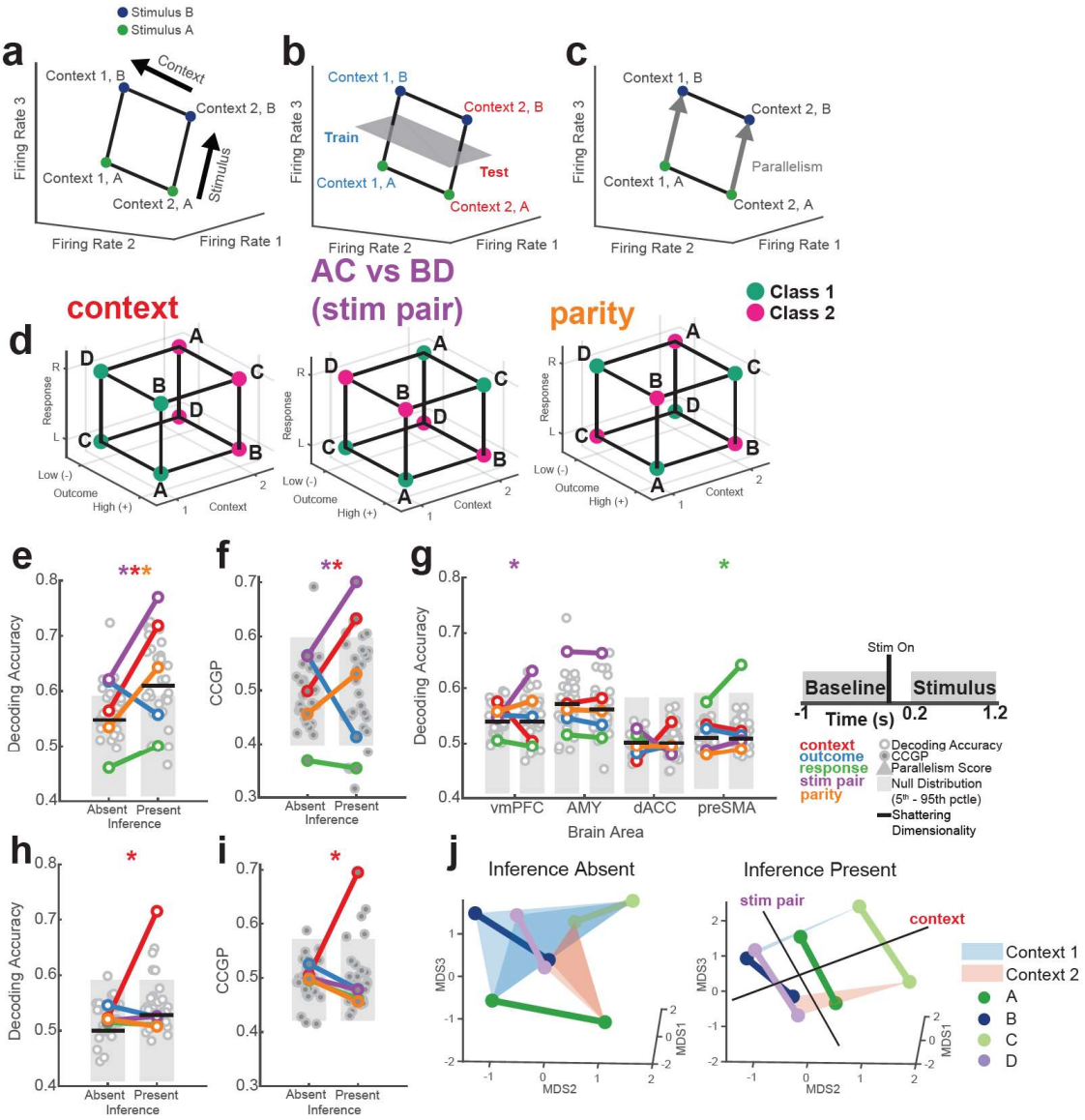

Figure 2.2. Emergence of multiple abstract variables in hippocampus supports inference.
(a) Simplified example of a neural state space where each axis is the firing rate of one neuron. Points correspond to the response of the neurons to different task states, i.e. two stimuli (green and orange) that elicit two responses (R and L) in two contexts. Note the coding vectors for response and context (black arrows) are not aligned with the axes as each neuron might respond to mixtures of variables. The axes of this state space differ from those shown in Fig. 2.1b, with the latter being defined by experimenter-selected variables rather than neural firing rates.
(b) Example of cross-condition generalization. A decoder is trained to classify context only on response "R" conditions (green) and is evaluated on its ability to decode context on response "L" conditions (purple). If context is represented in an abstract format (i.e. disentangled from response), then the decoder should generalize to the held-out response condition, yielding a high cross-condition generalization performance (CCGP) for context.
(c) Example of context parallelism. Coding vectors for context (gray arrows) are parallel, indicating that the coding direction for context is identical for different responses, and thus that context and

response are disentangled. Details for computing the Parallelism Score of a balanced dichotomy with 8 conditions are provided in the methods.

(d) Balanced dichotomies of task conditions that correspond to important task variables, including context (red), behaviorally-relevant stimulus grouping (stim pair, purple), and parity (orange). Class labels for binary classification are indicated with green and magenta. Class assignment is arbitrary, and labels can be inverted without loss of generality. See Fig. 2.E2 for a complete account of labeled balanced dichotomies.

(e-g) During the stimulus presentation period, context (red) and stimulus pair (purple) become decodable in inference present sessions in the HPC. Context is encoded in an abstract format. Decoding accuracy **(e)** and CCGP **(f)** are shown for all 35 balanced dichotomies during the stimulus period (0.2 to 1.2 s following stimulus onset, see inset). A subset of the dichotomies are named (color code) because they represent task specific variables (see Fig. 2.E2). Swarm plots for decoding accuracy and CCGP are light circles and dark circles respectively. Shattering dimensionality (average dichotomy decodability) is shown with horizontal black lines. Gray bars denote the 5^{th}-95^{th} percentile of the shuffle-null distribution for decoding accuracy and geometric null distribution for CCGP. Stars denote named dichotomies that are above chance in inference present sessions and are significantly different from their corresponding inference absent value ($p_{RS} < 0.05/35$, Ranksum Test, Bonferroni corrected for multiple comparisons across all dichotomies).

(g) Identical analysis to **(e)** showing decodability of balanced dichotomies from neurons recorded in other brain regions (except VTC, which is shown in Fig. 2.E3).

(h-i) Same as **(e,f)**, but for spikes counted during the baseline period prior to stimulus onset. Context (red) becomes decodable in inference present sessions and is in an abstract format. Trials are labeled according to the current trial. Decoding accuracy **(h)** and CCGP **(i)** computed in HPC for all balanced dichotomies with spikes counted during the pre-stimulus baseline period (-1 to 0s prior to stimulus onset, see inset). All plotting conventions identical to those in **(e-g)**, except Baseline analysis is conducted with task variables from previous trial.

(j) Three-dimensional projections of hippocampal neural responses to task conditions during the stimulus period in inference absent (left) and present (right) sessions generated by performing Multi-Dimensional Scaling on neural data. Points correspond to unique task conditions identified by the associated stimulus and context color consistent with Fig. 2.1b. Hypothetical decoders for stim pair and context are shown for schematic purposes (black lines).

Note: all reported geometric measures, decoding accuracies, or angles are the average of 1000 runs with condition-wise trial resampling as described in the methods. All null distributions are constructed from 1000 iterations of shuffled trial-resampling using either trial-label shuffling (shuffle null) or random rotations designed to destroy low-dimensional structure (geometric null). Also, neuron counts are balanced between inference absent and inference present sessions for every brain area to ensure that dimensionally-sensitive values (e.g. vector angles, decoding accuracies, etc..) are directly comparable. See methods for details.

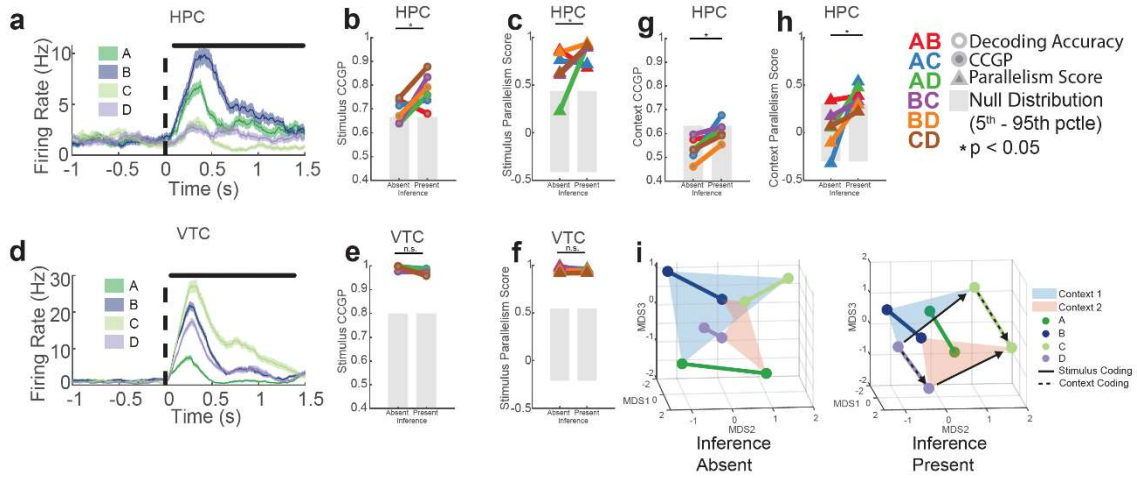

Figure 2.3. Stimulus representations become structured around context with inference in HPC but not VTC.

(a-c) Responses in HPC following stimulus onset carry information about stimulus identity. **(a)** Example PSTH of a neuron in the HPC that encoded stimulus identity.

(b,c) Stimulus geometry across contexts, with geometric analysis conducted over pairs of stimuli in each context. Data points shown correspond to different stimulus pairs (color coded, see right for legend). Significance of differences is tested using RankSum comparing inference absent and present over all stimulus pairs (* indicates $p < 0.05$, n.s. otherwise). All other conventions identical to those in Fig. 2.2.

(b) CCGP ($p_{RS} = 0.041$) and **(c)** parallelism score ($p_{RS} = 0.040$) for stimulus coding across contexts significantly increased in inference present compared to inference absent sessions.

(d-f) Same as **(a-c)**, but for VTC.

(d) CCGP ($p_{RS} = 0.15$) and **(e)** parallelism score ($p_{RS} = 0.39$) for stimulus coding across contexts does not differ significantly between inference absent and inference present sessions.

(g-h) Context encoding across stimulus pairs for HPC. Plotting conventions are identical to those in panels (b-c). **(g)** CCGP for context across stimuli ($p_{RS} = 0.012$) and **(h)** parallelism score for context coding vectors between pairs of stimuli ($p_{RS} = 0.015$) both significantly increase from inference absent to inference present sessions.

(i) Changes in neural geometry in HPC. MDS of condition-averaged responses of all recorded HPC neurons shown for inference absent (left) and inference present (right) sessions. Colored points are average population vector responses to a stimulus (point color) in each context (plane color). Stimuli of the same identity in either context are connected by a line of the same color. Abstract coding of stimulus across contexts (solid arrows) and context across stimuli (dashed arrows) are highlighted for a randomly selected pair of stimuli (C and D). This data in this plot is identical to Fig. 2.2j.

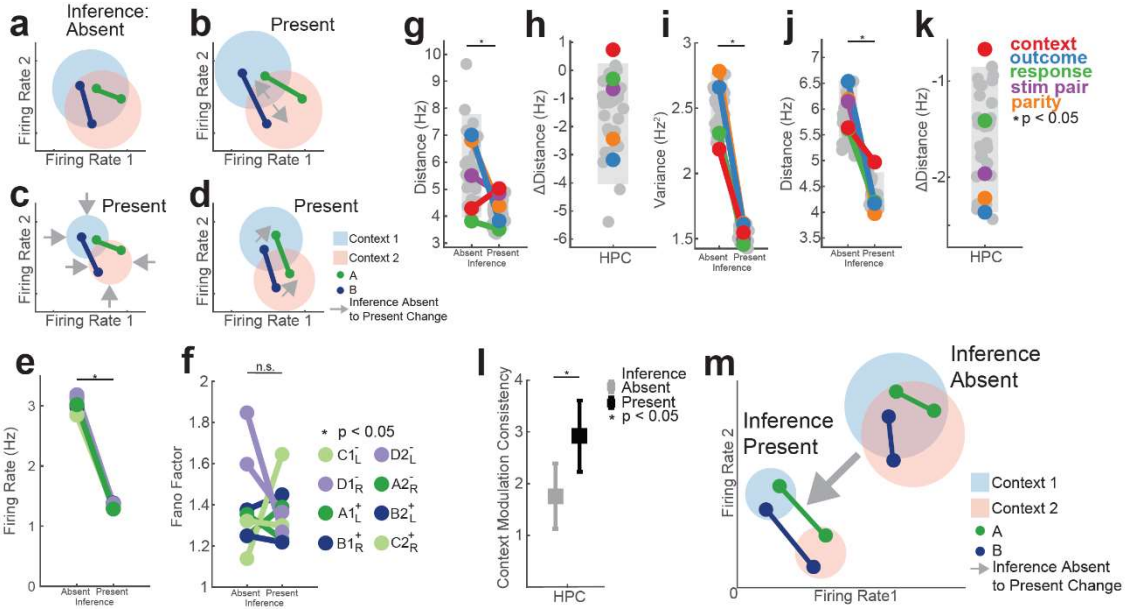

Figure 2.4. Firing rate properties underlying the observed changes of population-level hippocampal neural geometry.

(a-d) Illustration of different hypothesized firing rate pattern changes that could give rise to the observed population level geometry changes. The four different hypotheses are illustrated with two stimuli (A and B), each present in two different contexts (blue and red). Condition responses are defined by the firing rates of two hypothetical neurons. The solid-colored points represent the condition average for each stimulus and the larger shaded circles represent the trial-by-trial variation of the two neurons for each context. Gray arrows signify changes that have occurred in inference present plots **(b-d)** relative to the inference absent plot **(a)**.

These response of the neurons to the stimuli during inference absent sessions **(a)** can be shaped by **(b)** increasing the distance between the context centroids, **(c)** decreasing the variance along the coding direction in the absence of changes in distance, or **(d)** straightening the neural responses without changing distances/variances so that the geometry becomes more orthogonalized.

(e) Changes in hippocampal firing rate from inference absent to present sessions. Points correspond to the average firing rate over neurons in the HPC for each of 8 unique task conditions, and are colored according to stimulus identity in that condition (e.g. task condition $C1_L^-$ describes: stimulus C, context 1, outcome $-$, response L). Neuronal firing rates were lower during inference present compared to inference absent sessions ($p_{RS} = 8.3x10^{-5}$, RankSum over conditions).

(f) Same as **(e)**, but for condition-wise fano factors. Fano factor (FF) here is computed as the ratio of the condition-wise variance and the condition-averaged firing rate, computed by neuron and averaged over neurons. Points correspond to average FF over all hippocampal neurons. There was no significant difference (RankSum over conditions between inference absent and present sessions, $p_{RS} = 0.99$).

(g) Population distances between centroids for all 35 balanced dichotomies. The colored connected points represent distances for the named dichotomies indicated in the legend to the right. Gray bars indicate the 5th-95th percentile of the geometric null distribution. Across all dichotomies, distances decreases from inference absent to present ($p_{RS} = 2.9x10^{-8}$, RankSum over dichotomies).

(h) Context alone is the only dichotomy whose distance significantly increases from inference absent to present **(**red, $p_{\Delta Dist} = 0.040$**)**. The null distribution shown in (h) is the distribution of differences between the inference present and inference absent null distributions shown in (g).
(i) Average variance projected along the coding direction decreased on average between inference absent and inference present sessions ($p_{RS} = 6.5x10^{-1}$). Variance was computed in a cross-validated manner (see methods) resulting in a distribution of trial-by-trial population activity along each dichotomy coding direction from which the coding variance was computed. The 5^{th}-95^{th} percentile of the geometric null distribution was also used here for the null distribution.
(j,k) Same as **(g,h)**, but for spike counts during the baseline period and grouping trials by task state of the previous trial. Distance was significantly reduced across all dichotomies (**j**, $p_{RS} = 6.4x10^{-1}$, RankSum over dichotomies) and context alone exhibits a distance reduction that is smaller than would be expected by chance (**k,** red, $p_{\Delta Dis} = 0.027$)
(j) Change in the consistency of context-modulation for stimuli averaged over all neurons in HPC. Greater context modulation consistency for individual neurons results in greater parallelism score for context at the population level. HPC neurons on average exhibit a significant increase in context modulation consistency between inference absent and inference present sessions ($p_{RS} = 0.0039$) during the stimulus period.
(m) Illustration of implementational changes to neural state space using the conventions introduced in **(a-d)**. We find that, when comparing inference absent with inference present sessions, that (i) context dichotomy distance increased (indicated by the increased distance between the red and blue shaded circles), (ii) variance decreased due to a reduction in firing rate (indicated by decreased shaded circle radius and movement towards the origin of state space), and (iii) an increase in the consistency of stimulus modulation across contexts (indicated by lines becoming parallel).

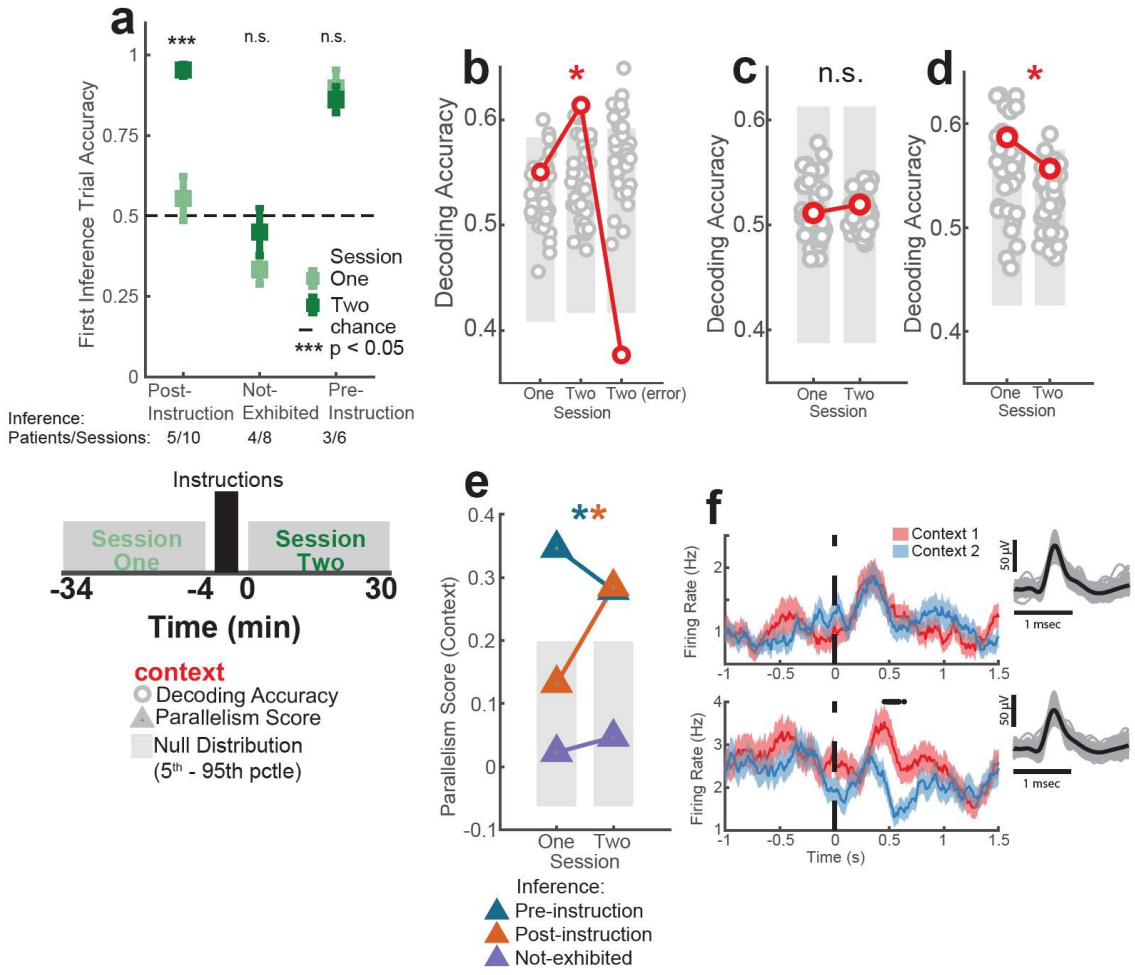

Figure 2.5. Abstract hippocampal representation of context is present following successful verbal instructions about latent context.

(a) Behavioral performance on the first inference trial shown for three separate groups of peri-instruction sessions where patients exhibited inference either after receiving high-level instruction (post-instruction inference), before receiving high-level instruction (pre-instruction inference), or who were never able to exhibit inference (inference not-exhibited). The session before and after high-level instructions are labeled as Session One (red) and Session Two (blue) respectively (See inset). Data are identical to the black-boxed data in Fig. 2.E1g-h.

(b,c,d) Encoding of context in the stimulus period in Sessions One during correct trials (One), Session Two correct trials (Two) and Session Two error trials (Two (error)). The first trial following a switch is excluded from this analysis. * indicates $p < 0.05$ against null in any column of a given geometric measure plot, and n.s. otherwise. **(inset)** Schematic of the recording procedure, showing Sessions One and Two shaded in gray (30 min duration), with a 4 minute inter-session break (mean duration = 241 s, range 102-524s) during which instructions detailing task structure were provided.

(b) Context emerges as significantly decodable in Session Two but not Session One in the post-instruction inference group in a task-performance-dependent manner ($p_{One} = 0.17, p_{Two} = 0.016, p_{RS} = 3.1x10^{-19}, p_{Two\ (error)} = 0.99$). **(c)** Context is not significantly decodable in the inference not-exhibited group neither Session one nor two ($p_{One} = 0.44, p_{Two} = 0.42$). **(d)** Context is decodable in the Pre-instruction inference group ($p_{One} = 0.014, p_{Two} = 0.17$).

(e) Summary of changes in parallelism score for context for all three session groups (pre-instruction green, post-instruction orange, not-exhibited purple). Neuron counts are sub-sampled to match across all groups so that parallelism score values are directly comparable. Significant increase in context parallelism score from session one to two is indicated for the post-instruction inference group ($p_{Post-Instruct\ \ ,One} = 0.20, p_{Post-Instruct\ \ ,Two} = 0.0028$), but not for the inference not-exhibited ($p_{Not-Exhibite\ ,One/Two} < 0.5$) and pre-instruction inference ($p_{Pre-Instruct\ \ ,One/Two} < 0.005$) groups.

(f) Example hippocampal neuron with univariate context encoding in the session after (bottom) but not before (top) instructions. (one-way ANOVA, $p_{One} = 0.40, p_{Two} = 0.010$).

Extended Data Figures:

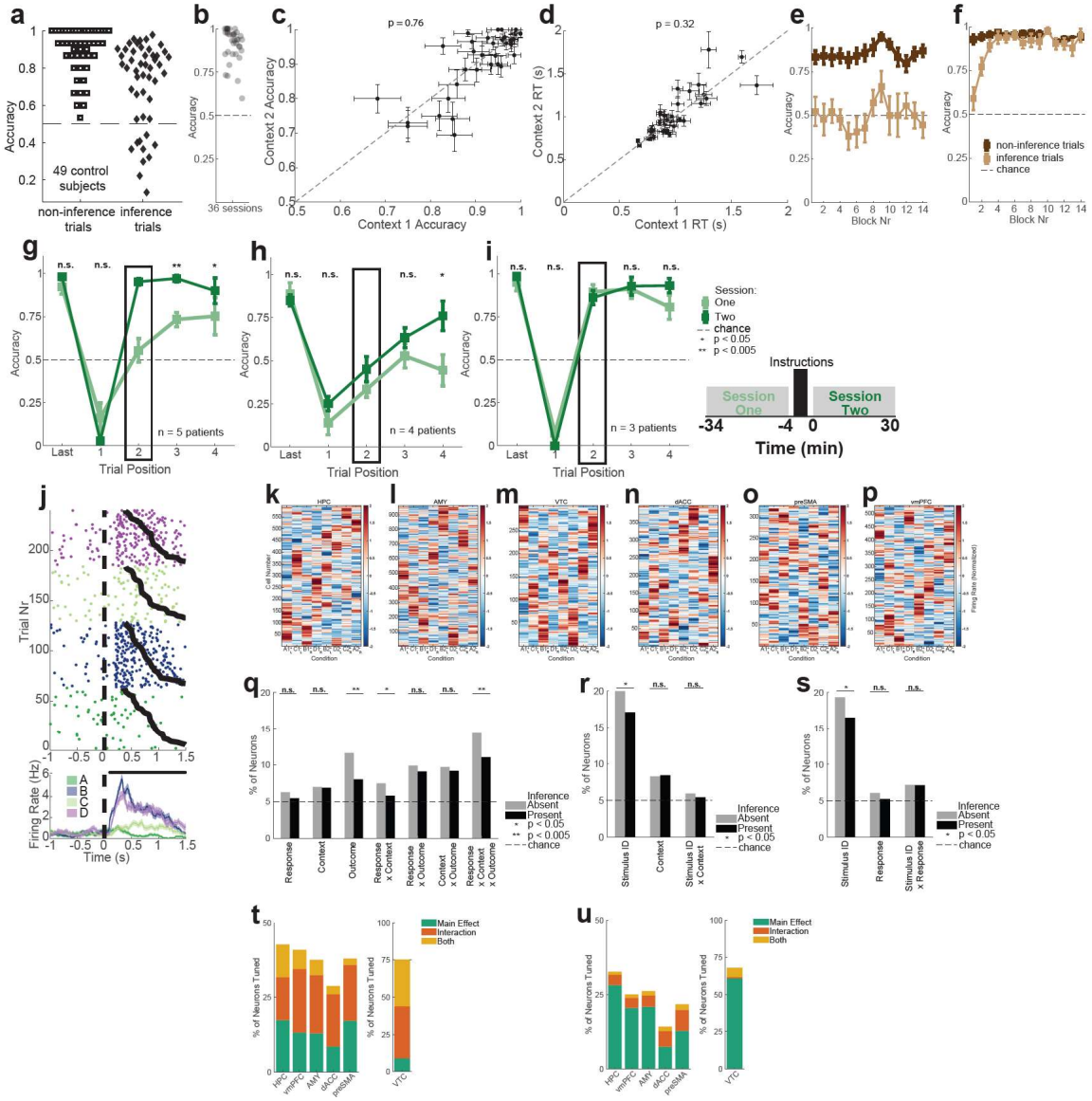

Figure 2.E1. Task behavior and single-neuron responses across all recorded regions.
(a) Task performance on individual sessions from 49 control subjects recruited through an online platform (Amazon MTurk). Accuracy is reported as an average for each subject over all non-inference trials (left) and inference trials (right). The horizontal gray dashed line corresponds to chance (50%). This task variant is equivalent to the first session of the task encountered by patients where they were given general instructions about learning stimulus-response mappings, but were not informed of the latent task structure. Subjects exhibited a variety of behaviors, with 46/49 subjects performing above chance on non-inference trials, indicating that the SRO maps were generally learnable despite the wide variation in performance on inference trials.
(b) Patients exhibited high accuracy on non-inference (baseline) trials. Each dot corresponds to the average non-inference trial performance over a single session. Black dashed line indicates chance. Only sessions where patients exhibited above-chance accuracy on non-inference trials are shown (36/42 sessions, $p < 0.05$, Binomial Test on all non-inference trials).

(c) Non-inference performance for context 1 is plotted against context 2 for each of the 36 sessions included in the analysis. Error bars correspond to SEM computed over blocks. The diagonal gray dashed line indicates identical block performance (y=x). The reported p-value is computed by paired t-test between the mean accuracies for Context 1 and Context 2 across all sessions.
(d) Same as (c), but with reaction time (RT), computed as time from stimulus onset to button press for every trial. Mean RT's are also computed by block.
(e-f) Task performance as a function of time in the task for the **(e)** inference absence and **(f)** inference present groups. Shown is the accuracy for the last non-inference trial before a switch (black) and the first inference trial after a switch (gray). Accuracy is shown block-by-block averaged over a 3-block window (mean ± s.e.m. across sessions).
(g-i) Behavioral performance plot similar to Fig. 1E. Plot shows performance on the last trial before the context switch, the first trial after the context switch, and for the first inference trial (Trial 2) averaged over all trials in each session (mean ± s.e.m. across sessions). Dashed line marks chance. Red and blue lines correspond to session performance before and after instructions detailing latent context are provided.
(g) This plot shows performance for the post-instruction inference session group – did not exhibit significant inference performance during Session One (before high-level instruction), but did exhibit inference performance during Session Two (following instruction, see Inset). First inference trial performance (block box) was used to classify patients, so difference significance is not computed. All trials where Session One/Two performance difference was insignificant ($p > 0.05$) are shown with n.s.
(h) Same as (g), but for the inference not-exhibited session group – did not exhibit significant inference performance during either Session One or Two.
(i) Same as (g), but for the pre-instruction inference session group – exhibited significant inference performance during both pre-instruction and post-instruction sessions.
(j) Example hippocampal neuron that encodes stimulus identity. Raster trials are reordered based on stimulus identity, and sorted by reaction time therein (black curves). Stimulus onset occurs at time 0. Black points above PSTH indicate times where 1-way ANOVA over the plotted task variables was significant ($p < 0.05$).
(k) Normalized activity for all neurons recorded from the hippocampus is plotted as a heat map by region after computing the trial-averaged response to each unique condition (8 total, specified by unique Response-Context-Outcome combinations). Z-scored firing rates are computed from 0.2s to 1.2s after stimulus onset for every trial. Each row of the heat map corresponds to the activity of a single neuron, and columns correspond to each of the 8 conditions. Neurons are ordered such that adjacent rows (neurons) are maximally correlated in 8-dimensional condition response space. This approach would allow for modular tuning to visibly emerge in the heat map if groups of neurons were clustered in their response profiles. Clearly, the responses here are very diverse.
(l) Same as (G), but for amygdala.
(m) Same as (G), but for ventral temporal cortex.
(n) Same as (G), but for dorsal anterior cingulate cortex.
(o) Same as (G), but for pre-supplementary motor area.
(p) Same as (G), but for ventromedial prefrontal cortex.
(q) Percentage of neurons across all areas that exhibit tuning to each of the three binary variables manipulated in the experiment. Tuning was assessed by fitting either a 2x2x2 (Response-Context-Outcome) ANOVA for every individual neuron's firing rate during a 1s window during the stimulus presentation period. Significant neurons were counted as $p < 0.05$ for main effects (Linear) or

interaction effects (Nonlinear) involving the stated variables. Significance in the change of percentage of neurons exhibiting modulation to each factor is determined via z-test, where "*" indicates $p < 0.05$, "***" indicates $p < 0.005$, and "n.s." indicates "not significant".

(r) Same analysis as **(q)**, but for a 4x2 ANOVA for stimulus identity and context.

(s) Same analysis as **(q)**, but for a 4x2 ANOVA for stimulus identity and response.

(t) Same analysis as Fig. 2.1j, but with percentages of tuned neurons shown separately for each region. Single-neuron tuning identified here using 3-Way ANOVA (Response x Context x Outcome), corresponding to column 1 (RCO) of Fig. 2.1j.

(u) Same as **(t)**, but single-neuron tuning identified here using 2-Way ANOVA (Stimulus ID x Context), corresponding to column 2 (SC) of Fig. 2.1j.

Note: the corresponding analysis between stimulus identity and outcome cannot be conducted since those variables are correlated by task construction.

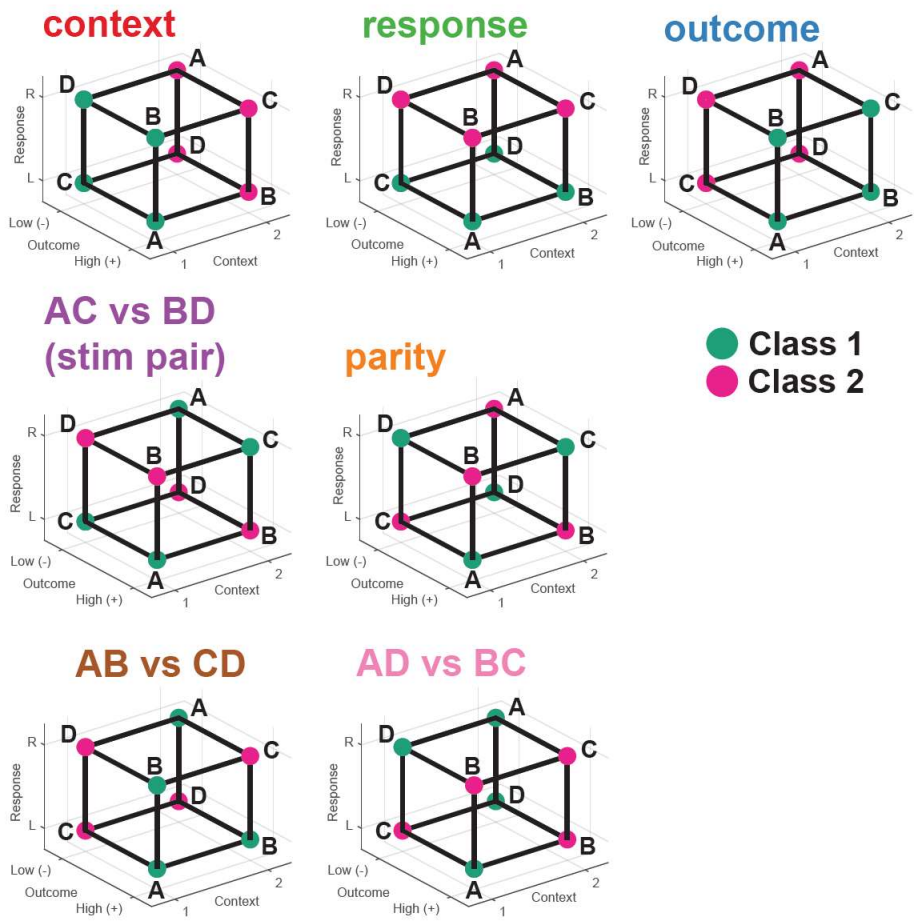

Figure 2.E2. Visual representation of all named balanced dichotomies.

Named balanced dichotomies correspond to condition splits that have clearly interpretable meaning with respect to the construction of the task when evaluating the decodability or disentanglement (Cross Condition Generalization Performance, Parallelism Score) for that dichotomy. For example, the context dichotomy (top left), arises from assigning all conditions for context = 1 to one class and all conditions for which context = 2 to the other class. Performing binary classification on neural responses with trial labels arranged in this way corresponds to decoding context from the neural population. The specific assignment of class labels 1 and 2 is arbitrary, and inverting the labels still corresponds to the same meaning for the dichotomy. All named dichotomies shown here are color coded to reflect their value in all Shattering Dimensionality, CCGP, and Parallelism Score plots, and this color code remains consistent throughout the paper whenever balanced dichotomies are considered.

The "stim pair" dichotomy corresponds to the special split of stimulus identities where the stimuli that have the same response in each context are grouped together (e.g. Stimuli A and C have Response L in Context 1 and Response R in Context 2, v.v. for Stimuli B and D). There are more balanced dichotomies that correspond to splits of stimulus identity (AB vs CD and AD vs BC). High decodability of any one of these balanced dichotomies reflects stimulus-id coding in the neural population.

The "parity" dichotomy is another special dichotomy that corresponds to the most difficult, or non-linear, dichotomy that can be constructed, and high decodability of this dichotomy is a

signature of a high-dimensional representation. The term parity is in reference to the fact that, if task states were represented as 3-bit binary words where each bit corresponds to the value of the response, context, and outcome variable that describes the state (e.g. 000 for Left, Context 1, Low, 111 for Right, Context 2, High), then one class of the parity dichotomy corresponds to all states with an even number of ones, and the other class corresponds to all states with an odd number of ones. Note, for this dichotomy, no node of a given class shares an edge with another node of the same class. If one views the faces of the cube, one can see that the standard 2D XOR dichotomy between class 1 and 2 is present on every face. The notion of parity, can be generalized to arbitrarily many dimensions.

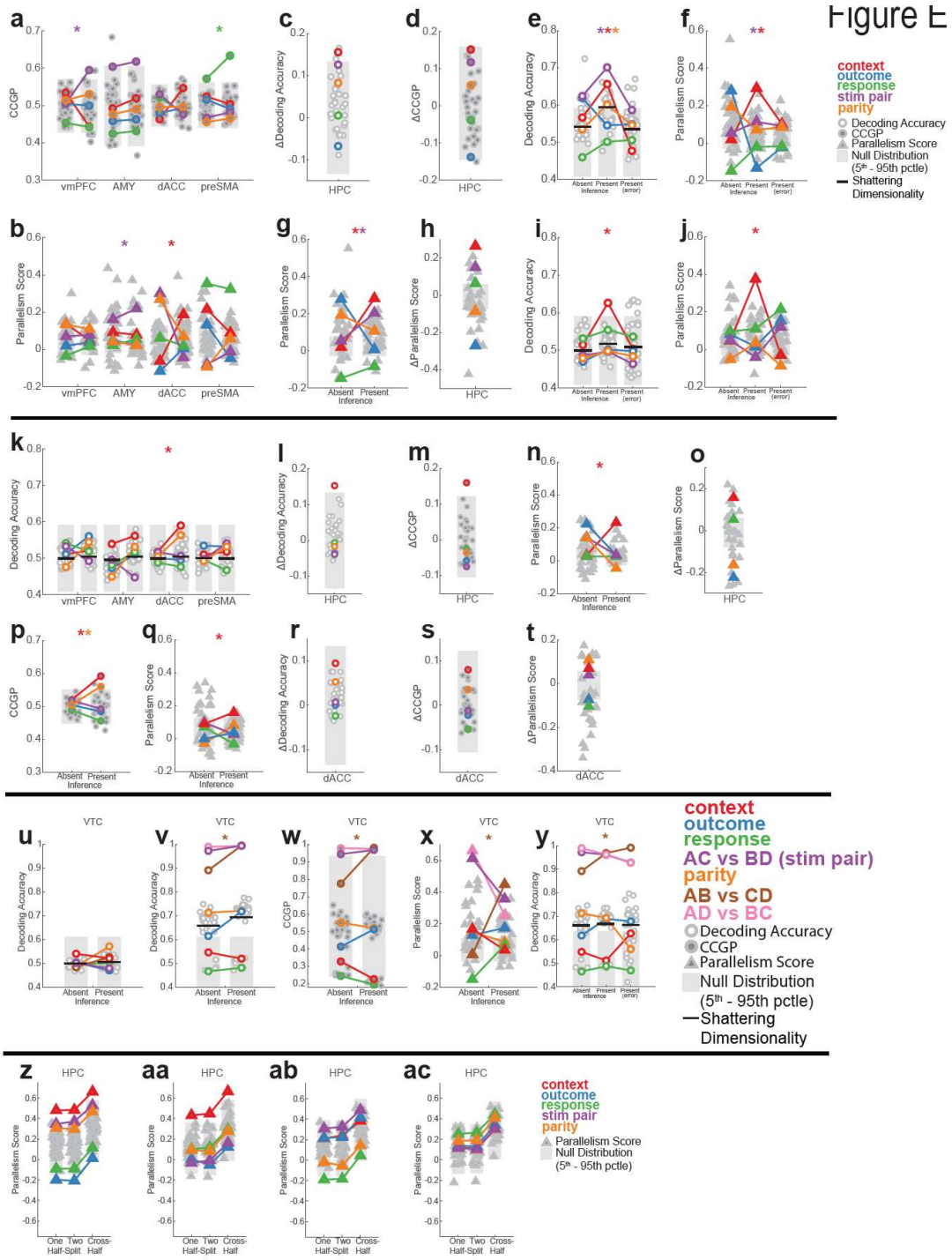

Figure 2.E3. Additional geometric analysis during stimulus processing and baseline periods. **(a)** Cross-condition generalization performance (CCGP) reported for other regions over balanced dichotomies. Each dot corresponds to the CCGP for a single dichotomy. The reported values are averages over 1000 repetitions of resampling of trials and neurons as described in the methods. For every region, the left column corresponds to inference absent sessions and the right column to inference present sessions. Colored lines are drawn to connect values for named dichotomies in inference absent and present sessions using the standard color-coding scheme. The gray background bars indicate the 5th (bottom) to 95th (top) percentile of the null distribution. Note that the null

distribution differs by area due to the different number of neurons present in each area. Significant named dichotomies are marked when the dichotomies are: above 95th pctle of null in inference present (i.e. significantly above chance during inference), significantly different between inference absent and present (RankSum p < 0.01/35, Bonferroni corrected for balanced dichotomies). Significant increases were observed in vmPFC for stim pair (purple, $p_{Absent} = 0.45, p_{Present} = 0.014$) and preSMA for response (green, $p_{Absent} = 0.045, p_{Present} = 0.0010$). Note that stim pair CCGP in AMY was above chance for both inference absent and present sessions (purple, $p_{Absent} = 0.050, p_{Present} = 0.039$).

(b) Same as (a), but for Parallelism Score. Significant increases in Parallelism Score were present for stim pair in amygdala (purple, $p_{Absent} = 1.3x10^{-4}, p_{Present} = 9.0x10^{-8}$) and context in the dorsal anterior cingulate (red, $p_{Absent} = 0.99, p_{Present} = 3.8x10^{-12}$).

(c) Change in decoding accuracy computed as Inference present – Inference absent for every balanced dichotomy. The gray shaded bar again indicates 5th-95th pctle of null, which is populated by computing the difference for 1000 random pairs of dichotomies in the null distributions of the inference present and inference absent sessions computed separately. In general, the null distribution for any difference plot is computed by drawing such samples from the inference absent and present null distributions of the associated area and metric (e.g. the null distribution here is computed using the inference present and absent nulls in Fig. 2.2e, the null distribution for Fig. 2.E3d is computed from Fig. 2.2f, the null distribution for Fig. 2.E3h is computed from Fig. 2.E3g, etc…).

(d) Same as (c), but for changes in CCGP from inference absent to present.

(e,f) Following stimulus onset, context (red) is not decodable **(e)** and not in an abstract format **(f)** in incorrect trials occurring during inference present sessions. Decoding accuracy and parallelism scores are estimated separately during correct and error trials in inference present sessions (inference present and present (error), respectively) and in correct trials only during inference absent sessions (absent). Horizontal black bars indicate shattering dimensionality (average over dichotomies). Stars denote named dichotomies that are above chance in the inference present trials and are significantly different from their corresponding inference absent value ($p < 0.05/35$, Ranksum Test, Bonferroni multiple comparison corrected across dichotomies).

(g) Parallelism score plot for hippocampus in inference absent and present sessions. Coloring, plotting, and significance conventions are identical to those used for Decoding Accuracy and CCGP plots (e.g. Fig. 2.2c,d). Null distribution here (gray background bars) is computed using the geometric null, the same procedure as CCGP. We note that the 5th and 95th pctle of null is quite narrow since the rotation procedure used to generate the null distribution produces approximately orthogonal coding vectors in high dimensional spaces. Here, context was significantly elevated in inference present compared to absent sessions (red, $p_{Absent} = 0.55, p_{Present} = 1.4x10^{-15}$), as was stim pair (purple, $p_{Absent} = 0.17, p_{Present} = 1.7x10^{-8}$).

(h) Same as (C), but for changes in Parallelism Score from inference absent to present sessions.

(i,j) Same as **(e,f)**, but estimated for spikes counted during the baseline period and task states from previous trial. Note: inference present values here are slightly lower than the values reported in Fig. 2.2 because additional neurons were removed by subsampling to equalize the number of neurons used for both correct and error trials.

(k) Decoding accuracy reported for regions other than hippocampus, analogous to Fig. 2.2g, but for the baseline period instead of the stimulus period. Additionally, trial labels here correspond to the conditions (Response, Context, Outcome) encountered during the previous trial. Details regarding the reported decoding accuracies and null distributions are identical to those described in Fig. 2.2e. Decoding accuracy for all balanced dichotomies is reported for the inference absent (Left) and

inference present (right) conditions, with color-coded lines for named dichotomies connecting dots for each region. Significant increase from inference absent to present was observed in dACC for context (red, $p_{Absent} = 0.37, p_{Present} = 0.049$) . No significant changes in shattering dimensionality were present (inference absent vs inference present $p_{RS} > 0.05$ for all areas).

(l) Change in decoding accuracy for all balanced dichotomies in hippocampus associated with the presence of inference, again computed during the baseline period. Procedure for all reported accuracies and null distribution construction identical to that described in Fig. 2.E3c, except that the analysis used baseline firing rates and condition labels from the previous trial instead of the current trial. Here, context is the only balanced dichotomy whose increase in decodability is significantly above null.

(m) Change in cross-condition generalization performance (CCGP) for all balanced dichotomies in hippocampus. See Fig. 2.E3x for plotting details. Context is also the only dichotomy for which the CCGP rises significantly more than the 95th pctle of the geometric null distribution.

(n) Parallelism score for hippocampus. Plot is analogous to Fig. 2.E3e, but computed during the baseline with previous trial labels instead of during the stimulus with current trial labels. Context is the only named dichotomy for which the Parallelism Score increased to significance above the geometric null in the inference present condition. (red, $p_{Absent} = 0.37, p_{Present} = 1.2x10^{-10}$)

(o) Same as **(c,d)**, but for parallelism score. Context is the only named dichotomy to increase significantly in inference present sessions, and the other un-named dichotomies that also significantly rise are correlated with context.

(p) CCGP for balanced dichotomies in the dorsal anterior cingulate cortex (dACC). Associated decoding accuracy for balanced dichotomies in inference absent and present sessions shown in **(k)**. Here in the dACC, context (red, $p_{Absent} = 0.26, p_{Present} = 0.018$) is also found to be in an abstract format.

(q) Parallelism score for balanced dichotomies in the dACC. Here, context (red, $p_{Absent} = 0.18, p_{Present} = 0.013$) emerges as significant in inference present sessions.

(r) Change in decoding accuracy for balanced dichotomies in dACC with inference. The associated plot for inference absent and present is in **(k)**. Though context is the dichotomy with the greatest increase in decoding accuracy, it is still below the null 95th pctle in this case.

(s) Same as **(m)**, but for CCGP. Here, context is also the dichotomy with the greatest increase, but is still below null 95th pctle.

(t) Same as **(n)**, but for Parallelism Score. Increase in Parallelism Score for parity is notably significant ($p_{\Delta} = 0.0016$). Context also significantly increases ($p_{\Delta} = 0.026$).

(u-y) Ventral temporal cortex (VTC) strongly encodes high-level features of visual stimuli, necessitating the introduction of two new dichotomies that capture stimulus identity, while not directly corresponding to any of the principal manipulated variables in the task (Response, Context, Outcome). The AB vs CD and AD vs BC dichotomies are “stimulus” dichotomies in that they represent systematic differences in coding between unrelated stimuli arbitrarily paired together, unlike the AC vs BD dichotomy which pairs stimulus identities for which responses are identical across the two contexts. That is, A/C response is L in Context 1 and R in Context 2, and v.v. B/D response is R in Context 1 and L in Context 2. These are the images whose correct responses “switch together” across contexts. Note that the new stimulus dichotomies are correlated with other named dichotomies: AB vs CD is correlated with outcome and AD vs BC is correlated with the parity dichotomy. Thus, high AB vs CD decodability will lead to increased outcome decodability. However, CCGP is robust to these dichotomy correlations, and will be low for correlated

dichotomies even if decodability is increased. These three dichotomies (AB vs CD, AC vs BD, and AD vs BC) are particularly relevant to vtc given its strong stimulus representations.

(u) Dichotomy decodability during pre-stimulus baseline. None of the balanced dichotomies are decodable during inference absent or present ($p > 0.05$ for all dichotomies). Shattering dimensionality does not significantly differ between inference absent and present sessions (0.50 vs 0.51, p_{RS} =0.34).

(v) Dichotomy decodability during the stimulus presentation period. All three named stimulus dichotomies are highly decodable both during inference absent and inference present sessions. Correlated dichotomies also demonstrated above-chance decodability. Horizontal black bars indicate shattering dimensionality (inference absent vs present, 0.66 vs 0.70, $p_{RS} = 0.0056$). Dichotomies: purple, $p_{Absent} = 6.8x10^{-1}$, $p_{Present} = 6.6x10^{-14}$, brown, $p_{Absent} = 2.2x10^{-9}, p_{Present} = 6.0x10^{-14}$, pink,$p_{Absent} = 1.1x10^{-13}, p_{Present} = 6.7x10^{-14}$. Notably, context is not significantly decodable in either inference absent or inference present sessions (red, $p_{Absent} = 0.24, p_{Present} = 0.38$).

(w) Dichotomy CCGP for VTC during the stimulus presentation period. Two stimulus dichotomies are in an abstract format in inference absent and all three are in an abstract format in inference present (purple, $p_{Absent} = 0.0054, p_{Present} = 0.00\ 36$, brown, $p_{Absent} = 0.057, p_{Present} = 0.0029$, pink, $p_{Absent} = 0.0030, p_{Present} = 0.0032$). All remaining dichotomy CCGP values are at chance apart from response and context, which are significantly below chance in inference present sessions (green, $p_{Absent} = 0.93, p_{Present} = 0.96$, red, $p_{Absent} = 0.84, p_{Present} = 0.94$).

(x) Dichotomy Parallelism Score for VTC during the stimulus presentation period. Again, two stimulus dichotomies are in an abstract format in inference absent sessions, and all three are in an abstract format in inference present sessions (purple, $p_{Absent} = 0, p_{Present} = 4.3x10^{-13}$, brown, $p_{Absent} = 0.73, p_{Present} = 0$, pink, $p_{Absent} = 0, p_{Present} = 5.9x10^{-7}$).

(y) Dichotomy decodability analysis for incorrect trials during the stimulus presentation period. Decoders are trained on correct trials and evaluated on error trials (balanced by condition) in inference present sessions. Plotting conventions are identical to those described in Fig. 2.E3f. Note that all three stimulus identity-related dichotomies are still highly significantly decodable during error trials in inference present sessions (purple, $p_{Present\,(error)} = 7.8x10^{-11}$, brown, $p_{Present\,(error)} = 1.1x10^{-13}$, pink, $p_{Present\,(error)} = 8.7x10^{-11}$) and shattering dimensionality does not decrease (black bar, inference present vs present (error), 0.67 vs. 0.66, $p_{RS} = 0.65$).

(z) Parallelism score for context computed during the stimulus period for random half-splits of the inference present sessions (Left, Middle column, 11 sessions in each half). Cross-half context parallelism is also computed through cross-session neural geometry alignment (Right Column, see Methods). Baseline context parallelism is significantly above chance within each half and across halves ($p_{Half-Split\ One} = 0.0081, p_{Half-Split\ Two} = 0.0098, p_{Cross-Half} = 0.033$).

(aa) Same as **(z)**, but for the baseline period. Stimulus context parallelism is again significantly above chance within each half and across halves ($p_{Half-Split\ One} = 0.0029, p_{Half-Split\ Two} = 0.0022, p_{Cross-Half} = 0.010$).

(ab) Same as **(z)**, but for the inference absent sessions (7 sessions in each half) during the stimulus period.

(ac) Same as **(ab)**, but for the baseline period.

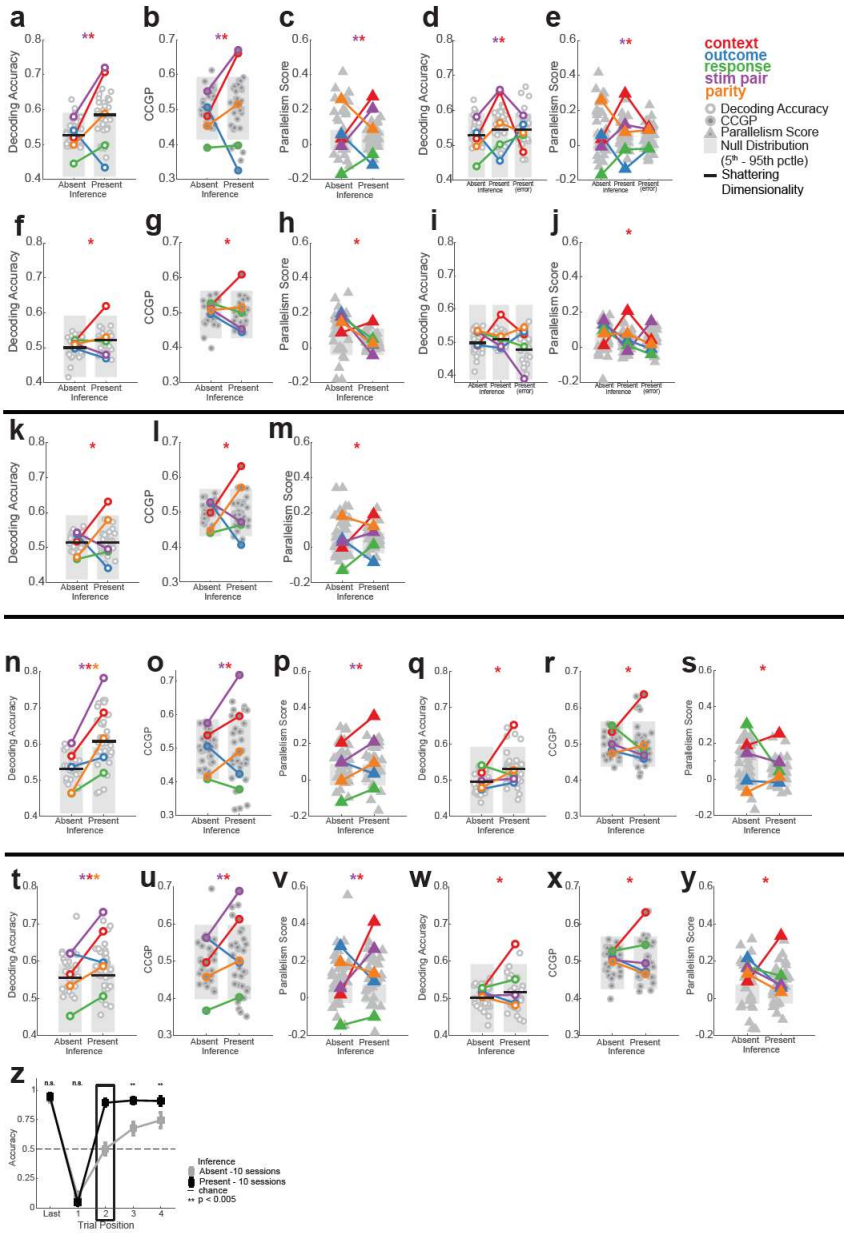

Figure 2.E4. Additional control analyses for Hippocampal representational geometry.
Identical analysis to the main geometric analysis shown in Fig. 2.2, except that hippocampal neurons are excluded from the analysis with the following criteria: in **(a-j)**, neurons with significant linear tuning for Context, Response, or Outcome (2x2x2 ANOVA, Any Main Effect $p < 0.01$), and in **(k-m)**, neurons with significant linear tuning for Stimulus Identity or Context (4x2 ANOVA, Any Main Effect $p < 0.01$).

Using the 3-Way ANOVA applied neuron-by-neuron, 455/494 neurons were retained for the stimulus period analysis **(a-c)** and 458/494 neurons were retained for the baseline period analysis **(d-f)**. All primary results for changes in hippocampal geometry were recapitulated apart from decodability of the parity dichotomy during the stimulus period **(a)**.

(a) Context decodability (red, $p_{Absent} = 0.36, p_{Present} = 0.0001, p_{RS} = 1.6x10^{-31}$). Stim pair decodability (purple, $p_{Absent} = 0.078, p_{Present} = 4.2x10^{-5}, p_{RS} = 6.6x10^{-31}$). Shattering dimensionality (black, 0.54 vs. 0.58, $p_{RS} = 0.0012$) during the stimulus presentation.

(b) Context CCGP (red, $p_{Absent} = 0.63, p_{Present} = 0.0016, p_{RS} = 5.2x10^{-34}$). Stim pair CCGP (purple, $p_{Absent} = 0.17, p_{Present} = 0.00095, p_{RS} = 5.3x10^{-34}$) during the stimulus presentation.
(c) Context Parallelism Score (red, $p_{Absent} = 0.40, p_{Present} = 3.7x10^{-13}$). Stim pair Parallelism Score (purple, $p_{Absent} = 0.83, p_{Present} = 1.2x10^{-7}$) during the stimulus presentation.
(d) Context decodability (red, $p_{Absent} = 0.36, p_{Present} = 0.0029, p_{Present\ (error)} = 0.64, p_{RS} = 1.5x10^{-2}$). Stim pair decodability (purple, $p_{Absent} = 0.071, p_{Present} = 0.0021, p_{Present\ (error)} = 0.062, p_{RS} = 2.0x10^{-5}$). Shattering dimensionality (black, inference present vs present (error), 0.56 vs. 0.55, $p_{RS} = 0.62$) during the stimulus presentation.
(e) Context Parallelism Score (red, $p_{Absent} = 0.40, p_{Present} = 4.6x10^{-15}, p_{Present\ (error)} = 0.012$) during the stimulus presentation.
(f) Context decodability (red, $p_{Absent} = 0.37, p_{Present} = 0.013, p_{RS} = 2.2x10^{-2}$) during the baseline. Shattering dimensionality (black, 0.50 vs. 0.52, $p_{RS} = 0.036$) during the baseline.
(g) Context CCGP (red, $p_{Absent} = 0.31, p_{Present} = 0.0044, p_{RS} = 1.9x10^{-3}$) during the baseline.
(h) Context Parallelism Score (red, $p_{Absent} = 0.12, p_{Present}$ =0.0055) during the baseline.
(i) Context decodability (red, $p_{Absent} = 0.55, p_{Present} = 0.12, p_{Present\ (error)} = 0.37$) during the baseline. Shattering dimensionality (black, inference present vs present (error), 0.51 vs. 0.49, $p_{RS} = 0.030$) during the baseline.
(j) Context Parallelism Score (red, $p_{Absent} = 0.66, p_{Present} = 8.5x10^{-9}, p_{Present\ (error)} = 0.30$) during the baseline.

Using the 2-Way ANOVA applied neuron-by-neuron, 412/494 neurons were retained for the stimulus period analysis **(k-m)**. The stim-pair dichotomy is no longer decodable after removal of all stimulus-identity tuned neurons, but context is still present in an abstract format.
(k) Context decodability (red, $p_{Absent} = 0.38, p_{Present} = 0.0088, p_{RS} = 4.1x10^{-28}$), Shattering dimensionality (black, 0.53 vs. 0.53, $p_{RS} = 0.69$) during the stimulus presentation.
(l) Context CCGP (red, $p_{Absent} = 0.51, p_{Present} = 6.0x10^{-4}, p_{RS} = 2.5x10^{-34}$) during the stimulus presentation.
(m) Context Parallelism Score (red, $p_{Absent} = 0.77, p_{Present} = 2.3x10^{-6}$) during the stimulus presentation.
(n-s) Seizure onset zone exclusion analysis. Identical analysis to the main geometric analysis shown in Fig. 2.2, except that hippocampal neurons neurons recorded in seizure onset zones (SOZs, post-hoc identification) were removed. 410/494 neurons were retained for analysis. The exclusion neurons recorded from SOZ hippocampi led to the full hippocampal geometric analysis being effectively identical to that reported in Fig. 2.2, with every significant named dichotomy increase during stimulus **(n-p)** and baseline **(q-s)** periods being recapitulated in the absence of SOZ hippocampal neurons.
(n) Context decodability (red, $p_{Absent} = 0.12, p_{Present} = 0.00044, p_{RS} = 1.0x10^{-26}$). Stim pair decodability (purple, $p_{Absent} = 0.034, p_{Present} = 2.0x10^{-7}, p_{RS} = 3.5x10^{-32}$). Parity decodability (purple, $p_{Absent} = 0.74, p_{Present} = 0.019, p_{RS} = 2.4x10^{-30}$). Shattering dimensionality (black, 0.54 vs. 0.62, $p_{RS} = 4.6x10^{-6}$) during the stimulus period.
(o) Context CCGP (red, $p_{Absent} = 0.74, p_{Present} = 0.019, p_{RS} = 1.1x10^{-31}$). Stim pair CCGP (purple, $p_{Absent} = 0.084, p_{Present} = 2.7x10^{-5}, p_{RS} = 1.2x10^{-33}$) during the stimulus period.
(p) Context Parallelism Score (red, $p_{Absent} = 3.5x10^{-7}, p_{Present} = 0$). Stim pair Parallelism Score (purple, $p_{Absent} = 0.027, p_{Present} = 1.6x10^{-7}$) during the stimulus period.

(q) Context decodability (red, $p_{Absent} = 0.35, p_{Present} = 0.0025, p_{RS} = 1.1x10^{-30}$). Shattering dimensionality (black, 0.50 vs. 0.53, $p_{RS} = 0.0013$) during the baseline.
(r) Context CCGP (red, $p_{Absent} = 0.20, p_{Present} = 0.00018, p_{RS} = 2.5x10^{-34}$).
(s) Context Parallelism Score (red, $p_{Absent} = 0.0022, p_{Present} = 2.0x10^{-5}$).
(t-z) Non-inference performance control analysis. Identical analysis to the main geometric analysis shown in Fig. 2.2, except that inference absent and inference present sessions were distribution-matched for non-inference trial performance. Pairs of inference absent and inference present sessions with at most 7.5% difference in non-inference trial performance were selected, prioritizing sessions with more hippocampal neurons. This matching process yielded 10 inference absent sessions (152 neurons) and 10 inference present sessions (187 neurons) whose average non-inference performances did not statistically significantly differ (92.8% v.s. 94.7%, $p_{RS} = 0.58$, RankSum over sessions). All main geometric findings were recapitulated for the stimulus **(t-v)** and baseline **(w-y)** periods.
(t) Context decodability (red, $p_{Absent} = 0.12, p_{Present} = 0.00051, p_{RS} = 7.6x10^{-7}$). Stim pair decodability (purple, $p_{Absent} = 0.014, p_{Present} = 1.2x10^{-5}, p_{RS} = 3.3x10^{-7}$). Parity decodability (purple, $p_{Absent} = 0.27, p_{Present} = 0.057, p_{RS} = 5.6x10^{-5}$). Shattering dimensionality (black, 0.57 vs. 0.58, $p_{RS} = 0.26$) during the stimulus period.
(u) Context CCGP (red, $p_{Absent} = 0.52, p_{Present} = 0.044, p_{RS} = 6.7x10^{-8}$). Stim pair CCGP (purple, $p_{Absent} = 0.17, p_{Present} = 0.0021, p_{RS} = 6.7x10^{-8}$) during the stimulus period.
(v) Context Parallelism Score (red, $p_{Absent} = 0.54, p_{Present} = 0$). Stim pair Parallelism Score (purple, $p_{Absent} = 0.15, p_{Present} = 2.7x10^{-15}$) during the stimulus period.
(w) Context decodability (red, $p_{Absent} = 0.32, p_{Present} = 0.0036, p_{RS} = 4.4x10^{-7}$). Shattering dimensionality (black, 0.51 vs. 0.52, $p_{RS} = 0.64$) during the baseline.
(x) Context CCGP (red, $p_{Absent} = 0.27, p_{Present} = 0.0013, p_{RS} = 6.7x10^{-8}$) during the baseline.
(y) Context Parallelism Score (red, $p_{Absent} = 0.015, p_{Present} = 0$).
(z) Distribution-matched behavior shown using conventions from Fig. 2.1, 2.E1.

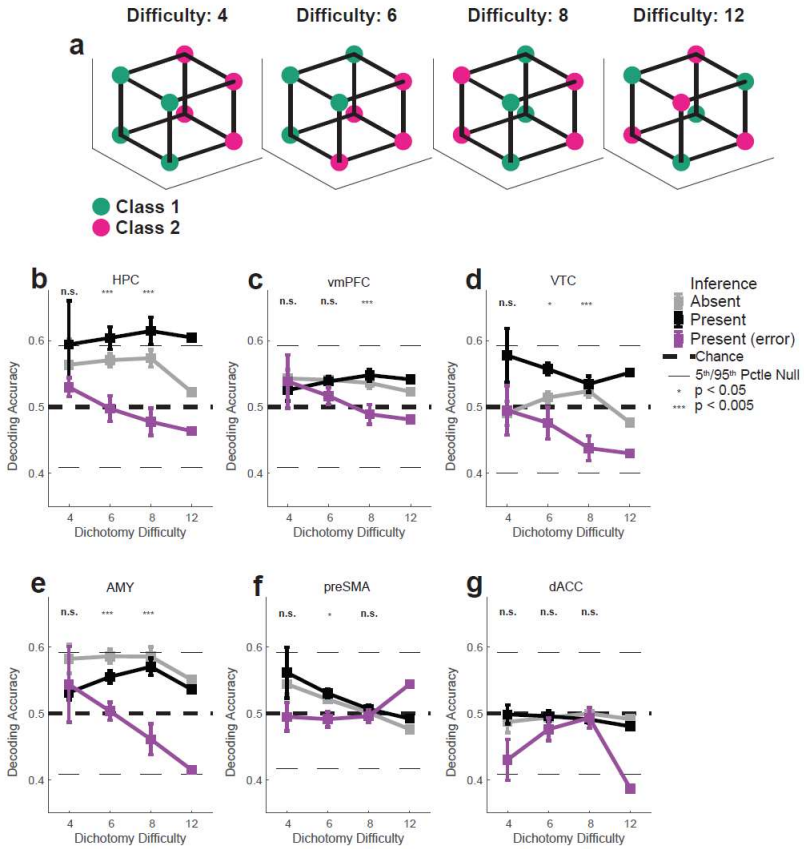

Figure 2.E5. Effect of inference and errors on shattering dimensionality as a function of dichotomy difficulty.

The signature for a high-dimensional representation is a greater degree of non-linear mixing of task variables. "Dichotomy difficulty" is a systematic measure that quantifies the relative amount of non-linear interaction of task variables needed in a population of neurons to make a given dichotomy decodable (see methods for detailed description). **(a)** Example schematics of dichotomies of increasing difficulty. The cubes here represent different unique task conditions realized by three binary variables, and node coloring represents membership of a condition to one of two arbitrary classes assigned for the purposes of dichotomy decoding (identical to Fig. 2.E2). Note: the difficulty 4 dichotomy corresponds to context and difficulty 12 dichotomy corresponds to parity (Fig. 2.E2). **(b-g)** Decoding accuracy as a function of dichotomy difficulty for different regions. Reported values (mean +/- SEM) are computed over dichotomy decoding accuracies, where the average decoding accuracy for each dichotomy is computed with 1000 repetitions of re-sampled estimation (see methods). The blue, red, and green curves correspond to correct inference absent trials, correct inference present trials, and inference present error trials respectively. Black dashed lines indicate chance level (50% for binary decoding), horizontal black lines indicate the 5th and 95th pctle of the null distribution. P-values are computed by conducting a one-way ANOVA over dichotomies independently for every dichotomy difficulty (Bonferroni MCC). This value is not meaningfully computable for difficulty 12, which contains a single dichotomy (the parity dichotomy), and is therefore not reported. Hippocampus alone **(b)** exhibits an increase in decoding accuracy from inference absent to inference present sessions, with more difficulty dichotomies rising above 95th pctle null in inference present sessions. A collapse in representational dimensionality on error trials

(purple curves) is present in the hippocampus **(b)**, and is also present in other areas, most prominently in the ventral temporal cortex **(d)** and the amygdala **(e)**.

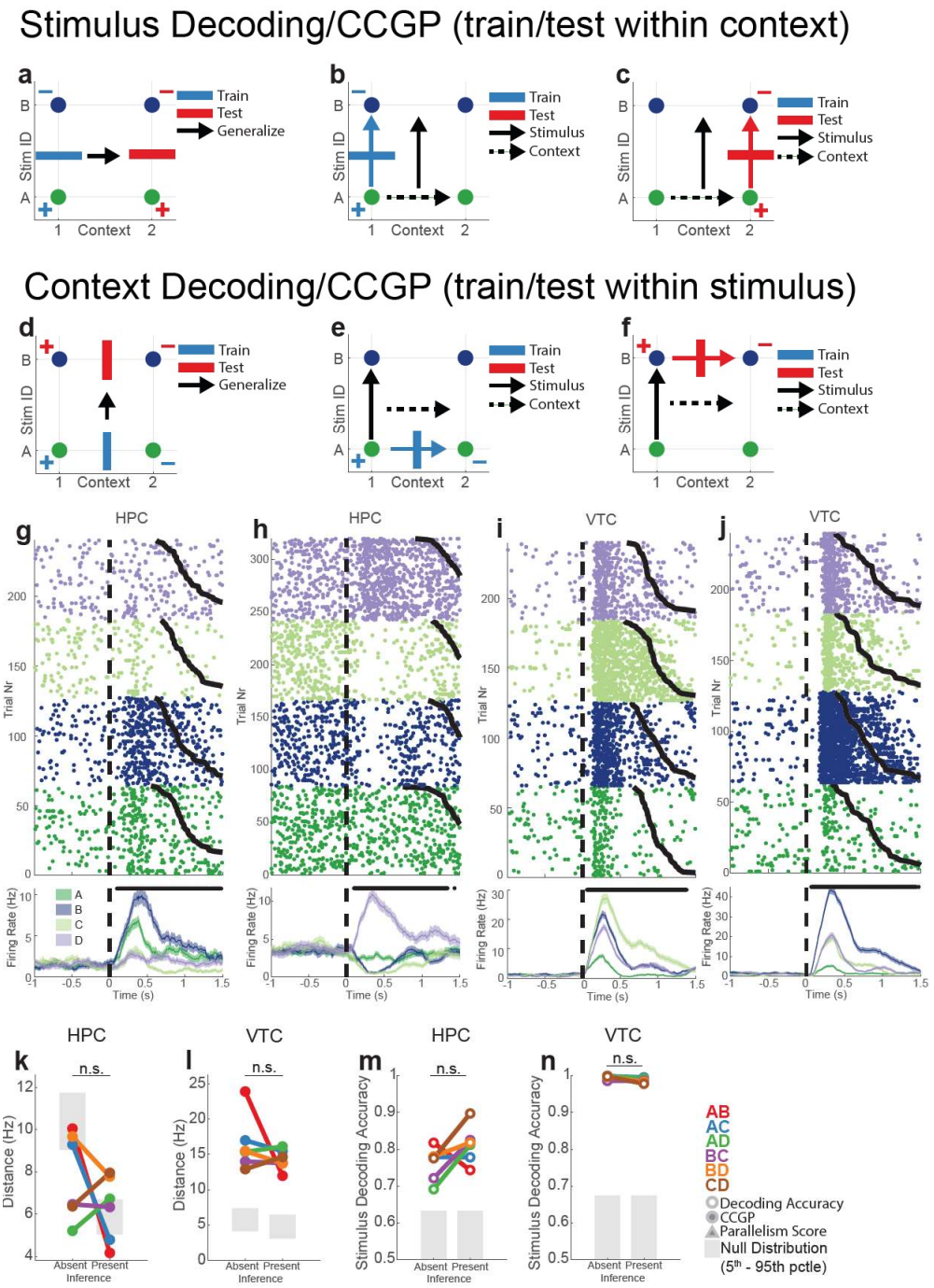

Figure 2.E6. Cross-condition generalization performance for stimulus identity and context defined over stimulus pairs.

To fully disentangle and study the interaction between stimulus coding and context, geometric analysis of balanced dichotomies must be replaced by new analyses that are defined for pairs of individual stimuli, thus allowing for the study of stimulus coding un-ambiguously without arbitrarily grouping together stimuli as is necessary in the balanced dichotomy approach. When considering a pair of stimuli (e.g. A and B) across two contexts (e.g. 1 and 2), there are four possible task conditions (A1, B1, A2, B2). On these points, stimulus (A1A2 vs B1B2) and context (A1B1 vs A2B2) can be decoded in a straightforward manner, but is not informative about the format in which

stimulus and context are encoded. The CCGP for stimulus across contexts **(a-c)** and for context across stimuli **(d-f)** provide information about the structure of the two variables and how they interact.

Consider within-context training/testing. The procedure is summarized in **(a)**, which shows a linear decoder (blue bar) trained between stimuli A and B in context 1 (blue + and – correspond to class labels for training). The decoder is then generalized to context 2, where stimulus identity is decoded (red bar, + and – for class labels). This procedure is broken down step-by-step for training in **(b)** and testing in **(c)**. In addition, arrows showing persistent stimulus and context coding vectors (black/dashed arrows) have been drawn alongside the vector orthogonal to the hyperplane learned during train/test (colored arrow passing through the bar). Note that, for this formulation of Stimulus CCGP, the stimulus coding vector and the normal vector to the hyperplane are parallel in **(b)** and **(c)**. Thus, in cases with high within-context train/test Stimulus CCGP, stimulus information is present in an abstract format across contexts.

The same procedures for computing CCGP can be applied for studying the format of context organizing across pairs of stimuli **(d-f)**, with schematic details identical to those described above for Stimulus CCGP. Here, high Context CCGP indicates that context is encoded abstractly across the different stimuli.

(g-j) Rasters and PSTHs of example neurons from hippocampus **(g,h)** and ventral temporal cortex **(i,j)** showing tuning for stimulus identity. Plotting conventions identical to those used in Fig. 2.E1j.

(k) Average distances between stimulus representations in hippocampus (HPC) in inference absent and inference present sessions. All plotted points correspond to named, interpretable groups of conditions defined by pairs of stimuli presented in both contexts. For example, the green dot in "inference absent" indicates the average distance (~5.1Hz) between the condition centroids for stimulus A and stimulus D (averaged over contexts). Distance is computed as Euclidean distance between the stimulus centroids, each of which is an N (# of neurons) dimensional vector of average firing rates during stimulus presentation. Neuron counts are balanced between inference absent and inference present sessions to allow for direct distance comparisons. Null distributions here are geometric nulls, and are identical to those used for CCGP and Parallelism Score. Significance of the difference between inference absent and inference present session inter-stimulus distances is established by RankSum test computed over stimulus pairs, and n.s. indicates $p > 0.05$.

(l) Same as **(k)**, but for VTC.

(m) Decodability of stimuli also did not significantly change between inference absent and inference present for HPC. Here, decoding accuracies are reported for each unique pair of stimuli with 1000 repetitions of trial sub-sampling. Null distributions are constructed with trial-label shuffling, and the gray bars correspond to the boundary of the 5^{th} to 95^{th} pctle of the null. Significance of the difference between inference absent and inference present decodability is also established by Ranksum test over average decoding accuracies and n.s. indicates $p > 0.05$.

(n) Same as **(n)**, but for VTC.

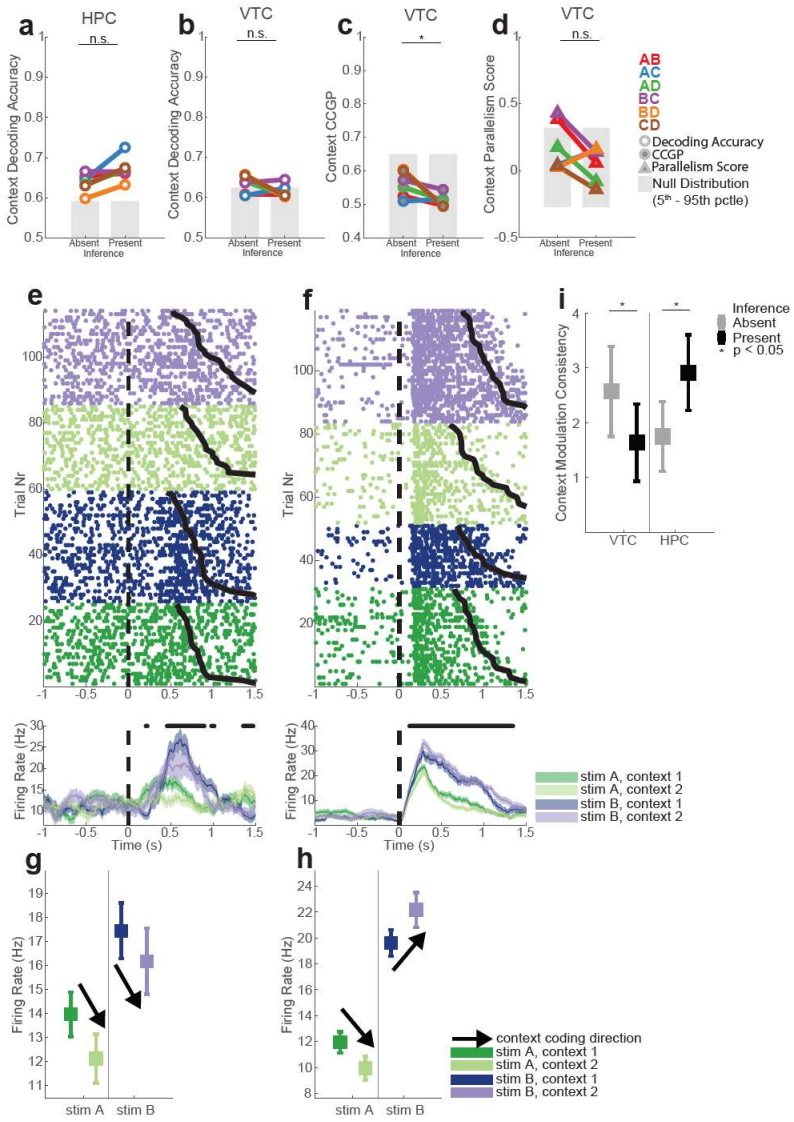

Figure 2.E7. Additional context CCGP analysis over stimulus pairs for hippocampus and ventral temporal cortex (stimulus period).

Change in context decoding accuracy from inference absent to inference present sessions evaluated over individual stimulus pairs is shown for the hippocampus **(a)** and ventral temporal cortex **(b)**. Individual points correspond to context decoding accuracy averaged over 1000 repetitions of decoding/CCGP/Parallelism Score estimation with trial re-sampling.

(c) and **(d)** show Context CCGP and Context Parallelism Score over stimuli for VTC. Analogous plots for HPC are Fig. 2.3g,h.

Two exemplar neurons are shown, one from HPC **(e,g)** and one from VTC **(f,h)** that feature both stimulus tuning and context modulation. Plotting conventions are identical to previous raster/psth plots apart from the colors and conditions plotted, which here are two stimuli (A and B) in the two contexts. Responses to task conditions for the two neurons are summarized in **(g,h)**, which show mean ± s.e.m. firing rates by condition for spikes counted on individual trials during the stimulus period (0.2s to 1.2s after stimulus onset). The same trials used to compute **(g)** and **(h)** are shown in **(e)** and **(f)** respectively, and condition colors are matched between the two sets of plots. Black arrows

indicate the direction in which the firing rate for a stimulus is modulated by a shift in context. The HPC neuron **(g)** shows consistent modulation by context since both arrows point downward, whereas the VTC neuron **(h)** shows inconsistent modulation by context since one arrow points downwards and the other points upward.

(i) Change in the consistency of context-modulation for stimuli averaged over all neurons in VTC and HPC. Context modulation consistency is the tendency for a neuron's firing rate to shift consistently (increase or decrease) to encode context across stimuli. This consistency can take on values between 0 (increase in firing rate to encode context for half of the stimuli, decrease in firing rate for the other half) and 4 (either increase or decrease in firing rate for all four stimuli). An interaction effect is observed between context modulation consistency for HPC neurons and VTC neurons in inference absent and inference present sessions in the absence of main effects (2 x 2 ANOVA, $p_{Area} = 0.36$, $p_{Inference} = 0.64$, $p_x = 4.5x10^{-5}$), revealing significant increases in context modulation consistency in HPC from inference absent to present with concurrent decreases in VTC.

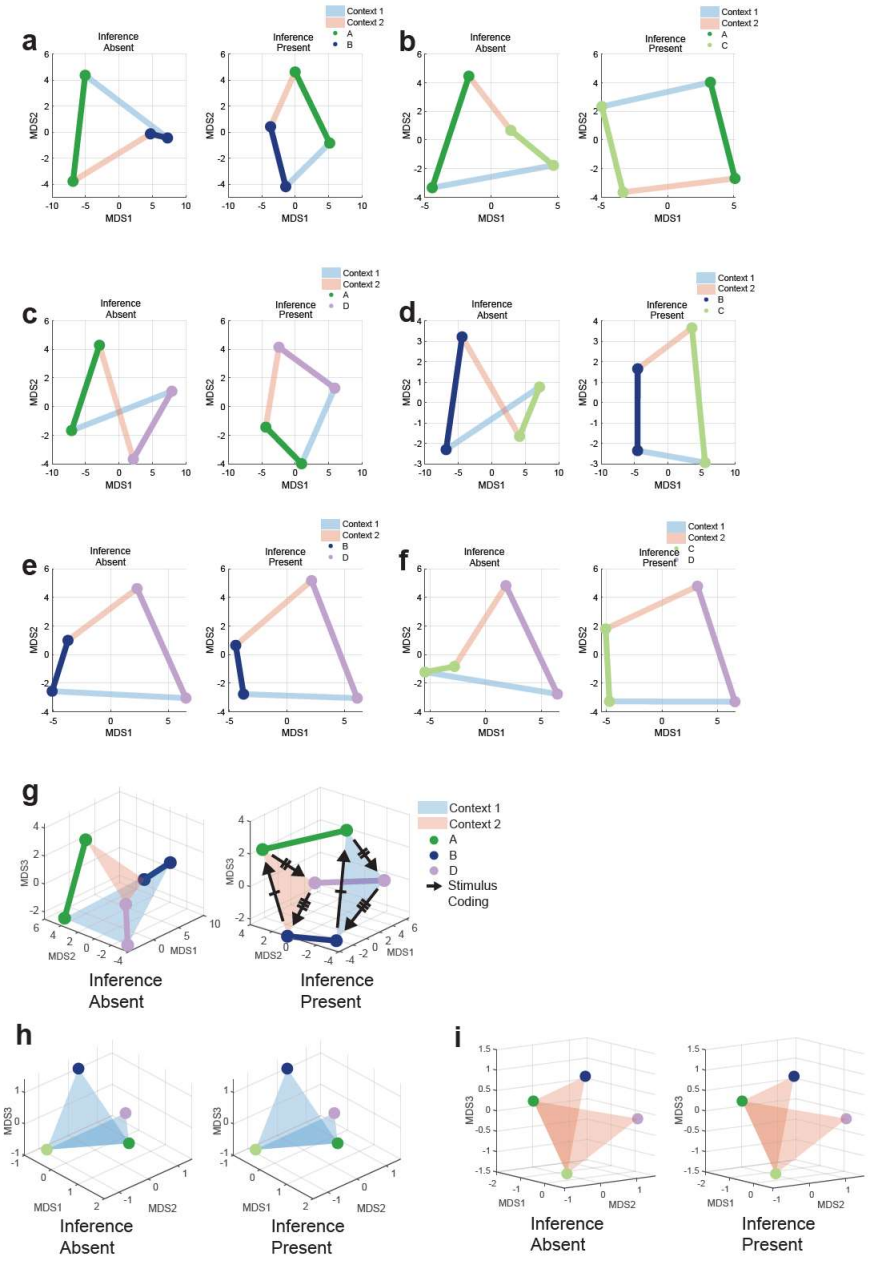

Figure 2.E8. Hippocampal MDS plots summarizing changes in stimulus and context geometry.

MDS plots analogous to that shown in Fig. 2.3i, but plotted in 2D for individual stimulus pairs. Colored points represent the mean condition response of all HPC neurons during inference absent or inference present sessions to a given stimulus in a given context. Stimuli are color coded according to identity (e.g. in A, green points are condition responses to stimulus A and orange points are condition responses to stimulus B), and are connected by a line of the same color to reflect the context coding direction for that stimulus. Stimuli in the same context are connected by shaded lines that are blue for context 1 and red for context 2. Since MDS here was conducted independently for inference absent and inference present, individual MDS axes are not directly comparable, but the relative distances are comparable since the number of neurons is matched between inference absent and present, and both are reduced to the same number of MDS dimensions (N_{dim} = 2). Condition averages are computed using only correct trials. Evidence of disentangling of context and stimulus

identity is present across most stimulus pairs, with the notable exception of the B/D stimulus pair **(e)**, which is perfectly correlated with outcome and therefore cannot be dissociated from outcome using CCGP. The emergence of quadrilaterals with approximately parallel sides for all other stimulus pairs **(a-d, f)** is a signature of disentangling of stimulus identity and context.

(g) Changes in neural geometry in HPC. MDS of condition-averaged responses of all recorded HPC neurons shown for inference absent (left) and inference present (right) sessions. All plotting conventions are identical to those in **(a-f)**, except MDS was applied with $N_{dim} = 3$, and three stimuli (A,B,D) are plotted simultaneously. Black arrows on the inference present plot highlight parallel coding of stimuli across the two context planes.

(h,i) MDS plots of HPC condition-averaged responses shown for context 1**(h)** and context 2 **(i)** separately. Axes are directly comparable here between inference absent and present due to alignment via CCA prior to plotting. Note that the stimulus geometry in each context is a tetrahedral (maximal dimensionality, unstructured) regardless of the presence or absence of inference behavior.

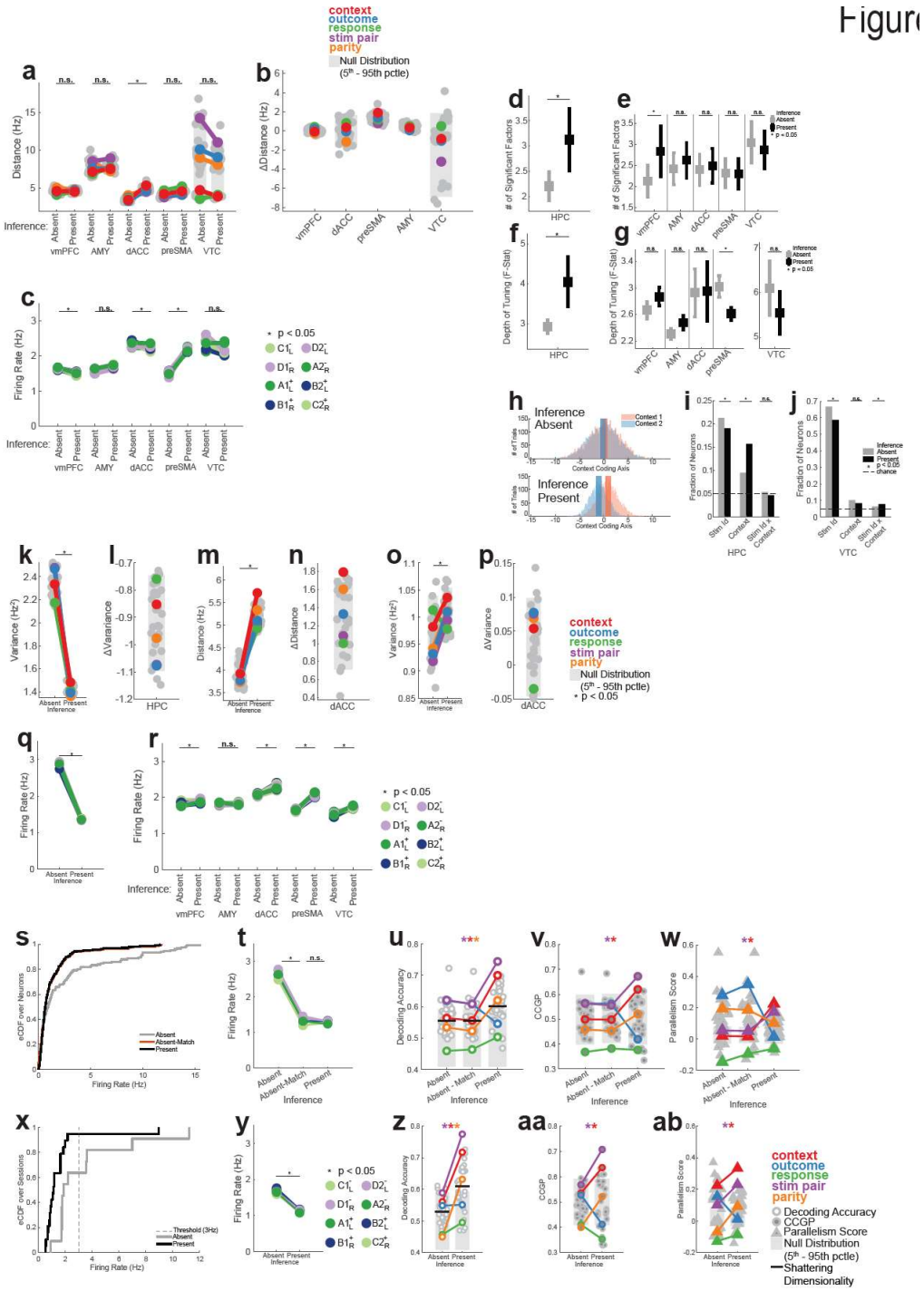

Figure 2.E9. Implementation of geometric changes in hippocampal representation.
(a-j) Stimulus period analysis.
(a) Distances between centroids for balanced dichotomies shown for all regions other than HPC. Plotting conventions are identical to those used in Fig. 2.4g. Note: neuron counts were only balanced across inference absent/present within-region, so distances in different regions are computed in spaces with different dimensionality and are therefore not meaningfully comparable. Significant change in average dichotomy separation determined through Bonferroni MCC RankSum where * indicates $p < 0.05$, and n.s. otherwise.

(b) Changes in inter-centroid distance for balanced dichotomies. Points in these plots are the differences between inference present and absent distances shown in **(a)**. No distances for named dichotomies increased or decreased more than would be expected by chance (outside 5th-95th pctle null).
(c) Firing rates for individual task conditions (8 total) for all regions other than HPC. Plotting conventions identical to those used in Fig. 2.4e. Task conditions are color coded based on the identity of the presented stimulus (same as Fig. 2.1b. 2.4e,f). Significant change in average dichotomy separation determined through Bonferroni MCC RankSum where * indicates $p < 0.05$, and n.s. otherwise.
Changes in hippocampal single-neuron tuning quantified by 3-way ANOVA (Response, Context, Outcome) with interactions. Significant factors ($p < 0.05$) were identified for every neuron and averages of both the number of factors per neuron **(d)** and the depth of tuning of those factors quantified through -ANOVA F-Statistic) **(f)** reported (mean ± s.e.m. across neurons) for the inference absent (red) and inference present (blue) sessions. Significance of difference between inference absent and present sessions for both the number of factors ($\mathbf{d}, p_{RS} = 0.041$) and the tuning strength ($\mathbf{f}, p_{RS} = 0.027$) was assessed by RankSum test over neurons between the two groups, and "*" indicates $p_{RS} < 0.05$.
(e) same as **(d),** but for all regions other than HPC.
(g) same as **(f),** but for all regions other than HPC.
(h) The change in the distribution of trials projected along the coding direction for context was visualized during inference absent (above) and inference present (below) sessions. The red and blue histograms are the distribution of projected trials from context 1 and 2 respectively, with the red and blue vertical lines indicating the mean of each distribution. Positive and negative values for projection were arbitrarily established by computing the coding vector as (context 1 – context 2).
(i) Plot showing the fraction of hippocampal neurons that exhibit task selectivity for inference absent (red) and inference present (blue) sessions. Selectivity is determined independently for every neuron using a 4x2 ANOVA (Stimulus Identity, Context), with a per-factor significance threshold of $p < 0.05$. Significant differences in tuned fractions between inference absent and inference present assessed with z-test. **(j)** Plot showing the fraction of hippocampal neurons that exhibit task selectivity for inference absent (red) and present (blue) sessions. Selectivity is determined independently for every neuron using a 4x2 ANOVA (Stimulus Identity, Context), with a per-factor significance threshold of $p < 0.05$. Significant differences in tuned fractions between inference absent and present sessions assessed with z-test.
(j) same as **(i)**, but for VTC.
(k-r) Stimulus period analysis.
(k) Average variance of individual trials projected onto the coding direction for every dichotomy. Plotting conventions identical to those in Fig. 2.4i. Average variance along coding directions decreased significantly between inference absent and inference present sessions ($p_{RS} = 6.5x10^{-13}$, RankSum over dichotomies).
(l) Changes in variance between inference present and inference absent for all dichotomies shown in **(k)**. No named dichotomies fall outside the 5th-95th pctle of the null distribution.
(m) Population distance between dichotomy centroids for dACC at baseline. All plotting conventions identical to those used in Fig. 2.4g. Average distance between dichotomy centroids increased when comparing inference absent to inference present sessions ($p_{RS} = 2.9x10^{-8}$, RankSum over dichotomies). Notably, context centroids emerged as significantly separated in inference present sessions than expected by chance ($p_{Absent} = 0.48, p_{Present} = 0.0065$)

(n) Changes in distance between inference present and inference absent sessions for all dichotomies shown in **(m)**. Context alone (red, $p_{\Delta} = 0.047$) exhibited a greater increase in distance than expected by chance.
(o) Same as **(k)**, but for projective variance in dACC during the baseline. Average variance along coding directions increased significantly between inference absent and present sessions ($p_{RS} = 6.0x10^{-3}$, RankSum over dichotomies).
(p) Same as **(l)**, but for differences in dACC variance during the baseline computed using **(o)**.
(q) Baseline firing rate averaged by condition (8 total) for the hippocampus. Plotting conventions are identical to those in Fig. 2.4e. Reduction from inference absent to inference present sessions is significant ($p < 0.05$, RankSum over conditions).
(r) Baseline firing rates averaged by condition for all regions other than hippocampus. Significance of change in firing rate also assessed by RankSum over conditions ("*" indicates $p < 0.05$, n.s. otherwise). Note: most regions (apart from AMY) exhibit slight, but significant increases in baseline firing rate during in inference present compared with inference absent.
(s-w) Stimulus period firing rate distribution-matched control analysis.
(s) eCDF of mean stimulus firing rate over all hippocampal neurons in the inference absent (gray) and inference present (black) sessions, as well as randomly thinned inference absent firing rates that distribution-match the inference present firing rates (orange).
(t) Mean stimulus firing rates over neurons reported by condition for inference absent, inference present, and distribution-matched inference absent firing rates. * indicates $p < 0.05$ for RankSum over conditions, and n.s. otherwise.
(u-w) Neural geometry measures compared for inference absent, inference present, and distribution-matched inference absent sessions. All plotting conventions for decoding accuracy **(u)**, CCGP **(v)**, and Parallelism Score **(w)** are identical to those used in Fig. 2.E2,3. No meaningful differences are present between inference absent and distribution-matched inference absent for any dichotomy/metric.
(x-ab) Stimulus period firing rate control analysis excluding high-hippocampal-firing-rate sessions.
(x) eCDF of mean hippocampal firing rate over inference absent (gray) and inference present (black) sessions. Each point in the distribution corresponds to the mean hippocampal firing rate over all neurons in a single session. Vertical dashed line indicates 3Hz threshold. Hippocampal neurons from all inference absent and inference present sessions above this threshold were excluded from analysis shown in **(y-ab)**. 131/169 inference absent neurons (10/14 sessions) and 318/325 inference present neurons (21/22 sessions) are retained.
(y) Same as **(t)**, but computed using all sessions with mean hippocampal firing rate < 3Hz.
(z-ab) Neural geometry measures re-computed excluding hippocampal neurons from high-firing-rate sessions. No meaningful differences apart from above-chance context Parallelism Score in inference absent sessions (Fig. 2.E9ab, red, $p_{Absent} = 2.2x10^{-8}$).

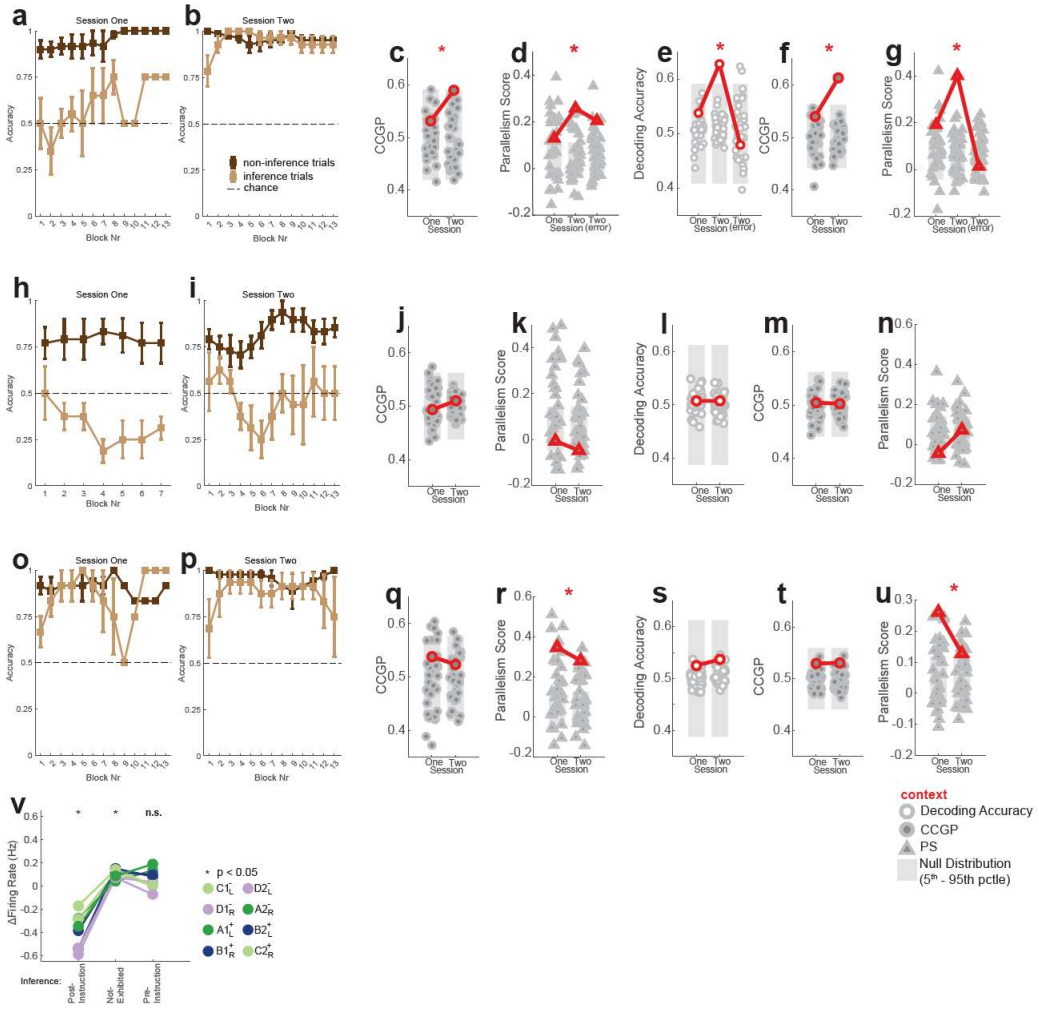

Figure 2.E10. Additional analysis of the effect of instructions on hippocampal neural geometry. **(a-g)** Additional behavior and neural analysis for the post-instruction inference session group. Behavioral plots **(a,b)** are similar to those shown in Fig. 2.E1e,f, except Session One **(a)** and Session Two **(b)** plots show performance over time for sessions recorded immediately preceding and immediately following verbal instructions describing latent task structure. Performance is shown for non-inference trials (black) and inference trials (gray). Average performance is computed as a moving average with a 3-block window on the last three trials before a context switch (non-inference) and on the first inference trial after a switch (inference). Error bars are standard errors computed over subjects. Chance performance is indicated with the dashed line at y = 0.5.

Geometric measures shown are computed over balanced dichotomies, and are plotted using the same conventions as discussed previously (see Fig. 2.2, methods for details), except for the left and right columns of each plot correspond to Session One and Session Two hippocampal neural geometry respectively. Only context is plotted as a named dichotomy for visual clarity.

(c) CCGP (context, red, $p_{One} = 0.27, p_{Two} = 0.046, p_{RS} = 1.4x10^{-31}$) and **(d)** Parallelism Score (context, red, $p_{One} = 0.029,\ p_{Two} = 3.5x10^{-6}, p_{Two\,(error)} = 0.0028$) for the post-instruction inference group during the stimulus period.

(e) Decoding accuracy (context, red, $p_{One} = 0.35, p_{Two} = 0.0014, p_{Two\,(error)} = 0.55, p_{RS} = 1.4x10^{-20}$), **(f)** CCGP (context, red, $p_{One} = 0.33, p_{Two} = 0.0037, p_{RS} = 3.0x10^{-3}$), and **(g)**

Parallelism Score (context, red, $p_{One} = 0.017, p_{Two} = 7.5x10^{-8}, p_{Two\ (error)} = 0.40$) for the post-instruction inference group during the baseline period.
(h-n) Same as **(a-g)**, but for the session group where patients never exhibited inference (inference not-exhibited).
(j) CCGP (context, red, $p_{One} = 0.56, p_{Two} = 0.39, p_{RS} = 0.004$) and **(k)** Parallelism Score (context, red, $p_{One} = 0.81, p_{Two} = 0.95$) for the inference not-exhibited group during the stimulus period.
(l) Decoding accuracy (context, red, $p_{One} = 0.45, p_{Two} = 0.45, p_{RS} = 0.68$), **(m)** CCGP (context, red, $p_{One} = 0.45, p_{Two} = 0.47, p_{RS} = 0.15$), and **(n)** Parallelism Score (context, red, $p_{One} = 0.93, p_{Two} = 0.30$) for the inference not-exhibited group during the baseline period.
(o-u) Same as **(a-g)**, but for the session group where patients exhibited inference from Session One, before they were explicitly instructed about the latent task structure (pre-instruction inference).
(q) CCGP (context, red, $p_{One} = 0.23, p_{Two} = 0.19, p_{RS} = 0.0045$) and **(r)** Parallelism Score (context, red, $p_{One} = 6.3x10^{-8}, p_{Two} = 4.5x10^{-7}$,) for the pre-instruction inference group during the stimulus period.
(s) Decoding accuracy (context, red, $p_{One} = 0.37, p_{Two} = 0.47, p_{RS} = 0.036$), **(t)** CCGP (context, red, $p_{One} = 0.30, p_{Two} = 0.50, p_{RS} = 5.9x10^{-7}$), and **(u)** Parallelism Score (context, red, $p_{One} = 1.7x10^{-5}, p_{Two} = 0.029$) for the pre-instruction inference group during the baseline period.
(v) Changes in hippocampal firing rates for the 3 different sub-groups of session pairs. Firing rate changes here are computed during the stimulus presentation period (0.2s to 1.2s after stim onset) from consecutive Session One and Session Twos. Points are average changes in condition-averaged firing rates (8 unique conditions). Changes in firing rate that significantly differed from zero (t-test, $p < 0.05/3$) are indicated with a "*". Post-instruction inference group alone exhibited significant decrease in firing rate. Inference not-exhibited group exhibited an increase in firing rate.

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

Table 2.S1. Tabulation of Patients, Sessions, Behavior, and Neurons.
Summary of patient information, the number of sessions performed, the behavioral classification at the patient and session level, and the number of recorded neurons per region per session. Patient behavior is defined with respect to instances of high-level verbal instructions (see Fig. 2.5), where: Pre – "pre-instruction inference achieved", NE – "Inference not exhibited", post – "post-instruction inference achieved", and N/A – "did not qualify for analysis". Session behavior is defined with respect to performance on the first available inference trial, where: IA – "inference absent", IP – "inference present", X – "at or below chance non-inference performance".

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

Table 2.S2. Definition of all balanced dichotomies.

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

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

Table 2.S3. Definition of all stimulus dichotomies.

Task condition assignment for stimulus dichotomies. These dichotomies are used in Fig. 2.3b,c,e,f and associated supplements whenever there is a reference to "Stimulus CCGP" or "Stimulus Parallelism Score". Partial and full correlations with other task variables are noted for each stimulus dichotomy.

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

Table 2.S4. Definition of all context dichotomies.
Task condition assignment for context dichotomies. These dichotomies are used in Fig. 2.3g,h and associated supplements whenever there is a reference to "Context CCGP" or "Context Parallelism Score". Partial and full correlations with other task variables are noted for each context dichotomy.

Supplementary Sections

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

We analyzed the baseline period preceding stimulus onset (Fig. 2.2, inset) to study the geometry of task representations that might persist from the previous trial (all labels were defined by the prior just completed trial for this analysis). Context was encoded as an abstract variable in the HPC in inference present and not in inference absent sessions. Unlike the stim epoch, context was the only named dichotomy to emerge as significantly decodable at baseline in the HPC for inference present sessions (Fig. 2.2h, red, inference absent vs. present, $p_{RS} = 1.1x10^{-33}, p_{Absent} = 0.35, p_{Present} = 6.4x10^{-5}$, E3l), indicating that any task condition information from the previous trial other than the context (i.e. outcome, response) was not significantly decodable from the population during this epoch. Nevertheless, the shattering dimensionality significantly increased in inference present compared to inference absent sessions (0.51 vs 0.53, $p_{RankSum} = 0.0079$), which is attributable to the rise in decodability of context and context correlated dichotomies. Context during baseline was also encoded abstractly as shown by the increase in CCGP (Fig. 2.2i, 2.E3m, red, Inference absent vs present, $p_{RS} = 2.4x10^{-34}$, $p_{Absent} = 0.45, p_{Present} = 7.7x10^{-6}$) and Parallelism Score (Fig. 2.E3n,o, red, $p_{Absent} = 0.37, p_{Present} = 1.2x10^{-10}$). These results indicate that an abstract representation of context that persisted from the previous trial also emerged in the hippocampus in sessions where patients performed inference.

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

To determine whether the geometric findings presented here arose as a consequence of strong classical (univariate linear) tuning, two ablation analyses were conducted in which all neurons that had significant univariate linear tuning to at least one task variable were excluded and all geometric measures were re-computed. In the first analysis, neurons tuned to one or several of response, reward, or context (3-Way ANOVA, any main effect $p < 0.01$) were excluded. Of the 494 neurons in HPC, 41 were excluded for exhibiting classic tuning prior to recomputing all measures. Qualitatively, findings remained largely unchanged by the exclusion of this population of specialized neurons (Fig. 2.E4a-e). The Parity dichotomy was no longer decodable above chance in inference present sessions (Fig. 2.E4a, orange, $p_{Present} = 0.056$), but the rise in Shattering Dimensionality (average over dichotomies) was still present (Fig. 2.E4a, $p_{RS} = 0.0012$, Ranksum over dichotomies). A similar analysis was also conducted for the baseline representation of context, and again the findings qualitatively remained unchanged (Fig. 2.E4f-j). Note that, although the decodability of context at baseline during inference present sessions did decrease in significance when conducting the error trial analysis (Fig. 2.E4i, red, $p_{Present} = 0.12$), it was still significantly increased compared to inference absent sessions ($p_{RS} = 5.7x10^{-1}$) and was significantly reduced in error trials ($p_{Present} = 0.37$, inference present correct vs error, $p_{RS} = 0.0002$). In the second analysis, neurons tuned to stimulus identity or context (2-Way ANOVA, any main effect $p < 0.01$) were excluded. In this analysis, 82 neurons were excluded, leading to a loss of decodability of stimulus dichotomies due to the removal of visually-selective neurons as expected (Fig. 2.E4k-m). Context, however, remained decodable and present in an abstract format. Together, these analyses indicate that the abstract context representation in the hippocampus is a highly distributed variable whose geometry is not a simple consequence of strong univariate tuning of a small population of "context" neurons.

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

To ensure that the reported findings were not influenced by pathology, in a further control analysis we excluded all HPC neurons (86/494) recorded from electrodes that were clinically identified to reside in medial temporal seizure onset zones (SOZs). Repeating our analyses on the neurons that remain revealed that results were qualitatively unchanged (Fig. 2.E4n-s), though the parallelism for context was now significantly above chance during inference absent sessions for both the stimulus (Fig. 2.E4p, red, $p_{Absent} = 3.5x10^{-7}$) and baseline (Fig. 2.E4s, red, $p_{Absent} = 0.0022$) periods. These findings suggest that neurons residing in hippocampal tissue with high disease burden do not meaningfully contribute to the task representation[17], thereby leading to marginally increased representation strength once these neurons are removed (note that the decodability and CCGP for context in inference absent sessions are still not significantly different from chance, confirming our finding).

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

To ensure that the reported geometric findings in the hippocampus did not arise due to non-inference performance differences between inference absent and inference present sessions, an additional control analysis was performed where the non-inference trial performance of included sessions was distribution-matched between sessions where inference was absent and present. Pairs of inference absent and inference present sessions with at most 7.5% difference in non-inference trial performance were selected, prioritizing sessions with more hippocampal neurons. This matching process yielded 10 inference absent sessions (152 HPC neurons) and 10 inference present sessions (187 HPC neurons) whose average non-inference performances did not statistically significantly differ (92.8% v.s. 94.7%, $p_{RS} = 0.58$, RankSum over sessions, Fig. 2.E4z). All main geometric findings were recapitulated using this session split during both the stimulus (Fig. 2.E4t-v) and baseline (Fig. 2.E4w-y) periods, indicating that the learning-dependent changes in the hippocampal representational geometry we observe cannot be explained by differences in non-inference trial performance.

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

To determine if changes in univariate tuning could partially explain the observed changes in the hippocampal representation from inference absent to inference present sessions, average tuning properties of single neurons were computed. Both the number of significant factors (main effects or interactions) per neuron (3-way ANOVA – Response, Context, Outcome, $p < 0.05$) and the tuning strength of neurons for those significant factors indexed by the F-statistic of those ANOVA factors were significantly elevated in inference present compared to inference absent (Fig. 2.E9d, $p_{RS} = 0.010$, 2.E9f, $p_{RS} = 0.0089$). These effects were only observed in HPC, with either no change or a decrease in tuning observed in the other areas except for a significant increase in vmPFC (Fig. 2.E9e). We also separately considered the linear and non-linear (interaction) terms for the 4 (Stimulus) x 2 (Context) ANOVA and found that, while the fraction of hippocampal neurons exhibiting significant ($p < 0.05$, Main Effect) stimulus identity tuning significantly decreased from absent to present inference sessions (Fig. 2.E9i, 21.3% vs. 19.1% of neurons, $p = 0.002$), the fraction of neurons exhibiting significant context tuning increased (Fig. 2.E9i, 9.5% vs. 15.7% of neurons, $p = 2.2x10^{-6}$), which could also partially explain the increase in centroid distance and resultant context decodability. Such "context neurons", however, cannot account for this effect fully as their removal from the pseudopopulation does not qualitatively alter the hippocampal population geometry (Fig. 2.E4). As a control, this analysis was also performed for VTC neurons, in which we did not find a significant difference in the percentage of neurons with univariate context tuning (Fig. 2.E9j).

2.S.6 Context representations outside of the hippocampus

The dorsal Anterior Cingulate Cortex (dACC) was the only other region to exhibit significant changes in its latent context representation as a function of patients' ability to perform inference. However, this context representation was limited to the baseline period, since geometric analysis for the stimulus period revealed an absence of significant context decodability and CCGP in inference absent and inference present sessions (Fig. 2.2g, 2.E3a). However, the Parallelism Score for context in the dACC did increase significantly (Fig. 2.E3b, red, inference absent vs. inference present, $p_{RS} = 2.4x10^{-19}, p_{Absent} = 0.99, p_{Present} = 3.8x10^{-1}$), reflecting an increase in context parallelism in condition averages that was not detectable with the other metrics (which are based on single trial decoding).

During the baseline, however, context also emerged as the only decodable dichotomy in inference present sessions in the dACC (Fig. 2.E3k, red, $p_{Absent} = 0.37, p_{Present} = 0.049$). We found this context variable was also represented in an abstract format, emerging as the dichotomy with the highest CCGP (Fig. 2.E3p,s, red, $p_{Absent} = 0.26, p_{Present} = 0.018$) and Parallelism Score (Fig. 2.E3q,t, red, $p_{Absent} = 0.18, p_{Present} = 0.013$) in inference present sessions, while being at chance for both metrics in inference absent sessions. As in HPC, the lack of decodability for previous trial outcome and response suggests that any variables encoded at earlier epochs during the previous trial (e.g. post-reply delay or outcome) were extinguished from the population by the onset time of the current baseline epoch.

Analysis of dichotomy centroid distances during the baseline period revealed context also emerged as the dichotomy with the greatest separation in the dACC during inference present session trials (3.9 vs. 5.7 Hz, $p_{Absent} = 0.48, p_{Present} = 0.0065, p_{\Delta Dist} = 0.046$, Fig. 2.E9m,n). However, this implementation of the context representation was achieved through significant increases in condition-wise firing rates (Fig. 2.E9r, dACC inference absent vs inference present, $p_{RS} = 0.049$) as opposed to the firing rate decrease observed in the hippocampus. Together, these analyses indicate that a weak, but nonetheless significant, representation of latent context encoded in an abstract format also emerged in the dACC, and that this variable was accommodated in the representation through a different implementational strategy.

Bibliography

1.Tolman, E. C. Cognitive maps in rats and men. *Psychological Review* **55**, 189–208 (1948).

2.Chung, S. & Abbott, L. F. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology* **70**, 137–144 (2021).

3.Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W. & Behrens, T. E. J. How to build a cognitive map. *Nat Neurosci* **25**, 1257–1272 (2022).

4.Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* **331**, 1279–1285 (2011).

5.Kemp, C. & Tenenbaum, J. B. Structured statistical models of inductive reasoning. *Psychological Review* **116**, 20–58 (2009).

6.McClelland, J. L. *et al.* Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* **14**, 348–356 (2010).

7.Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences* **14**, 357–364 (2010).

8.Ho, M. K., Abel, D., Griffiths, T. L. & Littman, M. L. The value of abstraction. *Current Opinion in Behavioral Sciences* **29**, 111–116 (2019).

9.Konidaris, G. On the necessity of abstraction. *Current Opinion in Behavioral Sciences* **29**, 1–7 (2019).

10.Vaidya, A. R., Jones, H. M., Castillo, J. & Badre, D. Neural representation of abstract task structure during generalization. *eLife* **10**, e63226 (2021).

11.Schapiro, A. C., Turk-Browne, N. B., Norman, K. A. & Botvinick, M. M. Statistical learning of temporal community structure in the hippocampus. *Hippocampus* **26**, 3–8 (2016).

12.Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).

13.Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013-1028.e14 (2017).

14.She, L., Benna, M. K., Shi, Y., Fusi, S. & Tsao, D. Y. The neural code for face memory. 2021.03.12.435023 Preprint at https://doi.org/10.1101/2021.03.12.435023 (2021).

15.Nogueira, R., Rodgers, C. C., Bruno, R. M. & Fusi, S. The geometry of cortical representations of touch in rodents. *Nat Neurosci* **26**, 239–250 (2023).

16.Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. Tuned geometries of hippocampal representations meet the demands of social memory. 2022.01.24.477361 Preprint at https://doi.org/10.1101/2022.01.24.477361 (2023).

17.Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology* **20**, 251–256 (2010).

18.Robert, S., Arno, V. & J.d, M. M. Distinct hippocampal and cortical contributions in the representation of hierarchies. *eLife* **12**, (2023).

19.Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).

20.Knudsen, E. B. & Wallis, J. D. Hippocampal neurons construct a map of an abstract value space. *Cell* **184**, 4640-4650.e10 (2021).

21.Aronov, D., Nevers, R. & Tank, D. W. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543**, 719–722 (2017).

22.Nieh, E. H. *et al.* Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).

23.Courellis, H. S. *et al.* Spatial encoding in primate hippocampus during free navigation. *PLOS Biology* **17**, e3000546 (2019).

24.Moore, J. J., Cushman, J. D., Acharya, L., Popeney, B. & Mehta, M. R. Linking hippocampal multiplexed tuning, Hebbian plasticity and navigation. *Nature* **599**, 442–448 (2021).

25.Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by human medial-frontal cortex. *Science* **368**, eaba3313 (2020).

26.Fu, Z. *et al.* The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* **376**, eabm9922 (2022).

27.Higgins, I. *et al.* Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun* **12**, 6456 (2021).

28.Higgins, I. *et al.* Towards a Definition of Disentangled Representations. Preprint at https://doi.org/10.48550/arXiv.1812.02230 (2018).

29.Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

30.Fried, I., Rutishauser, U., Cerf, M. & Kreiman, G. *Single Neuron Studies of the Human Brain: Probing Cognition*. (MIT Press, 2014).

31.Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* **3**, 946–953 (2000).

32.Gross, C. G. Single neuron studies of inferior temporal cortex. *Neuropsychologia* **46**, 841–852 (2008).

33.Decramer, T. *et al.* Single-Unit Recordings Reveal the Selectivity of a Human Face Area. *J Neurosci* **41**, 9340–9349 (2021).

34.Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]* (2014).

35.Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**, e253 (2017).

36.Ito, T. *et al.* Compositional generalization through abstract representations in human and artificial neural networks. Preprint at https://doi.org/10.48550/arXiv.2209.07431 (2022).

37.Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci* **22**, 297–306 (2019).

38.Johnston, W. J. & Fusi, S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nat Commun* **14**, 1040 (2023).

39.Muhle-Karbe, P. S. *et al.* Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron* (2023) doi:10.1016/j.neuron.2023.08.021.

40.Epstein, R. A., Patai, E. Z., Julian, J. B. & Spiers, H. J. The cognitive map in humans: spatial navigation and beyond. *Nat Neurosci* **20**, 1504–1513 (2017).

41.Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).

42.O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map*. *Oxford University Press: Oxford, UK. (1978)* (Oxford University Press, Oxford, UK, 1978).

43.Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron* **81**, 267–279 (2014).

44.Igarashi, K. M., Lee, J. Y. & Jun, H. Reconciling neuronal representations of schema, abstract task structure, and categorization under cognitive maps in the entorhinal-hippocampal-frontal circuits. *Current Opinion in Neurobiology* **77**, 102641 (2022).

45.Vaidya, A. R. & Badre, D. Abstract task representations for inference and control. *Trends in Cognitive Sciences* **26**, 484–498 (2022).

46.Morton, N. W., Schlichting, M. L. & Preston, A. R. Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proceedings of the National Academy of Sciences* **117**, 29338–29345 (2020).

47.Marr, D. Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci* **262**, 23–81 (1971).

48.McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev* **102**, 419–457 (1995).

49.Gluck, M. A. & Myers, C. E. Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus* **3**, 491–516 (1993).

50.Benna, M. K. & Fusi, S. Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proceedings of the National Academy of Sciences* **118**, e2018422118 (2021).

51.Tang, W., Shin, J. D. & Jadhav, S. P. Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *Cell Reports* **42**, 112246 (2023).

52.Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J. & Akam, T. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nat Neurosci* **25**, 1314–1326 (2022).

53.Wood, E. R., Dudchenko, P. A., Robitsek, R. J. & Eichenbaum, H. Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron* **27**, 623–633 (2000).

54.Grieves, R. M., Wood, E. R. & Dudchenko, P. A. Place cells on a maze encode routes rather than destinations. *eLife* **5**, e15986 (2016).

55.Frank, L. M., Brown, E. N. & Wilson, M. Trajectory Encoding in the Hippocampus and Entorhinal Cortex. *Neuron* **27**, 169–178 (2000).

56.Sun, C., Yang, W., Martin, J. & Tonegawa, S. Hippocampal neurons represent events as transferable units of experience. *Nat Neurosci* **23**, 651–663 (2020).

57.Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).

58.Hesse, J. K. & Tsao, D. Y. The macaque face patch system: a turtle's underbelly for the brain. *Nat Rev Neurosci* **21**, 695–716 (2020).

59.Tanaka, K. Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience* **19**, 109–139 (1996).

60.Axelrod, V. *et al.* Face-selective neurons in the vicinity of the human fusiform face area. *Neurology* **92**, 197–198 (2019).

61.Duncan, J. & Owen, A. M. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci* **23**, 475–483 (2000).

62.Dosenbach, N. U. F. *et al.* A Core System for the Implementation of Task Sets. *Neuron* **50**, 799–812 (2006).

63.Ma, L., Chan, J. L., Johnston, K., Lomber, S. G. & Everling, S. Macaque anterior cingulate cortex deactivation impairs performance and alters lateral prefrontal oscillatory activities in a rule-switching task. *PLOS Biology* **17**, e3000045 (2019).

64.Monosov, I. E., Haber, S. N., Leuthardt, E. C. & Jezzini, A. Anterior cingulate cortex and the control of dynamic behavior in primates. *Curr Biol* **30**, R1442–R1454 (2020).

65.Economides, M., Guitart-Masip, M., Kurth-Nelson, Z. & Dolan, R. J. Anterior Cingulate Cortex Instigates Adaptive Switches in Choice by Integrating Immediate and Delayed Components of Value in Ventromedial Prefrontal Cortex. *J Neurosci* **34**, 3340–3349 (2014).

66.Cohen, Y., Schneidman, E. & Paz, R. The geometry of neuronal representations during rule learning reveals complementary roles of cingulate cortex and putamen. *Neuron* **109**, 839-851.e9 (2021).

67.Fu, Z. *et al.* Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex. *Neuron* **101**, 165-177.e5 (2019).

68.Wang, S., Mamelak, A. N., Adolphs, R. & Rutishauser, U. Abstract goal representation in visual search by neurons in the human pre-supplementary motor area. *Brain* **142**, 3530–3549 (2019).

69.Insausti, R. & Muñoz, M. Cortical projections of the non-entorhinal hippocampal formation in the cynomolgus monkey (Macaca fascicularis). *Eur J Neurosci* **14**, 435–451 (2001).

70.Aggleton, J. P., Wright, N. F., Rosene, D. L. & Saunders, R. C. Complementary Patterns of Direct Amygdala and Hippocampal Projections to the Macaque Prefrontal Cortex. *Cereb Cortex* **25**, 4351–4373 (2015).

71.Wang, J., John, Y. & Barbas, H. Pathways for Contextual Memory: The Primate Hippocampal Pathway to Anterior Cingulate Cortex. *Cereb Cortex* **31**, 1807–1826 (2020).

72.Spellman, T. *et al.* Hippocampal–prefrontal input supports spatial encoding in working memory. *Nature* **522**, 309–314 (2015).

73.Kamiński, J. *et al.* Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat Neurosci* **20**, 590–601 (2017).

74.Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. *Proceedings of the National Academy of Sciences* **105**, 329–334 (2008).

75.Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat Neurosci* **18**, 1041–1050 (2015).

76.Zhou, J. *et al.* Evolving schema representations in orbitofrontal ensembles during learning. *Nature* **590**, 606–611 (2021).

77.Zhou, J. *et al.* Complementary Task Structure Representations in Hippocampus and Orbitofrontal Cortex during an Odor Sequence Task. *Current Biology* **29**, 3402-3409.e3 (2019).

78.Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: a new window into the human brain's unique capacity for flexible cognitive control. *Cogn Affect Behav Neurosci* **13**, 1–22 (2013).

79.Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433–436 (1997).

80.Tyszka, J. M. & Pauli, W. M. In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Human Brain Mapping* **37**, 3979–3998 (2016).

81.Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204–224 (2006).

82.Hill, D. N., Mehta, S. B. & Kleinfeld, D. Quality Metrics to Accompany Spike Sorting of Extracellular Signals. *J. Neurosci.* **31**, 8699–8705 (2011).

83.Courellis, H., Nummela, S., Miller, C. & Cauwenberghs, G. A computational framework for effective isolation of single-unit activity from in-vivo electrophysiological recording. in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)* 1–4 (2017). doi:10.1109/BIOCAS.2017.8325164.

84.Panzeri, S., Moroni, M., Safaai, H. & Harvey, C. D. The structures and functions of correlations in neural population codes. *Nat Rev Neurosci* **23**, 551–567 (2022).

85.Anderson, B., Sanderson, M. I. & Sheinberg, D. L. Joint decoding of visual stimuli by IT neurons' spike counts is not improved by simultaneous recording. *Exp Brain Res* **176**, 1–11 (2007).

86.Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology* **100**, 1407–1419 (2008).

Chapter 3

Temporally static and dynamic neural representations in human hippocampus and medial frontal cortex support persistent behavior

Abstract: Humans are capable of conducting tasks for extended periods of time after receiving a single instance of instruction. Such behavior necessitates active maintenance of a representation of task context that can dynamically shift as cognitive resources are updated in service of new goals. While previous studies have implicated both the medial frontal cortex and the anterior hippocampus in representations of task context, the implementation of that representation at the level of single neurons in the human brain remains an open question, particularly when context is not re-cued trial by trial and must be internally maintained over the course of minutes, far-exceeding the timescale considered by previous neural models of persistent activity. To clarify which encoding strategies are used by neurons in two medial frontal cortical structures: the dorsal anterior cingulate (dACC), pre-supplementary motor area (preSMA), and in the anterior hippocampus (HPC), we conducted single-neuron recordings in epilepsy patients who were instructed to alternate between cognitive tasks in two different experimental settings. In the first experiment, we recorded 970 neurons in 13 patients (33 sessions, dACC = 329, preSMA = 438, HPC = 203). We find a strong dissociation in coding strategy between these regions. HPC exhibits a temporally dynamic code, lacking neurons that stably encode individual task contexts independently of the progression of time through the experiment, thus leading to a population code for context that rapidly orthogonalizes. dACC and preSMA on the other hand stably encode task context over many minutes both during baseline and stimulus processing at the level of single units, leading to a population code that strongly generalizes across time throughout the experiment. To determine if temporal stability is an intrinsic property of the neurons in each region, we analyzed a second experiment which shared block- and trial-level structure with the first experiment, but featured very different trial-level task demands. In this experiment (17 patients, 42 sessions, 499 HPC neurons), we found that the hippocampal task context representation had stabilized, exhibiting comparable cross-temporal generalization properties to dACC and preSMA in the first experiment. These findings call attention to an intermediate-temporal scale upon which persistent single-neuron activity in the human brain gives rise to representations of task context variables, and indicates that the temporal stability of these representations is not an immutable region-specific property, but rather changes as a function of task demands.

Introduction:

Humans can engage in persistent behavior for extended periods that can be many orders of magnitude longer in duration than the stimuli that triggered them. For example, a single question conveyed over the course of several seconds could lead to a 2-minute trip to the kitchen to fetch some water, or a 20-minute quest to find lost keys depending on the nature of the posed question. To support temporally extended behavior, the human brain encodes a representation of the task or the environmental state that was signaled, either directly or indirectly, by the stimulus in question[1–5]. Here, we refer to such high-level, behavior-constraining variables as task context variables, whose value determines or modifies the actions deployed in response to many different sensory stimuli. Such task context variables can also contain information related to goals or targets, which typically specify individual stimuli or environmental states which must be attained. A task context variable may be specified directly by external stimuli, either symbolically or through language-based instruction[6,7], or it may be induced implicitly by the temporal statistics of the environment [8–10]. In the latter case, changes in context can also be inferred from feedback[11–14]. In either case, the state of the task context variable must be persistently represented in the brain to constrain behavior long after the stimulus signaling the change in task context has been removed. Numerous regions in the human frontal and temporal lobe have been implicated in the representation of such context variables[15–25], but many questions related to the implementation of these variables at the level of single-neuron activity remain open.

What computationally advantageous properties should these task context representations have to support flexible behavior? One potential property of these representations that has been given little attention is the degree of cross-temporal stability they exhibit on the timescale at which human behavior typically persists. There are many computational advantages for neurally representing task-relevant variables in a temporally static format, including facilitation of generalizing behavior flexibly across arbitrarily longer periods of time[26,27], and for generalizing behavior to new stimuli[28]. These advantages need to be balanced against the storage capacity benefits conferred by a dynamic code that evolves over time[29,30]. Considerable research has been dedicated to studying the presence of static and dynamic coding on the timescale of seconds in individual trials during working memory tasks[31], particularly in the human brain[32–34], and to the study of changes in neural tuning through representational drift on the timescale of days to weeks[35]. However, the neural underpinnings of persistent behavior on the timescale of ~1-10min, an interval which incidentally contains the average duration for human sustained attention on a task (~400s), remains elusive[36]. Here, we ask whether there are stable representation of task context that persist for many minutes after instructions have been given. We examine where in the brain such representations exist and how they are implemented at the level of single neurons.

Various brain structures in the frontal and temporal lobes are known to represent task context (task sets), but little is known about the temporal stability of such representations over minutes. Two structures in the medial frontal cortex (MFC), namely the dorsal Anterior Cingulate Cortex (dACC) and the pre-Supplementary Motor Area (preSMA) have been studied extensively for their role in flexible, temporally extended behaviors[37]. The dACC has been described as a "storage buffer" for the persistent representation of task context variables that mediate the behavioral policy deployed in a given environment[38]. The preSMA is also critical for the task-dependent selection of appropriate actions and the mediation of switching between action sets[39], both of which occur in a temporally extended manner. For explicitly instructed tasks, where MFC structures are known to causally mediate persistent behavior[40,41], one might hypothesize that task context representations emerge immediately following instruction and persist in a temporally static rate code in neurons until a new

task is instructed, potentially several minutes later. This hypothesis has never formally been tested. In the temporal lobe, there is also considerable evidence that the hippocampus forms representations of task context variables when constructing a cognitive map of the environment, mixing combinations of context variables and stimuli of different modalities that are behaviorally relevant in that environment. In recent work, we have demonstrated that an explicit representation of context emerges in hippocampal neurons as subjects learn to perform inference on a latent context variable whose state was never overtly signaled[12]. Notably, in this task, a context representation was almost entirely absent from the medial frontal cortex, raising the question of how the hippocampal implementation of a task context variable differs between this and other experiments where context representations are prominent in MFC.

In the hippocampus, the encoding of static task contexts must also be balanced against simultaneously present dynamic firing patterns that are thought to support episodic memory and allow for the readout of other variables such as the passage of time. At the single neuron level, the presence of time cells and ramp cells, whose activity simultaneously codes for the passage of time across several timescales ranging from milliseconds to hours, has been documented in several species including humans[42–45]. At the population level, evidence of dynamically coding for the continual passage of time is thought to be important for episodically fingerprinting continuous sensory experience[46]. These hippocampal neural dynamics form a "temporal context" that is reinstated during episodic recall of temporally remote memories[47]. How hippocampal neurons fulfill this computational role of representing a continually changing temporal context while simultaneously representing a temporally static and repeatedly encountered task context remains an open question.

Here, we ask the question of how the hippocampus and medial frontal cortex can simultaneously accommodate static and dynamic representations of task context variables of different kinds. We re-analyzed single neuron recordings from two experiments that jointly allow us to examine different aspects of this question. In both experiments, subjects dynamically change their responses to visual stimuli as a function of a high level contextual variable[6,12]. Based on the first experiment, we here show that the task context representation is implemented in a region-specific manner, with hippocampal neurons representing context in a manner that dynamically reorganizes itself across time throughout the experiment, thereby simultaneously multiplexing global temporal information and task context information. This encoding strategy stands in contrast to the dorsal Anterior Cingulate Cortex (dACC) and pre-supplementary motor area (preSMA) whose neurons encode a static task context representation that generalizes across time. Based on the second experiment, we then show that the temporally dynamic code for context is not always present in the hippocampus. Rather, we find that the hippocampal context representation exhibits considerable temporal stability in this second experiment across time throughout the experiment, notably in the absence of a simultaneous context representation in dACC and preSMA. Together, our findings provide insight into the implementation strategies used by different frontal and temporal brain structures to persistently represent task context variables that are needed to guide behavior over long time periods, and raise questions about the flexibility of that implementation as a function of task demands.

Methods:

Participants: The study participants were adult patients being surgically evaluated for invasive treatment of drug-resistant epilepsy (see Table 1). These patients were treated at Cedars-Sinai Medical Center (CSMC) and Toronto Western Hospital (TWH). All patients provided informed consent and subsequently volunteered to participate in this study. All research protocols were reviewed and approved by the institutional review boards of CSMC, TWH, and the California Institute of Technology.

Experiment 1:

One group of 13 patients (see Table 1) performed 33 sessions of an experiment that required alternation between two task contexts, one requiring semantic categorization of visual stimuli (categorization task), and the other requiring recognition memory of those same stimuli (memory task). Patients performed eight blocks of 40 trials, with the required task alternating between categorization and memory across consecutive blocks. Patients always began with categorization for the first block of the session. Text-based instructions for the required task in the current block were provided at the start of each block, and were not re-cued until the next block thus requiring patients to persistently remember the current task being executed. Patients could spend as much time as needed on instruction screens, and voluntarily proceeded into every new block after reading the instructions. Both tasks were formulated as binary (yes/no) questions of the form: "Is this an image of an X?" for the categorization task, where X was one of four unique semantic categories, and "Have you seen this image before?" for the memory task. The number of new and old images of each semantic category were balanced in each block to prevent response biases and to facilitate balanced decoding analyses. Individual trials consisted of a jittered pre-stimulus baseline (1s to 2s) followed by image presentation for a variable amount of time until the patient's response for that trial was provided. Patients provided trial responses using either a left/right button press on a CEDRUS binary response box or by saccade left/right to indicate True/False for each trial. The response modality was randomized over blocks and was re-cued every 20 trials. Following the response, the stimulus was removed from the screen and the baseline period for the next trial was initiated. Trial-by-trial feedback was not provided.

Experiment 1 Control Variant:

A control variant of the experiment described above, which was designed to disentangle stimulus processing from decision variables and motor plans, was utilized for a fraction of the sessions (5/13 patients, 6/33 sessions). In this variant, instead of being allowed to respond freely as soon as the image appeared on the screen, images were presented for a fixed interval (1s), then trial responses were allowed after a jittered delay (0.5s to 1.5s) when a response cue appeared on the screen. These task variations were uniformly applied to all trials, and thus could not have generated a bias in the representation of one task condition over another. Furthermore, none of the analyses shown here were performed on a response-aligned time window. Nevertheless, the core analyses in this work were re-conducted on neural pseudopopulations constructed exclusively from these control sessions, and all findings were re-capitulated (See Fig. 3.S4).

Experiment 2:

A separate group of patients 17 patients (see Table 1) performed 42 sessions of a second experiment (180-320 trials/session, 10-16 blocks/session) that shared key structural elements with the first experiment both at the trial level and at the block level that allowed for direct comparison of neurophysiological task responses across the two experiments. In sessions of this experiment, patients learned arbitrary stimulus-response-outcome (SRO) associations for four unique stimuli arbitrarily associated with either a left or right button press in one of two latent contexts. The contexts were related in that the required response for each stimulus was inverted between the two contexts (e.g. stimulus A was associated with left button press in context 1 and right button press in context 2, etc..). Blocks consisted of 15-32 trials in a given context before a covert switch to the other context. Trials consisted of a pre-stimulus baseline (1.5s to 2.5s), followed by a speeded response (left/right button press) provided with the onset of the stimulus, and the presentation of an outcome (either reward or "incorrect") for 1s following a fixed 0.5s delay. Rewards were provided deterministically such that, if a patient had learned the SRO map in a given context and suddenly encountered an incorrect trial, this was an unambiguous signal that the state of the latent context variable had changed. Patients could learn to perform inference on the state of the latent context variable such that, after a single incorrect trial, patients could infer that the context had changed and update all stimulus-response associations in accordance with the new context.

Cross-Experiment Comparison:

The two experiments considered here share a considerable amount of structure that facilitates cross-task comparison. Both experiments contained a binary task context variable that was designed to elicit different responses for the same stimuli depending on the state of the context variable. Both experiments have a blocked structure such that the context variable varies slowly with time, and many trials must be completed in a given block before the context changes. The current context is not re-cued in either experiment, with explicit instructions being provided once at the beginning of each block in experiment 1 and never being provided in experiment 2, thus requiring a persistent representation of task context in both experiments to achieve high performance. Trial structure is also very similar between the two experiments. In both cases, trials consist of a pre-stimulus baseline where a single gray fixation cross is present on the screen for ~2s. Trial onset is marked by the appearance of a single image subtending ~10 visual degrees. The image is removed when the patient provides a response for that stimulus in accordance with the currently instated context. Responses were formulated as binary in both experiments for all task contexts, with the Categorization and Memory tasks in experiment 1 formulated as yes/no questions and the two latent contexts in experiment 2 requiring left/right button presses. Thus, for the time periods analyzed here including baseline (-1s to 0s prior to stimulus onset) and stimulus processing (0.2s to 1.2s following stimulus onset), patients were engaged in cognitive tasks with roughly similar structure and comparable cognitive demands. Elaboration of the differences between these experiments and associated limitations is provided in the discussion.

Electrophysiology:

Electrode Placement and Recording: Extracellular electrophysiological recordings were conducted using microwires embedded within hybrid depth-electrodes (AdTech Medical Inc.) implanted bilaterally into the hippocampus, amygdala, dorsal anterior cingulate cortex, pre-supplementary motor area, ventromedial prefrontal cortex, in addition to variable unilateral or bilateral electrodes in ventral temporal cortex as determined by clinical needs. Broadband potentials (0.1Hz – 9kHz) were recorded continuously from every microwire at a sampling rate of 32kHz (ATLAS system,

Neuralynx Inc.). All subjects included in the study exhibited voltage waveforms consistent with well-isolated single-neuron action potentials in at least one implanted microwire.
Electrode Localization: Electrode localization was conducted using a combination of pre-operative MRI and post-operative CT using standard alignment procedures as previously described[6]. Electrode locations were also co-registered to the to the MNI152-aligned CIT168 probabilistic atlas[48] for standardized location reporting and visualization. Placement of electrodes in gray matter was confirmed through visual inspection of subject-specific CT/MRI alignment, and not through visualization on the atlas.

Spike Detection and Sorting: Raw electric potentials were filtered with a zero-phase lag filter with a 300Hz-3kHz passband. Spikes were detected and sorted using the OSort software package[49]. All spike sorting outcomes were manually inspected and putative single-units were isolated and used in all subsequent analyses. All processing and analysis of neural data was performed using MATLAB (The Mathworks, Inc., Natick, MA).

Analysis Periods, Single-Neuron Tuning, and Construction of Pseudo-populations:

All analyses were conducted on firing rates of neurons computed during two trial epochs the baseline period (base), defined as -1s to 0s preceding stimulus onset on each trial, and the stimulus period (stim), defined as 0.2s to 1.2s following stimulus onset on each trial. Firing rate vectors for every neuron were constructed during both trial periods.

Single-neuron tuning properties were assessed using univariate and multivariate ANOVAs applied to the firing rate vectors for each neuron independently unless otherwise stated. Task context was encoded as a categorical variable with two levels for both experiments. Semantic image category in experiment 1 and stimulus identity in experiment 2 were encoded as categorical variables with four levels. Any references to depth-of-tuning of a neuron for one of these variables or their interaction (e.g. Context x Stimulus identity) refer to the F-value of the variable in question when the ANOVA is performed on the trial-level firing rate vectors either during the stimulus or baseline periods. Note: all ANOVA analyses were performed using spikes counted during correct trials for the stimulus period and during baselines preceding correct trials. For some control analyses, an ANOVA F-statistics distribution matching procedure is also performed between neurons recorded in different regions. To match F-statistic distributions, valid pairs of neurons were identified, one from each region, whose F-statistics for the context variable were within 0.1. The candidate pair was removed from the pool of available neurons, and another pair was selected until no more valid pairs were present, at which time all remaining neurons were excluded from the subsequent distribution-matched analysis. This procedure creates two populations of neurons, one for each region participating in the balancing procedure, whose ANOVA F-statistic distributions are statistically indistinguishable.

Firing rate vectors including all trials for single neurons were concatenated to create neural pseudo-population matrices of dimension (# of trials x # of neurons) on which all subsequent decoding analyses were performed. These pseudo-populations only consisted of neurons that exhibited at least 0.1Hz firing rate averaged over the entire recording session. Repeated recording sessions of a given experiment with a given subject were typically separated by several days, and neurons recorded during these repeated sessions were treated as independent neurons in the pseudopopulation. No stimuli or stimulus-response pairings were ever re-used in repeated recording sessions for either experiment, thus preventing potential behavioral and neural confounds related to recognition memory signals across recording sessions.

Trial-Balanced Decoding Analysis:

Decoding analyses were performed using a linear support vector machine, and all decoding accuracies are reported out-of-sample using 5-fold cross-validation unless otherwise specified (e.g. cross-condition generalization). Model fitting was performed using the "templateSVM" with a linear kernel and the "fitecoc" model-fitting methods from the Stats toolbox of Matlab 2021b. Trial-balanced decoding of a task variable was conducted by concatenating firing rate vectors for neurons in a given region to construct firing rate matrices, i.e. the pseudopopulation response matrix, with dimensions KT x N, where K is the number of conditions (typically 2 apart from image category decoding, where there were 4 categories), T is the number of correct trials per condition, and N is the number of neurons. Neurons were excluded from the pseudopopulation if they there were fewer than 15 correct trials per condition in that recorded session. Neurons with more than the minimum number of correct trials had their trials randomly sub-sampled so that the number of correct trials per feature and per condition could be matched prior to decoding. To account for the presence of noise correlations between simultaneously recorded neurons, trials were also shuffled independently for each neuron within-condition prior to decoder fitting. To account sub-sampling and randomization bias, the trial sub-sampling and within-condition shuffling procedure were repeated 250 times, with reported decoding accuracies being the average over these repeats. All error bars shown are standard deviations over the distribution of decoding accuracies unless otherwise specified. Null distributions for decoding analyses were constructed by shuffling condition labels and reporting out-of-sample decoding performance, again with 250 repetitions. The significance of individual decoding accuracies was determined by reporting the p-value of the average decoding accuracy against the gaussian maximum likelihood fit of the null distribution. The significance of the difference between decoding accuracies (e.g. between two areas) was determined by reporting the p-value of the true difference against a null distribution of the difference constructed by computing all pair-wise differences between points in the null distributions for each of the two decoding accuracies being considered.

Decoder Cross-Condition Generalization:

Generalization analyses for decoders are performed by training a decoder to discriminate between two states of a variable in one condition and testing whether that decoder performs above chance in discriminating the same variable states in another condition. Critically, in order for such analysis to be performed, the variable in question must un-ambiguously be specified in the source condition (on which training is performed), and the target conditions (on which testing, or generalization is performed). In the case of task context, this generalization performance was performed over trial phases (baseline/stimulus), block phases (first block half/second block half), and over experiment phases (block pairs). Furthermore, generalization analysis requires that the training and testing feature spaces are aligned. In this case, the requirement of a minimum number of 15 correct trials per neuron per condition was also independently applied to the source and target conditions for these generalization analyses. For example, if training a context decoder on blocks 1/2 and testing on blocks 3/4, a neuron must have at least 15 correct trials per context in blocks 1/2 and 3/4 independently. Neurons that did not meet these criteria separately for both source and target conditions were excluded from the analysis. Since in a generalization analysis, decoding accuracy is out-of-sample by construction, cross-validation was not used, and all available trials were used for training and testing. However, since trial sub-sampling and within-condition shuffling were also

employed here, generalization decoding accuracies were also reported as the average over 250 repetitions.

To control for the amount of task variable information available to a decoder during training, a generalization index is reported that normalizes the performance of the decoder in the generalized conditions to the out-of-sample performance of the decoder in the training conditions. Specifically, the generalization index is computed as:

$$gi = \frac{\overline{g} - cl}{\overline{t} - cl}$$

where $\overline{g}$ = mean performance over all instances of generalization for all trained decoders, $\overline{t}$ = mean out-of-sample training performance for all decoders, cl = chance level. For example, when computing the generalization index for context decoders over block pairs, for the 8 blocks, $\overline{t}$ is the average over training performance of 4 context decoders, $\overline{g}$ is the average performance over 12 instances of generalization (note: generalization performance of decoders is not necessarily symmetric), and chance level is 0.5.

Coding Vector Angles:

The coding vectors used in all angle analyses are the β coefficients of decoders trained to decode task variables, typically task context, in different phases of each experiment. These coefficients are the weights returned by the SVM model fitting procedure. They reflect the relative contribution of each feature (neuron) to the decoder, with better-tuned neurons to the variable in question being assigned higher magnitude coefficients. The coefficients are also signed to reflect which of the conditions that feature prefers (e.g. the context assigned to +1 or -1 for binary classification). Class label assignment for classification was kept constant to allow for decoder generalization and meaningful estimation of the angle between coding vectors for different decoders. Angles between coding vectors were computed in by applying the definition of the dot product in N-dimensional neural state spaces, where N was the number of neurons included in the given analysis. N was matched for all regions within a given analysis so that angles reported in the same plot were directly comparable, and not computed using vectors with different dimensions. All angles between coding vectors were reported as the average over the 250 repetitions of decoder estimation described above. Null distributions were constructed by pooling together all possible pair-wise angles between shuffle-null decoders trained as described in the "Trial-Balanced Decoding Analysis" section. Angles are computed between: context decoders trained on different block pairs, context decoders trained on the baseline and stimulus processing periods, and image category decoders trained on different block pairs. In all these cases, the same neurons are used as the features in the two decoders between which the angle is being computed, so the neural state spaces are aligned by construction and the angle between coding vectors is readily interpretable as an overlap in the coding direction for the variable being decoded.

Population Vector Autocorrelation:

The trial-level autocorrelation of the neural population in each region was estimated computing the Pearson correlation between population vectors for every pair of trials present in each experiment. For experiment 1, all trials were included in this analysis leading to 320^2 dimensional autocorrelation matrices. Population vectors here were sub-sampled to match the smallest number of neurons available in any region as previously described so that correlation values were directly comparable across regions. For experiment 2, since block lengths were randomized for every block in every session, all blocks were sub-sampled in length to match the smallest available number of trials in a

given block. Re-sampled estimation of population vector autocorrelation for this experiment was simultaneously performed over neurons and trials-in-block. Reported autocorrelation heat maps are an average over 250 repetitions of re-sampled estimation, and are convolved with a 2D Gaussian filter with a standard deviation of 1 for visualization purposes. All subsequent analysis on population autocorrelations was performed on the un-smoothed maps. Block-wise decorrelation curves are computed by taking the average pairwise correlation between all trials within the same block for block distance 0, average pairwise correlation between all trials one block apart for block distance 1, and so on. On-diagonal correlations (trial with itself) are ignored to prevent artificial inflation of block distance – correlation. Even and odd block distances are colored differently to reflect the fact that even block distances correspond to trial-level correlations within the same task context and odd block distances correspond to trial-level correlations between different task contexts, since task contexts alternated at the block level in both experiments.

Two metrics are further derived from these curves: the decorrelation rate and the relative context modulation. The decorrelation rate for the neural population in each region was quantified by performing linear regression on the decorrelation curves and reporting the absolute value of the estimated slope. This slope was always negative for all neural populations and time periods considered as there was no instance in which the self-similarity of a neural population increased over time. The decorrelation rate is reported in units of $block^{-1}$ since the slope is an estimate of the change in linear correlation of the population (unitless) divided by the block distance, which is measured in units of blocks by definition. The relative context modulation is defined as the average absolute difference in linear correlation between trials 0 blocks apart and trials 1 block apart, normalized by the decorrelation rate. Since the absolute block 0-1 difference is also computed in units of $block^{-1}$, the relative context modulation is a unitless quantity that reports the effect of re-cuing task context on a population of neurons, normalized to the baseline tendency for that neural population to decorrelate over time. A relative context modulation of 1 indicates that the change in representation experienced by a neural population over the timespan of a block due to intrinsic decorrelation and due to explicit cueing of a different task are equivalent, with values greater than 1 and less than 1 indicating dominance of task-recuing effects and intrinsic decorrelation respectively. Reported values for both the decorrelation rate and the relative context modulation are averages over the 250 repetitions of re-sampled autocorrelation estimation, and error bars are s.e.m. over these repetitions.

Results:

The hippocampal context representation is temporally dynamic, whereas MFC is static.

In the first experiment we analyzed (henceforth Experiment 1), neurosurgical patients completed a blocked, context-dependent decision making task. Patients answered binary "Yes vs. No" decisions for each image shown according to the currently active context (Fig. 3.1A). The two task contexts required answering either a semantic categorization question ("Is the image a member of category X?") or a recognition memory question ("Have you seen this image before?"). These task contexts were explicitly provided to patients once at the beginning of every block, and needed to be remembered by the patient for the ensuing 40 trials in the block (lasting 115.85s ± 4.31s, mean ± s.e.m. over blocks). That is, successfully perfomring this task required working memory for the task context for up to 2 minutes. Patients completed 320 trials (8 blocks) in each session of this experiment, which lasted on average 1100.0s ± 37.1s (mean ± s.e.m. over sessions). The first block was always a categorization block and every ensuing block alternated between memory and categorization (Fig. 3.1A, bottom). Each trial consisted of a pre-stimulus baseline, followed by presentation of the stimulus, which was displayed until patients provided a response. Trial-by-trial feedback was not provided. Patients (n = 13) performed sessions (n = 33) of this experiment with high accuracy (83.6% ± 1.1%, mean ± s.e.m. over sessions) and with rapid trial-level response times relative to stimulus onset for each trial (1.34s ± 0.10s, mean ± s.e.m. over trials). Extensive experimental details and behavioral analyses have been provided in our previous work[6].

The activity of 970 single neurons was recorded from the Hippocampus (HPC, 203 neurons), dorsal Anterior Cingulate Cortex (dACC, 329 neurons), and pre-Supplementary Motor Area (preSMA, 438 neurons) across all sessions (Fig. 3.1B,C). Neurons were differentially responsive to several of the high level cognitive variables within the experiment, including task context and semantic image category (example HPC neurons shown in Fig. 3.1D,E). Univariate tuning analysis performed on firing rates estimated during the baseline (-1s to 0s prior to stimulus onset) and stimulus presentation (0.2s to 1.2s after stimulus onset) revealed above-chance tuning to task context during the baseline, and task context, image category, and interactions in all three considered brain regions (Fig. 3.1F, $p < 0.05$, full 2-Way ANOVA for context and category during stimulus, 1-Way ANOVA for context during baseline). Additional example neurons encoding task context during the baseline and stimulus periods from all three regions are shown in Fig. 3.S1. Thus, at the single-neuron level, there are neurons whose firing rate is significantly modulated by the instructed task context on average across the entire experiment during both stimulus processing and baseline periods.

To investigate the dynamics of task context representations across the entire duration of a block (~2 mins) and the entire task (~20 mins), we first conducted population-level analysis. We constructed pseudopopulations of neurons in each region pooled across sessions, and trained linear SVMs to decode task context on individual trials during the stimulus and baseline periods (Fig. 3.2A, inset, see Methods for details). Task context was significantly decodable in HPC ($60.3\%\ base, p_{base} = 0.007, 71.5\%\ stim, p_{stim} = 9.5x10^{-6}$), dACC ($83.9\%\ base, p_{base} = 1.4x10^{-13}$, $91.7\%\ stim, p_{stim} = 0$), and preSMA($82.9\%\ base, p_{base} = 6.1x10^{-15}$, $99.9\%\ stim, p_{stim} = 0$). All reported p-values are computed against a null distribution of retrained, trial-label shuffled decoders, only correct trials were used for decoder training/testing, and all decoders were matched for number of neurons and number of trials per condition through random sub-sampling unless otherwise specified.

Since the above analysis relied on pooling trials across the duration of each session, we next asked how stable the representation of task context was throughout each session. If the representation

were static, decoders trained to discriminate task context during one part of the experimental session should generalize to other parts of the session. If, on the other hand, the task context representation were dynamic, then context decoders trained in one part of the session should perform poorly when generalized to other parts of the session. Context decoders were trained on adjacent block pairs, each of which contained one task context due to the alternating structure of the experiment. Generalization to increasingly distant block pairs reveals a locally decodable context variable that does not generalize across time during stimulus processing in HPC (Fig. 3.2B; as indicated by higher on-diagonal vs. off-diagonal decoding accuracy). In contrast, the code for context in preSMA (Fig. 3.2C) and dACC (Fig. 3.2D) generalized well across time, with high decoding accuracy both on and off-diagonal (see below for quantification).

We quantified the relative degree of cross-temporal decoder generalization by computing a generalization index that captures decoder generalization performance normalized by cross-validated testing performance on the same period of time (see Methods). For the cross-temporal analysis, a generalization index of 0 indicates that none of the encoding of the task context variable in the training blocks is present in the blocks to which the decoder was generalized. A generalization index of 1, on the other hand, indicates that the context variable is as decodable in the generalized blocks as it was in the training blocks. Generalization index analysis revealed that both dACC (Fig. 3.2E, blue vs red, $p < 0.001$, Permutation test) and preSMA (Fig. 3.2E, blue vs red, $p < 0.001$, Permutation test) exhibited significantly greater cross-temporal context generalization than the HPC.

Since high decoder performance could, in principle, be driven by a small number of well-tuned neurons, we also computed the angle between context coding vectors in each block pair. We define the coding vector here to be the normal vector to the hyperplane learned by each decoder during training. In this analysis all neurons have equal weight, making it insensitive to tuning of only a small subset of neurons. We find that pairs of context coding vectors between any two block pairs significantly differ from orthogonal both during the stimulus period for dACC (Fig. 3.2F, blue, 76.5º, $p < 0.001$ against shuffle null) and preSMA (Fig. 3.2F, green, 59.0º, $p < 0.001$ against shuffle null), indicating significant context coding vector alignment across block pairs. In HPC, on the other hand, context coding vector angles did not significantly differ from orthogonal across block pairs (Fig. 3.2F, red, 88.5º, $p = 0.46$ against shuffle null). Note that angles here are computed in a 150-dimensional space constructed by sub-sampling neurons randomly over iterations. All of the above cross-temporal context generalization findings were also present during the baseline period for the three regions (Fig. 3.S2), with the notable exception that the cross-block pair generalization index did not significantly differ between dACC and preSMA (Fig. 3.S2D, blue vs green).

Together, these findings indicate that the code for task context is dynamic in the HPC. In contrast, in the MFC, context coding is static. What about the neuronal responses in the HPC causes the code to be dynamic? One possibility is that individual HPC neurons do not reliably represent a given task context throughout a session. This stands in contrast to preSMA and dACC, where generalizing context decoders imply that single neurons in these regions represent task context with a more static rate code. To test this prediction, we fit 2-Way ANOVAs on every neuron individually with two categorical regressors for block number and task context. We reasoned that for activity of neurons supporting cross-temporal generalization, the main effect of task context would explain more variance than the block number main effect (and vice versa for neurons supporting a dynamic code). We therefore next compared the amount of variance explained by the two main effects (ΔF-statistic, see methods). Variance in single-neuron responses in the hippocampus was on average better explained by the block-specific regressor during both stimulus (Fig. 3.2G, red, $p = 0.017$, Student's t-test) and baseline (Fig. 3.S2F, red, $p = 0.02$, Student's t-test) periods. In dACC and

preSMA, on the other hand, the task-context regressor explained significantly more variance during the stimulus (Fig. 3.2G, $p_{green}, p_{blue} < 0.001$), but not during the baseline (Fig. 3.S2F, $p_{green}, p_{blue} > 0.01$) period, though the general trend remained. Thus, the cross-block pair generalization pattern in these regions can be accounted for by different encoding strategies at the level of single-units. These differences in encoding strategy can be directly visualized by plotting rasters and PSTHs of neurons spanning entire blocks, revealing that dACC (e.g. Fig. 3.2H, 3.S3A) and preSMA (e.g. Fig. 3.2I, 3.S3B,C) neurons exhibit task context modulated firing that persists for the duration of entire blocks, while such features are absent from HPC neurons (e.g. Fig. 3.S3D-F).

These analyses together indicate that, while instructed task context is decodable in neural populations in frontal and temporal brain structures, the implementation of that context representation varies considerably. While medial frontal cortical context representations generalize across time, the task context representation in the hippocampus appears to reorganize dramatically across time to the point that it is orthogonalized at adjacent timepoints separated by the span of a few minutes.

Control analyses for context and image category representations.

The dynamic code for task context on the timescale of blocks in the hippocampus raised several questions we next addressed. First, were shifts in the context representation over time driven by abrupt changes at the beginning of each block, leaving a static context representation within each block? Alternatively, were there within-block changes in the context representation that could lead to context code orthogonalization even within a given block? Second, can the temporal stability/instability of the context representation in different regions can be explained by tuning strength differences of single-neurons and/or by recording instability?

To address the first question, we performed a block-half decoding and generalization analysis where decoders for task context were trained using data from the first half and second half of every block, then evaluated on the second half and first half respectively. If the hippocampal context representation exhibited within-block dynamics, then first- and second-block half decoders should fail to generalize. If, however, the context-code remains static within individual blocks, then a context decoder trained on the first block half should do approximately equally well on the second block half (and vice versa from second to first). We find that for HPC, dACC, and preSMA, context decoders trained on one block half generalized well to the other block half during both the stimulus period (Fig. 3.S4A-C) and the baseline period (Fig. 3.S4D-F). The block-half generalization indices were close to 1 for all three regions (Fig. 3.S4C,F). This data indicates that the context code within-block is stable for all three regions, suggesting that the dynamic code is due to changes that occur at the transition between blocks.

Next, to address the possibility that representational stability in the MFC arose due to neurons in dACC and preSMA being more strongly univariately tuned to task context to begin with, we matched the distribution of single neuron ANOVA F-statistics for the main effect of context across the three regions before re-computing cross-block pair context decoding (see Methods for details). Following distribution matching, 174 neurons remained in each area for the subsequent cross-block pair analysis. This analysis revealed that the cross-block pair generalization indices for context remain qualitatively unchanged for the three regions during both the stimulus (Fig. 3.S4G,H) and baseline (Fig. 3.S4I,J) periods, and thus that the stability of the context code in dACC and preSMA cannot be explained by stronger univariate context coding at the level of single neurons.

To demonstrate that the temporal stability of the context code in each region was robust to variation in trial and block duration within the experiment, we also re-conducted the cross-block pair

stability analysis in isolation on the subset of neurons recorded during sessions where a control variant of this experiment was used (6 sessions, 36 HPC, 75 dACC, 87 preSMA neurons). This variant required patients to wait for a 0.5-1.5-second period following stimulus offset and until the onset of a response cue to provide their response to each trial. This change was applied to both categorization and memory trials. For these experimental sessions, mean trial duration was 2.47s ± 0.04s (vs 1.08s ± 0.05s in non-control session) and mean block duration was 162.54s ± 1.83s (vs 105.47s ± 2.23s in non-control sessions). We reasoned that the ~250% increase in trial duration and ~60% increase in block duration might encourage stabilization of the hippocampal context representation as context-appropriate behavior needed to be maintained over even longer periods. Analysis of this data revealed, again, that dACC and preSMA exhibited significantly greater cross-block pair context stability both during stimulus (Fig. 3.S4K) and baseline (Fig. 3.S4L) periods, so the modifications in the control variant of experiment 1 did not qualitatively affect the temporal stability of the task context representation present in the three regions.

To address the question of recording stability, we analyze the geometry of a second task variable, image category, that is known from other work to be encoded by HPC neurons in a static manner[7,32,33]. We thus examined the temporal stability of the encoding of category as a control. To do so, we conduct an identical cross-block pair generalization analysis, but decoding image category. If the image category code, which is simultaneously encoded alongside the task context representation during stimulus processing, also appears to reorganize across block pairs, then the lack of decoder generalization shown here could trivially arise due to recording instability in the hippocampus. However, we find that image category was decodable from all three regions, most prominently in the HPC ($59.5\%, p = 0,$ against shuffle null; chance=25%, Fig. 3.S5A,B). The image category code was uniquely static in the hippocampus, with a significantly greater generalization index than MFC (Fig. 3.S5C-F), and significantly overlapping image category coding vectors between adjacent block pairs (Fig. 3.S5G). Furthermore, the image category coding vectors were orthogonal to the context coding vector during the stimulus period in all three regions (Fig. 3.S5H). This analysis shows that, in the same group of neurons, image category is encoded in a static manner without reorganization over time. Thus, the temporally dynamic encoding present in the same group of neurons is a property specifically of the task context variable.

Taken together, these analyses suggest that the code for context is relatively static within a given block in all regions including the hippocampus. In contrast, the code for context changes across consecutive block pairs in the hippocampus, and this effect cannot be explained by weaker context tuning at the level of single-neurons, or by an overall lack of recording stability as other simultaneously encoded task variables did exhibit cross-temporal stability.

Hippocampal neural population exhibits faster temporal dynamics than MFC.

The hippocampus is frequently studied for its role in contributing to episodic memory through the representation of temporal context[46,47,50]. One way temporal context representations are implemented is through slow, gradual drifts in neural population activity within the hippocampus. We next asked whether such slowly changing temporal context representations were present in our data, and if so, how they were related to our finding of the encoding of task context. To do so, we examined neural population dynamics with single-trial resolution over the timescale of the experiment (~20 mins) in a decoder-agnostic manner. The autocorrelation of the population response (see methods) revealed a striking pattern: during the stimulus period, the HPC neural population response gradually and continually decorrelated as indicated by positive near-diagonal population vector correlations and increasingly negative correlations with increasing trial distance (Fig. 3.3A). The preSMA, on the

other hand, exhibited a checkerboard-like autocorrelation pattern, with alternating groups of trials corresponding to individual blocks exhibiting positive within-context correlation and negative across-context correlation at trial-lags spanning the entire experiment (Fig. 3.3C). A qualitatively similar checkerboard pattern emerged in the dACC population-vector autocorrelation (Fig. 3.S6E). The same analysis performed during the baseline revealed qualitatively similar, but less visually pronounced results in all three areas (Fig. 3.S6A,C,G). These plots qualitatively track their respective cross-temporal generalization plots in Fig. 3.2, 3.S2, and were reproducible in a single patient (P44CS) who happened to provide enough simultaneously recorded neurons (37 HPC, 62 preSMA) such that the analysis could be performed for some regions.

Using the population-vector autocorrelation plots, we computed block-averaged autocorrelation curves (decorrelation curves) for each region, which plots the average correlation of pairs of trials as a function of the block distance of those trials, with 0 indicating trials in the same block, and so on (see Methods). These decorrelation curves show that the population response decorrelated in all brain areas (Fig. 3.3B,D) during both the stimulus and baseline period (see Fig. 3.S6 for dACC). Thus, all three regions exhibit some degree of gradual decorrelation in their context representation at the population level. However, the HPC neural population decorrelated (see methods) significantly more rapidly than the dACC and preSMA during both the stimulus and baseline periods (Fig. 3.3E, red vs green and red vs blue, $p_{stim} < 0.001, p_{base} < 0.001$ RankSum). In contrast, in the MFC, decorrelation speed did not differ significantly between dACC and preSMA (Fig. 3.3E, blue vs green, $p_{stim} > 0.05, p_{base} > 0.05$ RankSum). We also repeated the estimation of decorrelation rate separately for blocks in which the task was the categorization or memory task. This revealed that that decorrelation rate was not significantly different between the categorization and memory task contexts for any region during both stimulus and baseline periods (Fig. 3.S6I-N, blue vs red, all $p > 0.05$ RankSum). Thus, while there were systematic differences in the decorrelation rate between HPC and MFC, these differences did not depend on the specific task being performed at any individual point during the experiment.

Did the gradual decorrelation in the hippocampal neural population cause the temporally dynamic context code we found? If so, the task context representation of blocks of the same task (two blocks apart) would be twice as decorrelated as the representation of the opposite task (one block apart). To test whether this was the case, we computed the relative context modulation, which we defined as the average reduction in correlation from block distance 0 to block distance 1 normalized by the decorrelation rate. Relative context modulation is a unitless indicator of the degree to which explicit changes in task context shift the neural representation while accounting for simultaneously occurring decorrelation (see Methods for details). We found that all three areas significantly differed from each other in their relative context modulation, with the preSMA and dACC exhibiting stronger context modulation effect (Fig. 3.3F, green, RCM = 20.0, blue, RCM = 12.8) and the HPC exhibiting the weakest effect (Fig. 3.3F, red, RCM = 3.2). These values indicate that the effect of task context switching is ~3 times greater at driving changes in the HPC neural population as intrinsic decorrelation, whereas in medial frontal cortical structures the task-switching is ~10-20 times stronger. A similar qualitative pattern was observed during the baseline, but with the dACC exhibiting greater relative context modulation than the preSMA (Fig. 3.S6O). Thus, while the decorrelation rate of the hippocampal neural population is high, it exists alongside an even larger task context encoding effect. Decorrelation alone can therefore not explain the cross-temporal context instability we find in our data in the HPC.

These population-level autocorrelation analyses demonstrate that, while on short timescales within a block, all regions show a high degree of self-similarity, the hippocampus decorrelates more

rapidly and is less consistently modulated than the medial frontal cortex by the re-entry into previously encountered task contexts.

The dACC context representation generalizes between stimulus and baseline periods.

The context representations present in HPC, dACC, and preSMA during the baseline and stimulus periods share many properties with respect to their long-range (experiment-level) temporal dynamics, but the relationship between the two remains unclear at the timescale of a single trial (~2 seconds). Jointly tuned context neurons during the stimulus and baseline that retain their task selectivity (e.g. Fig. 4A), would support a context code that generalizes between the two trial periods. However, a plethora of context-tuning properties is present including neurons that are context tuned in one trial period and not the other (e.g. Fig. 3.S1A-D), and neurons that invert their preferred task between trial periods (e.g. Fig. 3.S1E-F). Thus, to clarify the relationship between the baseline and stimulus context codes, we directly compare the two at the population level.

We first compared the fraction of neurons that were jointly context tuned during the baseline and stimulus periods using separate 1-Way ANOVAs for context ($p < 0.05$ significance) in each time period. The fraction of jointly tuned neurons was greater than would be expected by chance in all areas (Fig. 3.4B, purple, HPC = 18.0%, dACC = 19.1%, preSMA = 15.8%, all $p < 0.05$ using Fisher's Exact Test). We next performed cross-trial period generalization analysis using decoders trained for context between the baseline and stimulus. Mean generalization decoding accuracy, reported as an average over baseline-to-stimulus and stimulus-to-baseline generalization, was significantly above chance for all three regions (Fig. 3.4C, $p < 0.001$ against Shuffle Null). When computing the generalization index, we found that the dACC was significantly greater than HPC (Fig. 3.4D, blue vs red, $p < 0.05$, permutation test) and preSMA (Fig. 3.4D, blue vs green, $p < 0.0\,5$, permutation test), indicating that dACC exhibited the most temporally stable context representation within the span of a trial. This finding was further confirmed by the angle analysis performed between the baseline and stimulus context decoders, which revealed that only dACC context coding vectors significantly differed from orthogonal (Fig. 3.4E, blue, $p = 0.0005$, against shuffle null), whereas HPC (Fig. 3.4E, red, $p = 0.02$, against shuffle null) and preSMA (Fig. 3.4E, green, $p = 0.02$, against shuffle null) context coding vectors weakly significantly differ from orthogonality when comparing baseline and stimulus. Together, these findings indicate that, while there may be some shared neural substrate between the baseline and stimulus context representations in all three areas, the dACC uniquely features a context representation that generalizes on the within-trial timescale, a property that separates it electrophysiologically from being grouped with the preSMA. In the hippocampus, the high baseline-stimulus generalization index indicates that the task context representation is largely common between these two time periods.

The hippocampal context representation stabilizes under different experimental conditions.

Is the temporal stability of context representations in the brain an immutable property intrinsic to each region, or can it vary as a function of experimental setting? Our analysis indicates that, within a given experiment, the temporal stability of the context representations does not change across alternating tasks (Fig. 3.S6I-N). Does this trend hold true in the limit of a completely different experiment? We next examined data from a second experiment (experiment 2) that, though structurally similar to experiment 1 (baseline, stimulus periods, blocks of trials, changing context, binary responses to visual stimuli, etc…), featured several key design differences that place processing demands on the hippocampus that differed significantly from experiment 1.

In experiment 2, patients were rewarded for providing correct binary responses to visual stimuli, with the specific stimulus-response associations learned through trial and error. There were two latent contexts, each specified by a different stimulus-response-outcome map, that alternated covertly in a blocked manner, and patients learned to perform inference on the current context through outcome signals provided after every trial (Fig. 3.5A). No explicit context-switching instructions were provided during experiment 2 sessions. Both experiments featured a binary context variable, block structure of similar trial length, a stimulus-identity related variable with four levels, and comparable baseline and stimulus trial periods on which single-neuron and population-level analyses could be performed. These commonalities in the trial-level and block-level structure between experiment 2 and 1 allow for a direct comparison of the encoding strategy employed by the brain at the single-neuron level to represent task-context variables of different kinds that are present in different experimental settings (see Methods, Discussion for detailed description of similarities and differences).

In experiment 2, 17 patients completed 42 sessions (180-320 trials/session, 10-16 blocks/session), with novel stimuli and stimulus-response-outcome maps that needed to be re-learned at the start of each session. Of these, only sessions where patients exhibited a significant behavioral signature of performing inference on the state of the latent context following covert context switches were considered for analysis. We only considered hippocampal neurons here because our prior work shows that latent context is only represented in the hippocampus (and not the MFC) during both the stimulus and baseline periods in this experiment (see [12]). Based on these constraints, 325/499 recorded HPC neurons from 12/17 patients in 19/42 sessions of experiment 2 were included for analysis. HPC neurons exhibited tuning to context during both the baseline (1-Way ANOVA, $p < 0.05$ significance, example in Fig. 3.S8A), and to context and stimulus identity during the stimulus period (2-Way ANOVA with interactions, $p < 0.05$ significance, examples in Fig. 3.S8A,B). The percentage of HPC neurons tuned to context and stimulus identity was not significantly different across the two experiments (Fig. 3.S8C, base context, 10.3 vs 13.2%, $p = 0.32$, stim context, 22.1 vs 16.3%, $p = 0.09$, stimulus identity, 17.2 vs 18.5%, $p = 0.72$, Chi-square test). Furthermore, the average ANOVA F-statistic for tuned neurons to context during the baseline or stimulus period was not significantly different between the two experiments (Fig. 3.S8D , $p_{base}, p_{stim} > 0.05$, RankSum over neurons). These analyses indicate that single hippocampal neurons generally exhibited similar univariate tuning properties across the two matched trial periods in the two experiments considered here.

To compare the population-level context code employed by the hippocampus in the two experiments, balanced decoding analysis was once again performed during the baseline and stimulus periods of both experiments while matching the number of neurons and correct trials per condition across the two experiments. Task context was significantly decodable from the hippocampus in all four conditions (Fig. 3.5D, $p_{base} = 0.009$, $p_{sti} = 1.6x10^{-6}$, $p_{base} = 9.4x10^{-7}$, $p_{stim2} = 3.7x10^{-5}$, using shuffle null distribution). Decodability of context from Exp 1 stimulus, Exp 2 stimulus, and Exp 2 baseline all did not differ significantly from each other ($p_{stim1,base} > 0.05$, $p_{stim\ ,stim2} > 0.05$, $p_{base2,stim} > 0.05$, Permutation Test).

Cross-temporal context decoder generalization (Fig. 3.S8E-H) and subsequent generalization index analysis revealed significantly greater generalization indices in experiment 2 when comparing baseline (Fig. 3.5E, baseline, red vs red, $p < 0.01$, permutation test) and stimulus (Fig. 3.5E, stimulus, blue vs blue $p < 0.01$, permutation test) across experiments respectively. Notably, cross-temporal generalization indices for context during the stimulus period are greater in experiment 2 despite the fact that univariate tuning to context was significantly greater on average at the single-

unit level in experiment 1 (Fig. 3.S8D), and context decoding accuracy did not significantly differ between the two (Fig. 3.5D, stimulus, blue vs blue $p > 0.05$, permutation test). These analyses indicate that, from a decoding standpoint, the hippocampal code for context is significantly more temporally stable across blocks in experiment 2 compared to experiment 1 during both stimulus and baseline periods, and these effects do not arise from a greater number or more strongly univariately context tuned neurons in experiment 2.

The increased cross-temporal stability of the hippocampal context representation in experiment 2 suggests that the self-similarity of the neural representation is increased at longer timescales when compared to experiment 1. This prediction was formally tested by performing population-vector autocorrelation analysis on both experiments and comparing both the decorrelation rate and the relative context modulation as was previously performed in experiment 1. These analyses revealed that the hippocampal decorrelation rate was significantly slower in experiment 2 than in experiment 1 during both the stimulus and baseline periods (Fig. 3.5F, 3.S8I-L). The relative context modulation was also significantly elevated in experiment 2 compared to experiment 1 during both the stimulus (Fig. 3.5G) and baseline (Fig. 3.S8N) periods. Taken together with the cross-temporal decoding analyses, these findings indicate that the hippocampal neural population was significantly more stable, exhibiting less decorrelation over the timescale of the experiment and maintaining a more stable representation of the task context.

Given the decodability of context during both the stimulus and baseline periods of experiment 2, we next compared the format between the two time periods through baseline/stimulus context decoder generalization analyses. We found that the context baseline-stimulus generalization index was significantly greater for the hippocampus in experiment 2 than in experiment 1 (Fig. 3.5G, Exp 1 vs Exp 2, $p = 3.9x10^{-4}$ RankSum), with coding vectors that deviated significantly from orthogonality (90 deg) for the experiment 2 and weakly in experiment 1(Fig. 3.2G, angle vs. chance $p_{Exp\,1} = 0.036, p_{Exp\,2} = 3x10^{-4}$). These findings suggest that the persistent representation of context generalizes across time periods within an individual trial in experiment 2. Thus, taken together, these analyses indicate that the latent context variable in experiment 2 is encoded in in a more temporally stable manner simultaneously at the timescale of a single trial (~2s) and at the timescale of the experiment (~20min) in the hippocampal representation.

Discussion:

The encoding strategies employed by different regions in the human frontal and temporal lobe can vary considerably for even simple, binary cognitive task variables whose representation must be persistently maintained to support behavior on long timescales. Here, we have demonstrated that neural populations in the human medial frontal cortex form a temporally stable representation of instructed task context that persists over many minutes in the absence of re-cuing both during baseline periods and stimulus processing periods (Fig. 3.2). The hippocampus, on the other hand, employs several surprising implementational strategies at the level of single-neurons to encode a representation of task context. We have shown that representations of instructed task context are encoded dynamically in the hippocampus when considering relatively long (~30 minute) time periods of persistent behavior, unlike the temporally static task context representations present in the medial frontal cortex (Fig. 3.2). Dynamic changes in task context encoding occurred rapidly at task boundaries with stable coding within blocks (Fig. 3.S4), and is not an artefact of weakly tuned neurons (Fig. 3.S4). Also it is not a result of recording instability, as the image category is stably encoded across the task in the HPC (Fig. 3.S5). While neurons in all areas were found to exhibit decorrelation over experiment time-scales (Fig. 3.3), the effect was slower in medial frontal cortex, with MFC neurons being statically modulated by task context across blocks (Fig. 3.2, 3.3, 3.S3) and across trial periods (Fig. 3.4), unlike the hippocampus. A temporally dynamic context representation is not an immutable feature of hippocampal neurons, however, as under different experimental conditions, the hippocampal context representation simultaneously stabilized across trial periods and across experimental blocks (Fig. 3.5).

Medial frontal cortical neurons representing task context

Medial frontal cortical structures in the primate brain have long been appreciated for their role in maintaining representations of task context variables that support persistent behaviors. However, previous single-neuron studies in non-human primates frequently provide task context cues on a trial-by-trial basis, thus obviating the need to maintain the instructed variable beyond a single trial. Various computational models have been proposed that account for the static and dynamic codes employed by these frontal cortical neurons, but again only apply to the dynamics for a single trial on the timescale of 1-2 seconds. Here, subjects maintained an instructed task context over many 10's of trials and many minutes without re-cuing. In experiment 1, we demonstrated that the encoding structure for instructed task context shares many commonalities between neurons in the two regions, including cross-temporal stability of the context representation across long experimental periods during both stimulus and baseline, a slow rate of population-level decorrelation over time within-context, and a relatively large degree of neural population modulation with context re-cuing. There are clear computational advantages for a network to employ such a static context representation, most notably the ability of a downstream region to read out the current task context arbitrarily long after the cue has been provided. Thus, the presence of medial frontal cortical context representations that generalizes across arbitrarily long and variable time periods could facilitate the ability of the individual to flexibly maintain persistent behavior accordingly. Various neural network architectures employed over the last decade, including different kinds of recurrent neural networks and transformer-based networks[51,52], famously struggle with generalization to sequence lengths outside of their training distribution, a flaw which could be ameliorated by encouraging the learning of temporally disentangled representations similar to those we observe in the medial frontal cortex. We note that, in experiment 1, even though the block length was predictable, the long block length (40 trials) and the lack of end-of-block anticipatory behavioral effects suggests that patients were not

actively keeping track of their progress through each block, and were prepared to continue responding appropriately beyond 40 trials.

What mechanism could allow for a neural code of task context to be persistently maintained for such long periods of time? Note that assuming a spike train autocorrelation time constant τ_s = 350msec for MFC[53], the ratio to the average block duration B/τ_s, during which the context representation was stable, is on the order of 300 (450 for the control variant) in experiment 1. Various circuit-level mechanisms, including recurrent excitation and short-term synaptic plasticity[54,55], do provide potential explanations for the increased window of temporal integration exhibited by primate frontal cortical neurons, and account for both static and dynamic codes those neurons exhibit. However, these models and analyses are limited to the duration of single-trials in working memory tasks, and representations of task context variables do not need to be maintained for more than 2-3 seconds during those delays. Neurons in some of these models can exhibit long time constants (up to 4 seconds in Area 24)[56], but it is unclear if such models can explain task variable coding activity that persists 2 orders of magnitude longer. Our work here calls attention to the lack of neural circuit-level models that are matched to this intermediate timescale of instructed human behavior and persistent neural activity.

In several cognitive tasks, neurons recorded in the dACC and preSMA of humans have been found to respond similarly to task variables, and are summarily grouped for population-level analysis. Here, one feature in which these two regions strongly diverge is their degree of baseline-stimulus context generalization, with dACC neurons exhibiting considerably greater context coding direction alignment between the two trial periods when compared to preSMA. These findings support the role of the dACC as a temporal storage buffer for context variables that influence behavior in a temporally extended way, since one might expect that a storage buffer would need to stably encode the variable it has buffered to facilitate flexible readout. When specifically compared to the preSMA, the difference in temporal stability of the context code across trial phases could result from intrinsic differences in recurrent excitation and spike train autocorrelation of neurons in these regions, which have been observed to be longer in the Anterior Cingulate Cortex when compared to more caudal frontal cortical regions[53,55,57]. The lack of baseline-stimulus context generalization could also arise from fundamental differences in circuit-level computation in the preSMA that lead to stronger non-linear interactions between persistently maintained context variables and incoming stimulus information, as choice signals were observed to be more prominent in the preSMA than dACC in our previous work[6]. More extensive psychophysical experimentation and widespread frontal cortical recording is needed to provide answers for such questions.

<u>The logic underlying static and dynamic hippocampal codes for task variables</u>

The cognitive map formed by hippocampal neurons has been extensively studied for its encoding of a wide variety of variables in support of flexible behavior. Here, we considered variables that can be split into two different categories: task context variables, which determine the appropriate response for many different stimuli, and stimulus variables, which encode identity or category information about a currently presented visual stimulus. The task context variables in these experiments also differed from stimulus variables in that their value was not re-cued trial by trial, and needed to be remembered for many minutes at a time in order to complete both experiments. We mainly focused our analysis on the hippocampal representation of task context variables, and found that the code for task context in experiment 1 was dynamic, orthogonalizing over time as patients transitioned between task contexts across blocks. This feature of the task context code stands

in contrast to the temporally static codes observed in MFC, and despite the fact task context was significantly decodable in all three regions.

Given the computational benefits of a temporally static representation discussed in the previous section, why would the hippocampus employ such a different code for task context? Several non-mutually exclusive potential explanations exist. First, it is possible that the presence of static MFC task context codes could obviate the need for temporal stability in the hippocampal context representation. If a static code persists elsewhere, the hippocampus is free to return to its "default" state of internally generated cell assembly sequences, which do not cross-temporally generalize[43]. This explanation could also account for the increased cross-temporal stability in the hippocampal context code observed in experiment 2, in which frontal cortical context representations were largely absent[12]. A related explanation pertains to the task context explicitly being signaled in experiment 1, whereas in experiment 2 context was a latent variable whose state was inferred through feedback. The hippocampus is known to support various kinds of inference behaviors in animals[11,58–60], and bilateral temporal lobectomy patients are unable to perform tasks with inferred rules such as the Wisconsin Card Sorting Task[61], which is similar to experiment 2. Clearly, these patients are nonetheless able to encode and follow language-based instructions similarly to experiment 1. Thus, the hippocampal context representation may stabilize across time specifically when it is needed to support persistent behavior, i.e. when the task context variables are latent and must be inferred and are not persistently encoded elsewhere in the brain.

The hippocampus plays a prominent role in memory formation, and the presence of episodic memory demands in experiment 1 could create a demand for the hippocampus to encode the passage of time, i.e. the current temporal context, alongside the instructed contexts in experiment 1. In experiment 1, the activity of single hippocampal neurons was better accounted for by block-specific tuning, and as a population the neurons exhibited continual decorrelation on the timescale of minutes that was comparable in magnitude to the effect of switching tasks between blocks (hippocampal RCM = 3.2). The encoding of time in the human hippocampus is achieved at the single-neuron level through time cells and ramp cells[42], and at the population level by sequential firing of these neurons so that the passage of an interval of time can be decoded from these neurons by a downstream readout. Furthermore, our group has previously demonstrated that temporal context reinstatement effects are present in hippocampal neurons[47]. Here, the hippocampus multiplexed temporal context information with task context information such that both variables were simultaneously decodable from the same neural population, possibly reflecting the association of multiple behaviorally relevant high-level context variables[62]. Such temporal context encoding may have been absent from experiment 2 due to the small number of stimulus-response associations to be remembered and the lack of behavioral demand to perform episodic recall. Unfortunately, the data and experiments here are unable to arbitrate between the above potential explanations for the differences in task context representations across the three areas, and new experiments are needed to address these points as well as other potential confounding factors (see Experimental Limitations). Nevertheless, characterizing the differences in task context encoding strategy employed by the hippocampus across different experiments and when compared to MFC is an important first step in characterizing unique properties of the hippocampal cognitive map.

What strategies are employed by the hippocampus for organizing the simultaneous representation of task context variables and stimulus variables in the same neural state space? Evidence from modern systems neuroscience points towards the hippocampus simultaneously exhibiting high and low dimensional properties in its state space representation of the environment, where combinations of these variables are mixed to varying degrees, allowing for flexible readout

for many downstream tasks while retaining some advantageous geometric properties that allow for generalization of one variable across others[11,12,63,64]. Here, we have demonstrated in two different experiments that hippocampal representations of stimulus variables are orthogonally encoded with respect to the blocked task context variables and the passage of time. Could the segregation of stimulus and context variables into orthogonal subspaces be a general feature of hippocampal representations? Place-field-like coding across conjunctions of variables is frequently seen in the hippocampal representations of rodents and non-human primates, arguing against this segregation as a general property of the hippocampus[65–67]. However, this may strictly be a property of human hippocampal representations, and the disentangling of stimulus and context codes may underlie the rapid learning and generalization behaviors exhibited by humans when compared to other species[11,12,68].

Experimental Limitations

Several features of the experiments and data used in this study limit our ability to draw conclusions about the cause of observed differences in neural activity, particularly in the hippocampus. The comparison between the two experiments shown here relies on similarities in high-level structure (i.e. same number of stimulus categories, blocked contexts, trial structure, etc…), but there are many salient differences between the two that could have contributed to the differences in task context representation that we observed. The first difference is the increased block length (40 trials/block, 8 blocks) and fully predictable context switches in experiment 1 compared to shorter blocks and more frequent, un-predictable switches in experiment 2 (15-32 trials/block, 10-16 blocks). More frequent, less predictable switches in environmental context variables could encourage a disentangled, or compositional hippocampal task representation, including disentangling from the passage of time throughout the experiment. The second difference relates to the language-based prompting of task contexts in experiment 1 and not experiment 2. Language-based instructions are invariant by construction, and support the emergence of systematically structured frontal cortical BOLD fMRI responses that have been studied extensively[69,70]. It is possible that the presence of such structured representations in the cortex obviates the need for forming such structured representations in the hippocampus for cortex to subsequently read out, which would not be possible in experiment 2 since the two contexts in that case have no associated language-based prompts. Third, the need to perform inference on the current state of the context through feedback in experiment 2, a behavioral process for which the hippocampus is thought to be necessary, might also encourage context stabilization across time. The fourth difference is the behavioral demand for episodic recall in experiment 1 and not in experiment 2. The behavioral pressure to reinstate many distinct old images could have encouraged the tracking and mixing of temporal context with the task context representation in experiment 1, unlike experiment 2 where memory of stimulus-response associations was needed, but not episodic recall per se. Fifth, the continual presentation of new images in addition to old images in experiment 1 could also have influenced the decorrelation rate, as each task block is more episodically distinct by virtue of encountering novel stimuli, whereas in experiment 2 the same four stimuli were used throughout each session. A sixth consideration for the data analyzed here is that the patients groups that completed the two experiments were strictly non-overlapping, with recordings for experiment 2 commencing several months after those for experiment 1 had concluded. Even though electrodes were implanted in the same general location of the anterior hippocampus using the same modern standard-of-care surgical procedure, individual differences between the two patient groups could still correlate with the mean difference in hippocampal neuronal activity observed between the two experiments.

We were also unable to perform temporal context reinstatement analysis in experiment 1 for two reasons. First, the lack of patient confidence ratings and the repeated presentation of "old" stimuli. In previous work[47], we utilized trial-level confidence rating to disambiguate memory task trials that were solved through episodic recollection instead of familiarity, which could also be used to complete the memory task without episodic recall. Here, patients did not give trial-level confidence ratings, preventing us using this approach to identify instances where the hippocampal temporal context might have been reinstated. Second, the repeated presentation of "old" images, once each in both the categorization and memory blocks, creates an ambiguity as to which episode of previous presentation is being recalled in the event of episodic recall. For example, during block 8, a patient might be recalling their having seen the image in block 7, or in block 1, and different patients might recall presentations of the same image from different blocks when answering a given trial. Thus, both single-patient and pseudopopulation approaches to temporal context reinstatement analyses are confounded here.

Conclusion and Future Directions

In this work, we leveraged the existence of large, pre-existing human single-neuron datasets recorded from neurosurgical patients performing different, but structurally well-aligned, psychophysical experiments to study differences in task variable encoding employed by the hippocampus, dACC, and preSMA. Our findings indicate that, in experiment 1, the same high-level task context variable can vary considerably in its encoding structure over time depending on the region being considered. In experiment 2, the hippocampus was the sole region to explicitly encode a latent context variable, with the representation of that variable becoming stabilized at multiple timescales. Of course, these findings weakly specify a small fraction of logic that governs the encoding strategies used by these regions for different context variables in complex environments with changing task demands. Experiments that require simultaneous encoding of multiple task context variables, all of which contribute equally to instantaneous task demands, and that vary in their re-cuing rate, predictability, and need for episodic recall, are needed to further clarify the principles used by the hippocampus and the frontal cortex to encode task context variables in different conditions.

Figures:

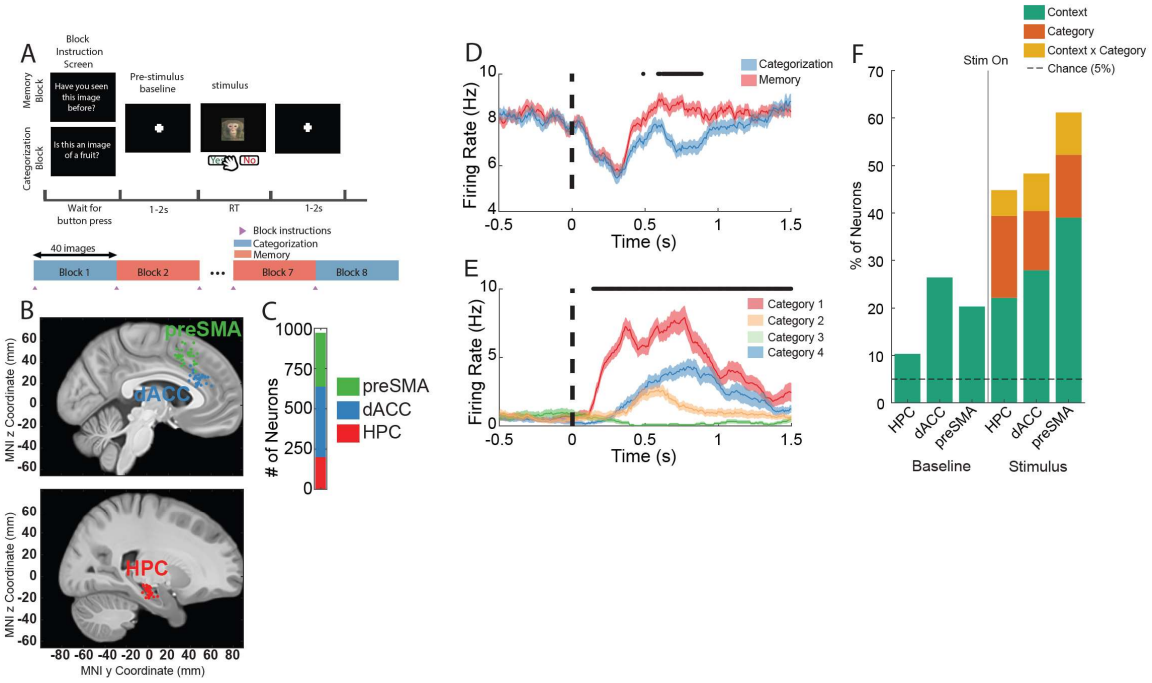

Figure 3.1. Single neurons are tuned to task variables during instructed task switching.
(A) The first experiment consisted of eight blocks of 40 trials where the task context alternated between a categorization task and a recognition memory task. Text-based instructions were provided to patients only once at the start of each block but applied for all trials until the next set of instructions. Tasks were formulated as yes-no questions in response to presented stimuli. Trials consisted of a pre-stimulus baseline with a central fixation cross, followed by the presentation of a single stimulus (image) to which the patient would respond yes or no according to the current task in a speeded manner. Following the response, the stimulus was removed from the screen, and the next trial would commence after a jittered delay (1-2s). No trial-by-trial performance feedback was provided.
(B) Electrode locations for the pre-Supplementary Motor Area (preSMA, green), dorsal Anterior Cingulate Cortex (dACC, blue), and anterior hippocampus (HPC, red). Each dot corresponds to the implant site of a microwire bundle for a single patient. All implants were bilateral and electrodes are shown on the same hemisphere for visualization purposes.
(C) Number of single units recorded across the three brain areas (970 neurons total, dACC, blue = 329, preSMA, green = 438, HPC, red = 203).
(D-E) Example PSTHs of neurons recorded in HPC that are differentially selective for task context **(D)** and semantic category of the visual stimulus **(E)** during stimulus presentation throughout the task. Stimulus onset occurs at time 0. Black points above the PSTH indicate times where a sliding-window 1-way ANOVA (250 msec width) over the considered task variables was significant ($p < 0.05$).
(F) Percentage of neurons that exhibit tuning to task variables during the Baseline (-1 to 0s prior to stimulus onset) and Stimulus periods (0.2 to 1.2 following stimulus onset). Neurons are considered tuned during the stimulus period to either a main effect (context – green, category – orange), or the interaction (orange) if the associated factor in a 2x4 ANOVA (Context x Category) was significant ($p<0.05$). Baseline tuning is limited to a 1-way ANOVA for context since the visual stimulus is not yet present. Horizontal dashed line indicates chance level. Vertical line marks the boundary between Baseline and Stimulus.

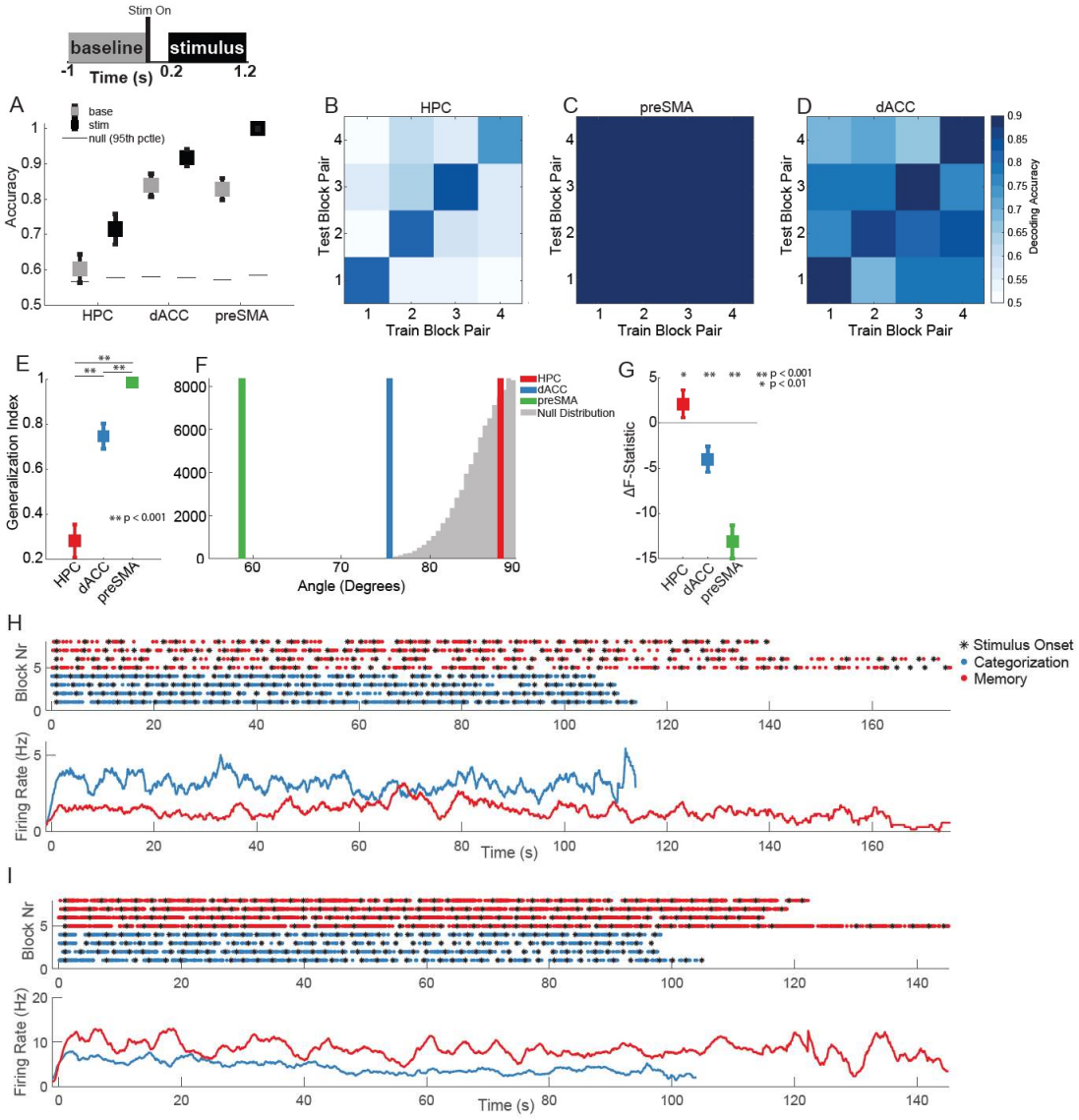

Figure 3.2. Task context representation temporally generalizes in MFC, not in Hippocampus
(A) Context decoding accuracy during the baseline (gray) and stimulus (black) periods (see inset) using correct trials from the entire experiment. Black horizontal lines indicate 95th percentile of null distribution. Chance decoding accuracy is 0.5 (two contexts).
(B-D) Cross-temporal decoding plots indicating the out-of-sample decoding accuracy of decoders trained to decode context from correct trials in adjacent block pairs during the stimulus period. X-axis indicates which block pairs are used to train the context decoder, and y-axis indicates the block pairs on which the decoder is evaluated. On-diagonal decoding accuracies (train/test on same block pair) are reported with 5-fold cross validation. Off-diagonal decoding accuracies use all available trials for training and testing. The colormap shown for dACC **(D)** also applies for HPC **(B)** and preSMA **(C)**.
(E) Generalization index (see methods) computed for the cross-temporal generalization of context decoding across block pairs. Index values range from 0 to 1, indicating no generalization and perfect generalization of context coding respectively. P-values are computed using Wilcoxon Rank-sum test.

(F) Angles computed between vectors normal to the hyperplanes of the block-pair context decoders. Angles here were estimated in an n = 150-dimensional space to facilitate direct comparison of angle values between regions. The mean angle averaged over all pairs of decoders is reported for HPC (red), dACC (blue), and preSMA (green). The null distribution (gray) is populated by the angle between randomly selected pairs of trial-shuffled context decoders.
(G) Comparison of ANOVA F-statistics fit using block number and task context in single-neurons. Values are reported are mean ± s.e.m. ΔF-statistic computed over neurons. P-values are computed using a two-sided t-test.
(H) Example raster (above) and PSTH (below) for a neuron in the dACC that exhibited persistent firing rate context modulation throughout entire blocks. An individual row in the raster (above) corresponds to the activity of a single neuron plotted for a block. Each point corresponds to one spike discharged by the neuron. Black stars indicate stimulus onset times. Blocks are re-ordered according to task context (categorization = blue, memory = red), and are aligned to the stimulus onset time of the first trial in each block. PSTH (below) shows mean firing rate computed over blocks.
(I) Same as **(H)**, but for a neuron in preSMA.
Note: All instances of plots with squares and error bars indicate mean ± s.e.m. of the computed metric (e.g. decoding accuracy, generalization index, etc…) over 250 iterations of bootstrapped re-sampling unless otherwise specified (see methods). Null distributions were also computed with 250 iterations of trial-label shuffling followed by re-computing the metric in question.

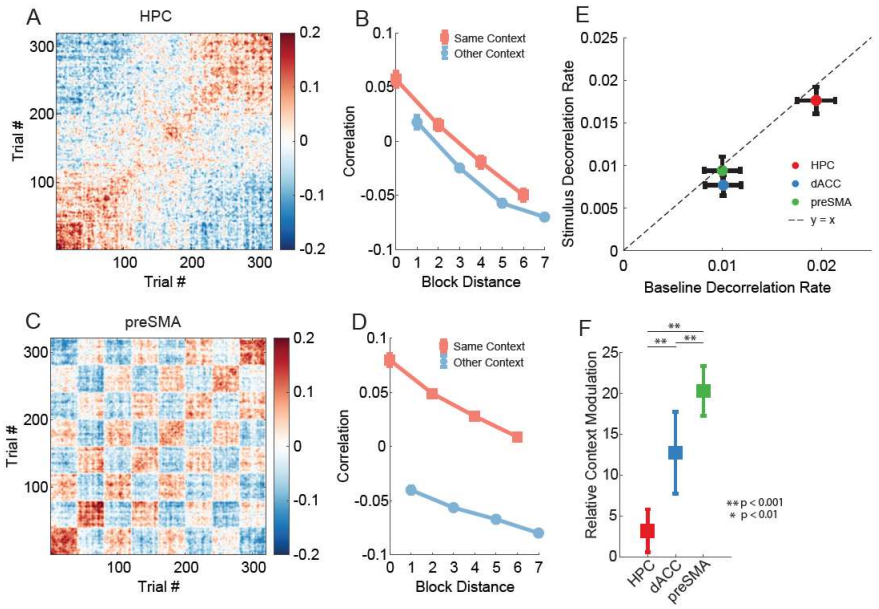

Figure 3.3. Hippocampal neural population exhibits temporal decorrelation.

(A) Hippocampal population-vector autocorrelation matrix showing Pearson correlation for all possible pairs of trials. Correlations are computed between pseudo-population firing rate vectors computed for each trial. Diagonal values are removed for visualization purposes. Matrix shown here is an average over 250 iterations of sub-sampled estimation to match the number of neurons between regions.

(B) Mean cross-block correlation computed over all possible pairs of trials that are in increasingly distant blocks for hippocampal neurons. For example, Block Distance 0 reports the average population-vector correlation between all pairs of trials in the same block, Block Distance 1 reports the average correlation between all pairs of trials exactly one block apart, etc. Even block distances correspond to blocks of the same task (light red), odd block distances correspond to blocks of the opposite task (light blue). Values are reported as mean ± s.e.m. over trial pairs.

(C,D) Same as (**A,B**), but for pre-SMA.

(E) Baseline vs Stimulus population decorrelation rate plotted for each of the three regions. Decorrelation rate is estimated as the absolute value of the slope of the least-square fit to the cross-block decorrelation curves, e.g. shown in **(B)** and **(D)**. All slopes were negative, so increasing values indicate increasing rate of decorrelation with block distance. Circles and error bars correspond to mean and s.e.m. decorrelation rate computed over iterations of neuron sub-sampling.

(F) Relative context modulation (cross-context correlation difference normalized by decorrelation rate) reported for the three areas during the stimulus period. Values are reported as mean ± s.e.m. over iterations of decorrelation curve estimation. P-values are computed by permutation test.

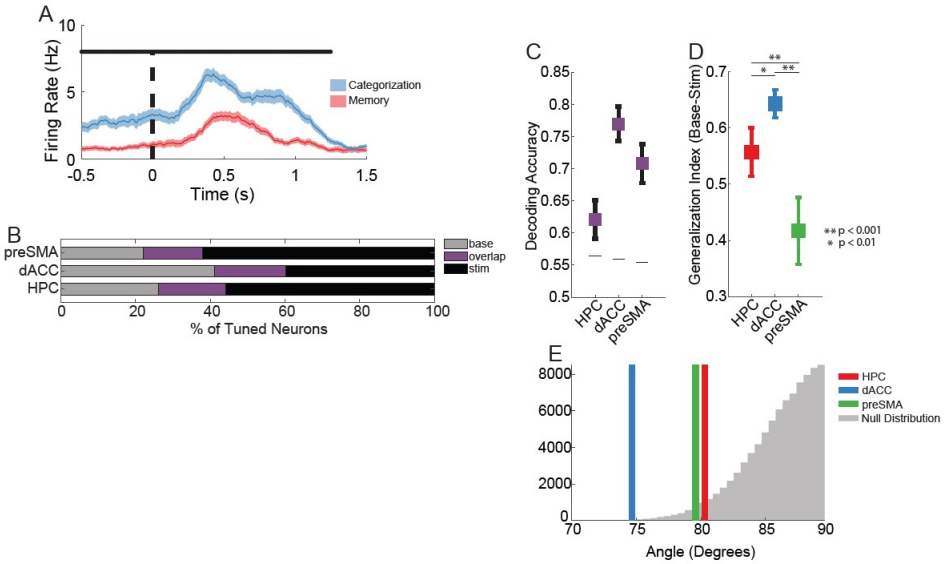

Figure 3.4. Task context representations generalize between baseline and stimulus periods.
(A) Example PSTH of a neuron in dACC that exhibits context tuning during the baseline period (Time < 0s) and during the stimulus period (Time > 0s). Vertical dashed line indicates stimulus onset. Blue and red curves indicate mean ± s.e.m. firing rate over categorization and memory task trials respectively. Black horizontal line indicates time period where firing rate significantly differs between contexts (1-Way ANOVA, $p<0.05$).
(B) Fraction of context-tuned neurons determined using 1-Way ANOVA for context during the baseline (base) and stimulus (stim) periods. Neurons are considered context tuned if $p<0.05$ for either base or stim. Neurons are either context modulated only during baseline (gray), only during the stimulus period (black), or during both (overlap, purple).
(C) Decoding accuracy averaged over both baseline trained/stimulus tested and stimulus trained/baseline tested context decoders. Error bars indicate s.e.m. as previously described. Horizontal black lines are 95th percentile of null distribution.
(D) Generalization index for baseline-stimulus context generalization. Values here are computed using the baseline/stimulus context generalization shown in **(C)**, and the within-stimulus and within-baseline context decoding accuracy reported in Fig. 3.2A. Values reported are mean ± s.e.m. generalization index computed for Hippocampus (red), dorsal Anterior Cingulate Cortex (blue), and pre-Supplementary Motor Area (preSMA). P-values are computed using Wilcoxon Rank-sum test.
(E) Angles between the baseline and stimulus context-decoding hyperplanes. Plotting conventions identical to Fig. 3.2C.

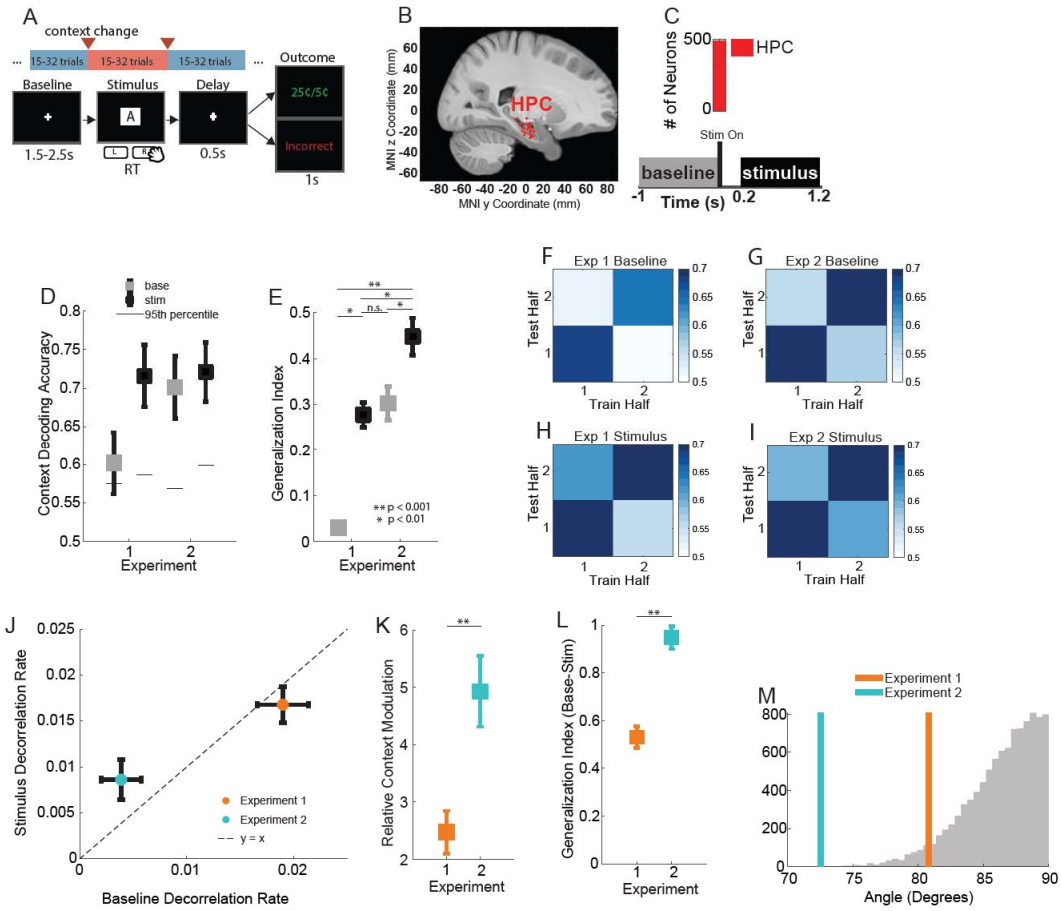

Figure 3.5. Hippocampal context representation temporally stabilizes when context is latent.
(A) Experiment 2 consisted of blocks of 15-32 trials where a latent context variable was specified by arbitrary, deterministic stimulus-response-outcome associations. Trials consisted of a pre-stimulus baseline with a central fixation cross, followed by the presentation of a single stimulus (image) to which the patient would respond with a "Left" or "Right" button press according to the current stimulus and context in a speeded manner. Changes in context were covert, but could be inferred from feedback provided during the "outcome" or feedback screen of every trial. Following feedback, the next trial would commence after a jittered delay (1.5-2.5s).
(B) Electrode locations for the anterior hippocampus (HPC, red). Plotting conventions identical to Fig. 3.1B.
(C) Number of single units recorded in the anterior Hippocampus (HPC, red = 499 neurons).
(D) Context decoding accuracy from HPC during the baseline (gray) and stimulus (black) periods (see inset) using correct trials from Experiment 1 (left) and Experiment 2 (right). Black horizontal lines indicate 95th percentile of null distribution. Chance decoding accuracy is 0.5 (two contexts). Values are reported as mean ± s.e.m.
(E) Cross-temporal context generalization index reported for Experiment 1 (left) and Experiment 2 (right). Values reported are mean ± s.e.m. generalization index computed for the baseline (red) and stimulus (blue) periods of each task. P-values are computed by permutation test.
Cross-temporal decoding plots for task context computed across experimental halves instead of across block-pairs are shown for during the baseline **(F,G)** and stimulus **(H,I)** periods for the two experiments.

(J) Baseline vs Stimulus population decorrelation rate reported for Experiment 1 (orange) and Experiment 2 (teal). Values are reported as mean ± s.e.m. in each dimension. Dashed line indicates y=x.
(K) Relative context modulation reported for the hippocampus during the stimulus period of experiments 1 and 2. Values are reported as mean ± s.e.m. over iterations of decorrelation curve estimation. P-value is computed by permutation test.
(L) Baseline-stimulus context generalization index reported for Experiment 1 (orange) and Experiment 2 (teal). Values are reported as mean ± s.e.m. P-value is computed by permutation test.
(M) Angles between the baseline and stimulus context-decoding hyperplanes for Experiment 1 (orange) and Experiment 2 (teal). Plotting conventions identical to Fig. 3.2C.

Supplementary Figures:

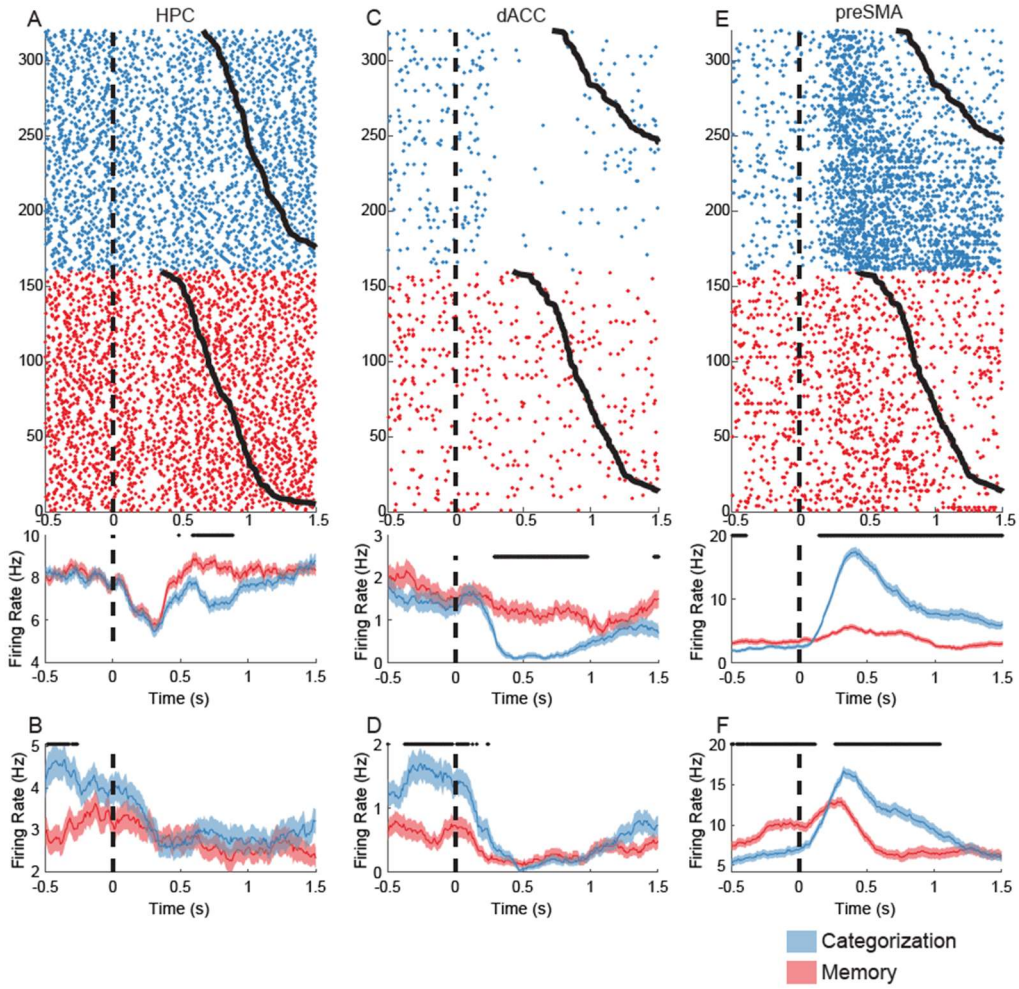

Figure 3.S1. Example neurons recorded in Task 1 that exhibit context tuning during stimulus and baseline periods.
(A) Example raster (above) and PSTH (below) for a neuron in anterior hippocampus (HPC) that was context-tuned during the stimulus presentation period. Trials in the raster are re-ordered according to task context (categorization = blue, memory = red), and are sorted according to reaction time therein, as indicated by the black curves on the right. Vertical dashed line denotes stimulus onset. PSTH shows mean ± s.e.m. firing rate computed over trials. The black dots above the plot indicate time periods where firing rate significantly differs between contexts (1-Way ANOVA, $p<0.05$).
(B) PSTH shown for a different neuron in HPC that exhibited significant context tuning during the baseline period prior to stimulus onset (i.e. to the left of the vertical dashed line).
(C,D) Same as **(A,B)**, but for dorsal Anterior Cingulate Cortex (dACC).
(E,F) Same as **(C,D)**, but for pre-Supplementary Motor Area (preSMA).

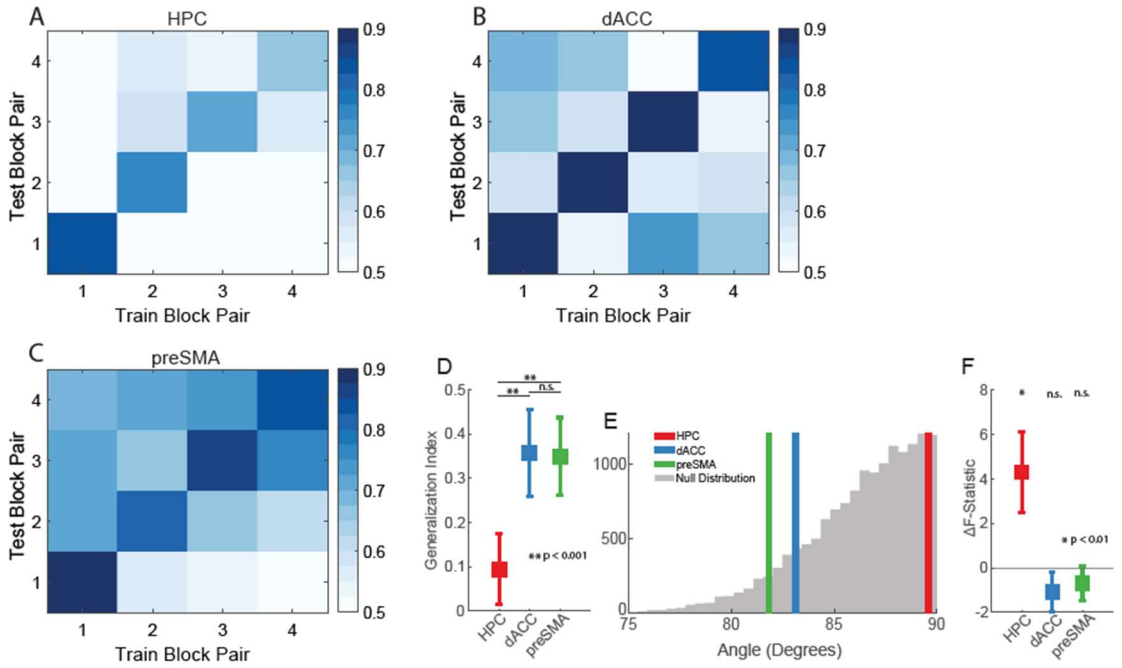

Figure 3.S2. Temporal generalization of task context representation during the baseline period.

(A-C) Cross-temporal decoding plots for task context computed during the baseline period (-1s to 0s prior to stimulus onset). X-axis indicates which block pairs are used to train the context decoder, and y-axis indicates the block pairs on which the decoder is evaluated. Plots are shown for HPC **(A)**, dACC **(B)**, preSMA **(C)**. Plotting conventions identical to those in Fig. 3.2B-D.

(D) Generalization index computed during the baseline for the cross-temporal generalization of context decoding across block pairs. Plotting conventions identical to those in Fig. 3.2E. P-values are computed by permutation test.

(E) Angles computed between vectors normal to the hyperplanes of the baseline block-pair context decoders. Plotting conventions identical to those in Fig. 3.2F.

(F) Single-unit model comparison of ANOVA F-statistics for block number vs task context. Plotting conventions identical to those in Fig. 3.2G.

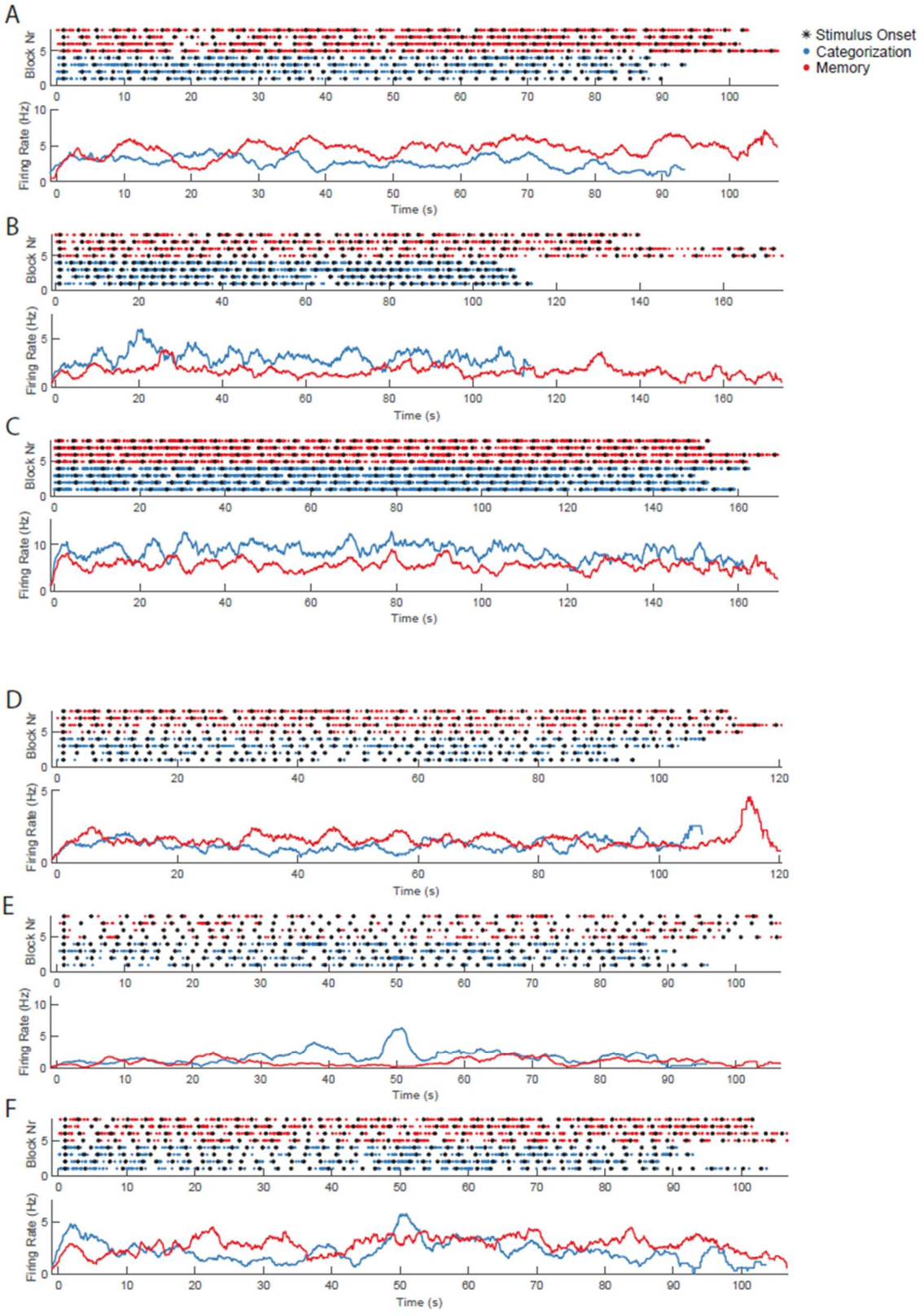

Figure 3.S3. Single-unit rasters/PSTHs showing persistent activity over entire blocks.

(A) Example raster (above) and PSTH (below) for a neuron in the dorsal Anterior Cingulate Cortex (dACC) that exhibited persistent firing rate context modulation throughout entire blocks. An individual row in the raster (above) corresponds to the activity of a single neuron plotted for a block. Each point corresponds to one spike. Black stars indicate stimulus onset times. Blocks are re-ordered according to task context (categorization = blue, memory = red), and are aligned to the stimulus onset time of the first trial in each block. PSTH (below) shows mean firing rate computed over blocks. Since block durations differ, due to randomization of inter-trial intervals and variability in patient responses, a blocks ceases to contribute to PSTH after the final spike in that block is discharged.
(B-C) Same as **(A)**, but for pre-SMA.
(D-F) Same as **(A)**, but for HPC.

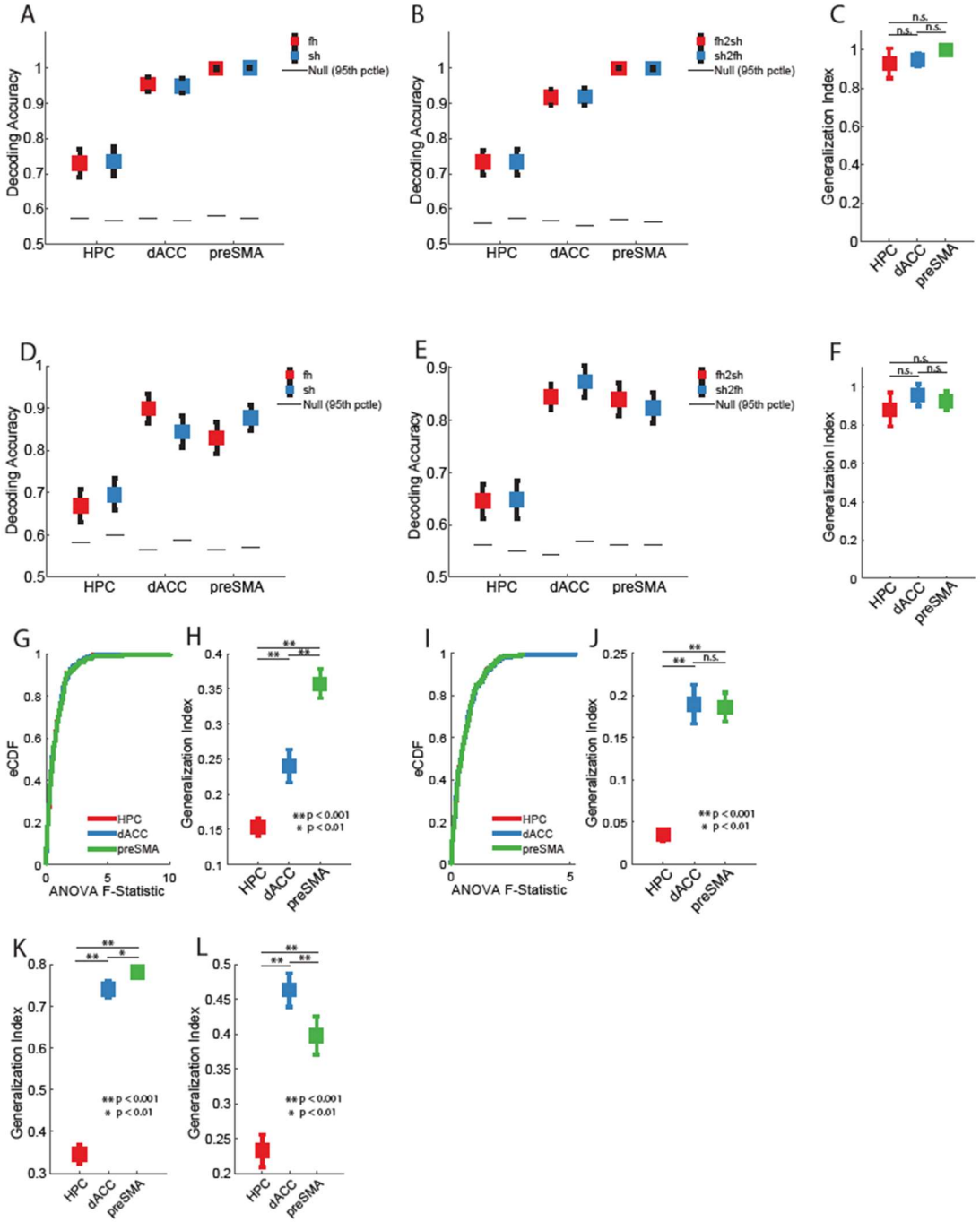

Figure 3.S4. Control analyses for temporal generalization of context representation.
(A) Task context decoding during the stimulus processing period using the first-half (fh, red) and second-half (sh, blue) of every block pair to demonstrate within-block context decoding stability. Plots show mean decoding accuracy ± s.e.m. over bootstrap iterations. Horizontal black lines indicate 95th percentile of shuffle null.
(B) Same as **(A)**, but for generalization decoding accuracy of the context decoder from the first block half to the second block half (fh2sh, red) and from the second block half to the first block half (sh2fh, blue).

(C) Generalization index computed for context decoding across block-halves. Plots show mean ± s.e.m. block-half generalization index computed over bootstrap iterations for HPC (red), dACC (blue), and pre-SMA (green).
(D-F) Same as **(A-C)**, but for context decoders trained and tested during the baseline period.

(G-J) Cross-temporal generalization analysis control with ANOVA F-Statistic distribution matching between regions to ensure that increased temporal stability is not simply a consequence of stronger univariate context tuning at the single-unit level. eCDF of single-unit ANOVA F-statistics for each area are shown during the stimulus **(G)** and baseline **(I)** periods after performing distribution matching. Note: eCDFs for different regions are not clearly visible on the plots since they are practically identical after distribution matching. Cross-temporal generalization indices for the stimulus **(H)** and baseline **(J)** are recomputed using the matched distributions and presented as mean ± s.e.m. over bootstrap iterations for HPC (red), dACC (blue), and pre-SMA (green).

(K-L) Cross-temporal generalization index for context computed on the control task variant where the trial response was given after a fixed delay instead of in a speeded manner. Plots show mean ± s.e.m. cross-temporal generalization index computed over bootstrap iterations for HPC (red), dACC (blue), and pre-SMA (green). Analysis is shown for both the stimulus period **(K)** and the baseline period **(L)**.
Note: all p-values reported in this figure are computed by permutation test.

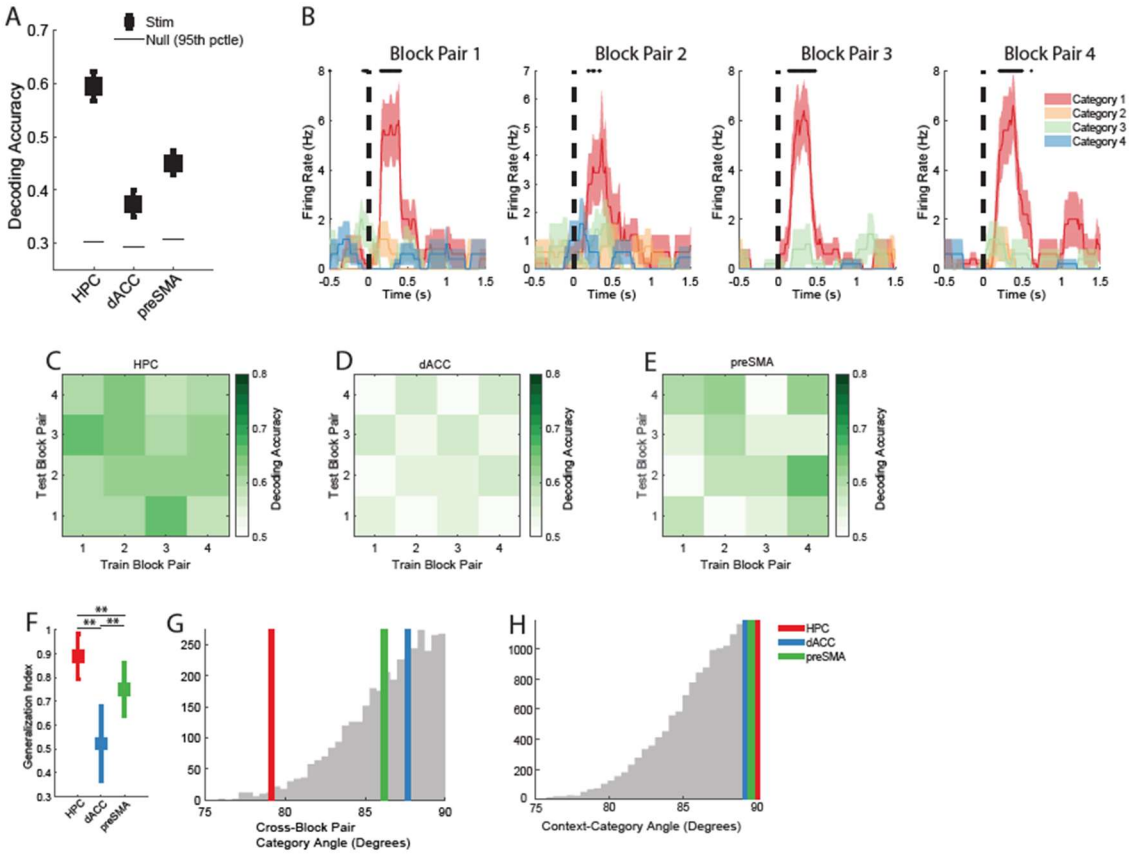

Figure 3.S5. Hippocampal stimulus representation generalizes across time.
(A) Decoding accuracy for semantic category of the presented stimulus during the stimulus period (0.2s to 1.2s following stimulus onset). Chance is 25% (4 categories). Plot shows mean decoding accuracy ± s.e.m. computed over bootstrap iterations. Horizontal black lines indicate 95th percentile of shuffle null.
(B) Example PSTHs for a single neuron exhibiting stable category selectivity (Category 1 preferred) plotted separately for every block pair in the experiment. All plotting conventions identical to those used for PSTHs in Fig. 3.S1.
(C-E) Cross-temporal decoding plots indicating decoding accuracy for decoders trained to decode image category from correct trials in adjacent block pairs during the stimulus period. All plotting conventions are identical to those in Fig. 3.2D. Cross-temporal decoding of image category is reported for HPC **(C)**, dACC **(D)**, and preSMA **(E)**.
(F) Cross-temporal generalization index for image category computed using decoding accuracies reported in **(C-E)**. Plot shows mean decoding accuracy ± s.e.m. computed over bootstrap iterations for HPC (red), dACC (blue), and preSMA (red). P-values are computed by permutation test.
(G) Angles computed between vectors normal to the hyperplanes of image category decoders for different block pairs. All plotting conventions are identical to those used in Fig. 3.2C.
(H) Angles computed between vectors normal to the hyperplanes of image category decoders and the stimulus period context decoders for each region. All plotting conventions are identical to those used in Fig. 3.2C.

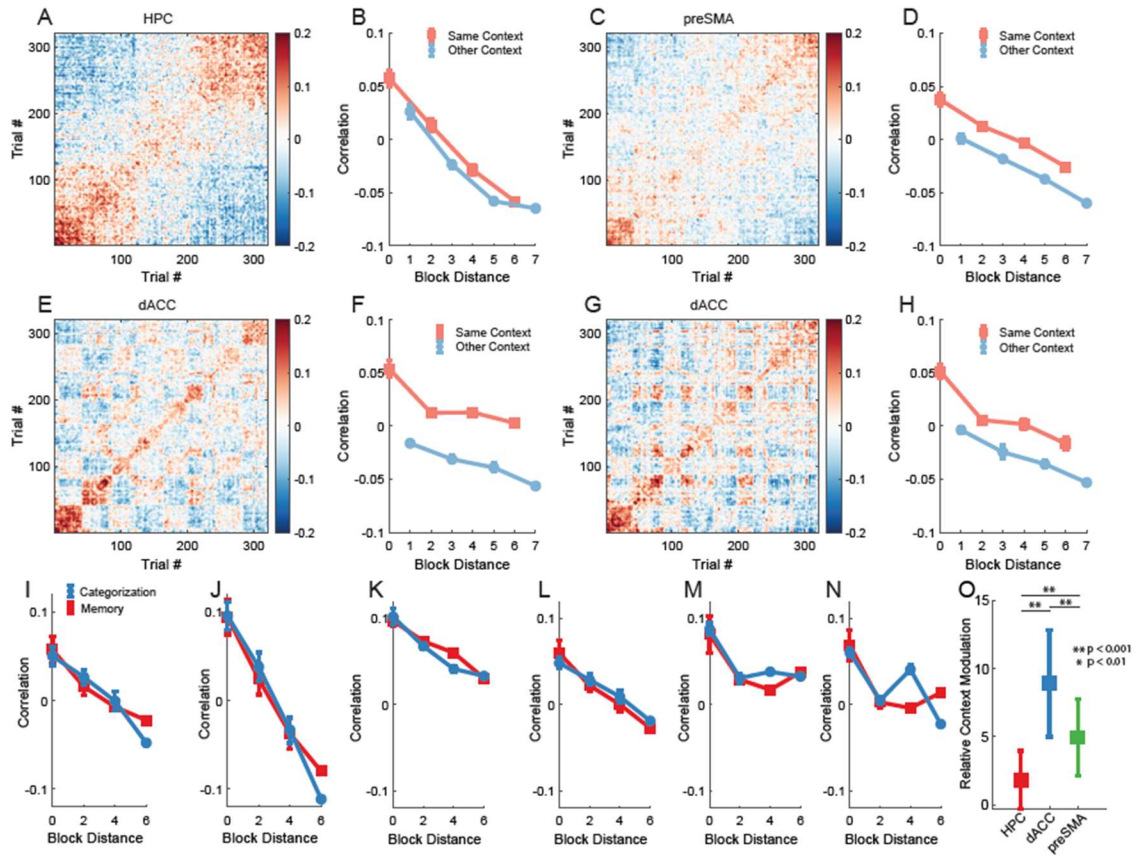

Figure 3.S6. Additional population-vector autocorrelation analyses.
(A-H) Trial-wise population vector autocorrelation plots and cross-block correlation curves shown for HPC during baseline **(A,B)**, preSMA during baseline **(C,D)**, and for the dACC during both stimulus **(E,F)** and baseline **(G,H)** periods. All plotting conventions identical to those used in Fig. 3.3.
(I-N) Cross-block correlation curves reported separately for the categorization task (red) and the memory task (blue). Note in this case, since tasks always alternate, only even block distances can be computed and reported since there is no task block that is an odd number of blocks away from a block of the same task. All other plotting conventions identical to those used in Fig. 3.3B,D.
Task-specific cross-block decorrelation curves are shown for the stimulus and baseline periods respectively in HPC **(I, J)**, preSMA **(K, L)**, and dACC **(M, N)**.
(O) Relative context modulation reported for the three areas during the baseline period. Values are reported as mean ± s.e.m. over iterations of decorrelation curve estimation. P-values are computed by permutation test.

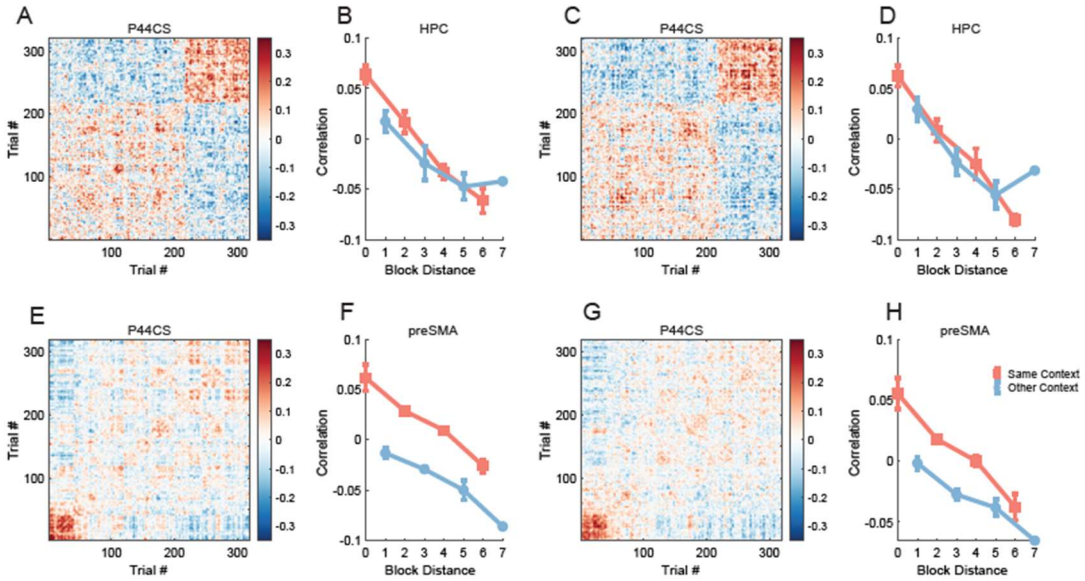

Figure 3.S7. Single-subject recapitulation of temporal decorrelation effect.
Recapitulation of area-dependent temporal decorrelation effect in neurons recorded in a single subject (P44CS). Trial-wise population vector autocorrelation plots and cross-block correlation curves are shown for HPC during stimulus **(A,B)** and baseline **(C,D)** periods and for preSMA during stimulus **(E,F)** and baseline **(G,H)** periods. All plotting conventions identical to those used in Fig. 3.3.

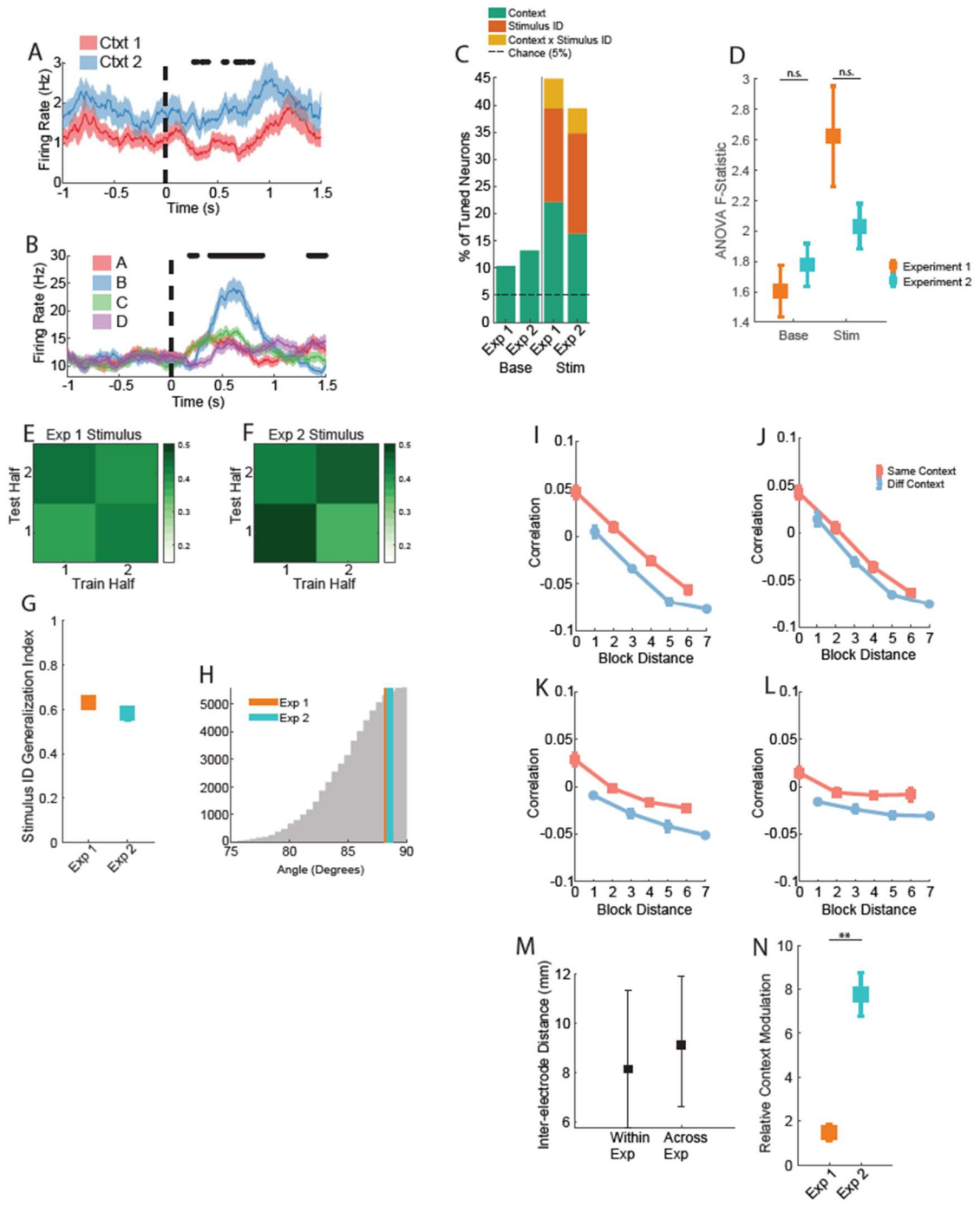

Figure 3.S8. Single-unit properties and cross-temporal context decoding for Experiment 2.
(A-B) Example PSTHs of two hippocampal neurons recorded from patients performing experiment 2. Neurons were modulated by the latent context variable **(A)** and by the identity of the stimulus presented on the screen **(B)**. Plotting conventions identical to those used in Fig. 3.S1.
(C) Percentage of hippocampal neurons that exhibit tuning to task variables during the Baseline (base, -1 to 0s prior to stimulus onset) and Stimulus periods (stim, 0.2 to 1.2 following stimulus onset). "Context" and "Stimulus" correspond to the task context variable and stimulus variable as applicable to each of the two experiments. 2-way (context x stimulus) ANOVAs are performed on

firing rates for single neurons, where context and stimulus correspond to 2-level and 4-level categorical regressors respectively in both experiments. All other plotting conventions identical to those used in Fig. 3.1F.

(D) ANOVA F-statistics for task context main effects shown for single neurons recorded in experiment 1 (left) and experiment 2 (right) during the baseline (red) and stimulus (blue) periods. Reported values are mean F-Statistic ± s.e.m. computed over neurons. P-values are computed by permutation test, and n.s. indicates $p > 0.05$.

Cross-temporal decoding plots for image category in experiment 1 **(E)** and stimulus identity in experiment 2 **(F)** across experimental halves are shown for during the stimulus period.

(G) Cross-temporal generalization index for the image category decoders reported for Experiment 1 (left) and stimulus identity decoders reported for Experiment 2 (right). Values reported are mean ± s.e.m. generalization index computed for the baseline (red) and stimulus (blue) periods of each task. P-values are computed by permutation test.

(H) Angles computed between vectors normal to the hyperplanes of the image category decoder and the stimulus period context decoder in experiment 1 (orange), and of the stimulus ID decoder and the stimulus period context decoder in experiment 2 (teal). All plotting conventions are identical to those used in Fig. 3.2C.

Cross-block correlation curves computed during the baseline **(I)** and stimulus **(J)** periods for experiment 1. Plots here are computed using the same data as those shown in Fig. 3.**3B** and **S5A**.

(K,L) Same as **(I,J)** but for Experiment 2.

(M) Distribution of all pair-wise inter-electro de distances within hemisphere computed within each experiment and pooled across the two experiment (Within Exp) and computed between all electrode pairs across the two experiments (Across Exp). Distances are reported as median with lower and upper error bars indicating 10^{th} and 90^{th} percentile respectively.

(N) Relative context modulation reported for the hippocampus during the baseline period of experiments 1 and 2. Values are reported as mean ± s.e.m. over iterations of decorrelation curve estimation. P-value is computed by permutation test. ****** indicates $p < 0.001$.

Patient ID	Experiment ID	Session ID	Session Behavior	HPC Neurons	dACC Neurons	preSMA Neurons
P41CS	1	1	N/A	1	2	3
	1	2	N/A	2	7	2
	1	3	N/A	1	0	1
P42CS	1	1	N/A	4	16	16
	1	2	N/A	3	18	24
P43CS	1	1	N/A	15	0	0
	1	2	N/A	21	0	1
	1	3	N/A	19	0	0
P44CS	1	1	N/A	10	24	35
	1	2	N/A	5	13	27
P47CS	1	1	N/A	0	1	7
	1	2	N/A	0	0	7
	1	3	N/A	4	1	5
P48CS	1	1	N/A	14	20	19
P49CS	1	1	N/A	1	1	2
	1	2	N/A	5	0	2
P51CS	1	1	N/A	12	28	10
	1	2	N/A	13	15	3
	1	3	N/A	14	17	4
	1	4	N/A	14	17	4
	1	5	N/A	10	12	2
P53CS	1	1	N/A	0	0	11
	1	2	N/A	1	1	12
P56CS	1	1	N/A	3	9	14
	1	2	N/A	1	4	7
	1	3	N/A	1	5	1
P57CS	1	1	N/A	5	15	8
	1	2	N/A	6	14	20
	1	3	N/A	6	14	20
P58CS	1	1	N/A	2	23	52
	1	2	N/A	2	23	52
	1	3	N/A	1	20	33
P61CS	1	1	N/A	7	9	34
P61CS	2	1	X	N/A	N/A	N/A
	2	2	IP	15	16	16
P62CS	2	1	IA	7	1	4
	2	2	IP	7	1	4
	2	3	IA	11	2	10
	2	4	IP	4	0	4
P63CS	2	1	IA	34	8	16
	2	2	IA	29	3	8
	2	3	IP	33	4	9
P65CS	2	1	IP	19	0	0
	2	2	IP	19	0	0
	2	3	IA	7	0	0
P67CS	2	1	IA	7	7	6
	2	2	IP	7	7	6
	2	3	IP	7	3	3
	2	4	IP	7	3	3
P69CS	2	1	X	N/A	N/A	N/A
	2	2	X	N/A	N/A	N/A
P70CS	2	1	X	N/A	N/A	N/A
	2	2	IP	2	0	8
P71CS	2	1	IA	0	36	8
	2	2	IA	0	36	8
P73CS	2	1	X	N/A	N/A	N/A
	2	2	IA	5	7	15
	2	3	IP	9	5	24
	2	4	IP	9	5	24
P74CS	2	1	IA	0	0	15
	2	2	IP	0	0	15
P76CS	2	1	IP	28	34	10
	2	2	X	N/A	N/A	N/A
P78CS	2	1	IA	32	0	0
	2	2	IP	32	0	0
P79CS	2	1	IP	50	15	21
	2	2	IP	50	15	21
	2	3	IP	22	26	14
TWH162	2	1	IP	2	0	0
TWH163	2	1	IP	0	37	3
	2	2	IP	0	37	3
TWH165	2	1	IA	10	0	0
	2	2	IA	10	0	0
TWH172	2	1	IP	12	0	0
	2	2	IP	12	0	0

Table S1. Tabulation of Patients, Behavior, and Neurons.
Summary of patient information, the number of sessions performed for each experiment, the behavioral classification at the session level for experiment 2, and the number of recorded neurons per region per session. Patient behavior in experiment 2 is defined with respect to instances of high-level verbal instructions, where: Pre – "pre-instruction inference achieved", NE – "Inference not exhibited", post – "post-instruction inference achieved", and N/A – "did not qualify for analysis". Session behavior is defined with respect to performance on the first available inference trial, where: IA – "inference absent", IP – "inference present", X – "at or below chance non-inference performance". Such definitions of patient behavior do not apply to experiment 1, and are listed as "N/A".

Bibliography

1. Badre, D. & Nee, D. E. Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences* **22**, 170–188 (2018).

2. Duncan, J., Emslie, H., Williams, P., Johnson, R. & Freer, C. Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognitive Psychology* **30**, 257–303 (1996).

3. Miller, E. K. & Cohen, J. D. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience* **24**, 167–202 (2001).

4. Fuster, J. M. The Prefrontal Cortex—An Update: Time Is of the Essence. *Neuron* **30**, 319–333 (2001).

5. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).

6. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by human medial-frontal cortex. *Science* **368**, eaba3313 (2020).

7. Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat Neurosci* **18**, 1041–1050 (2015).

8. Niv, Y. Learning task-state representations. *Nat Neurosci* **22**, 1544–1553 (2019).

9. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology* **20**, 251–256 (2010).

10. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol Rev* **120**, 190–229 (2013).

11. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).

12. Courellis, H. S. *et al.* Abstract representations emerge in human hippocampal neurons during inference behavior. 2023.11.10.566490 Preprint at https://doi.org/10.1101/2023.11.10.566490 (2023).

13. Saez, A., Rigotti, M., Ostojic, S., Fusi, S. & Salzman, C. D. Abstract Context Representations in Primate Amygdala and Prefrontal Cortex. *Neuron* **87**, 869–881 (2015).

14. Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019).

15. Jamali, M. *et al.* Single-neuronal predictions of others' beliefs in humans. *Nature* **591**, 610–614 (2021).

16. Khanna, A. R. *et al.* Single-neuronal elements of speech production in humans. *Nature* **626**, 603–610 (2024).

17. Smith, E. H. *et al.* Widespread temporal coding of cognitive control in the human prefrontal cortex. *Nat Neurosci* **22**, 1883–1891 (2019).

18. Wang, S., Mamelak, A. N., Adolphs, R. & Rutishauser, U. Abstract goal representation in visual search by neurons in the human pre-supplementary motor area. *Brain* **142**, 3530–3549 (2019).

19. Sheth, S. A. *et al.* Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature* **488**, 218–221 (2012).

20. Aponik-Gremillion, L. *et al.* Distinct population and single-neuron selectivity for executive and episodic processing in human dorsal posterior cingulate. *eLife* **11**, e80722 (2022).

21. Kragel, J. E. *et al.* Distinct cortical systems reinstate the content and context of episodic memories. *Nat Commun* **12**, 4444 (2021).

22. Qasim, S. E. *et al.* Memory retrieval modulates spatial tuning of single neurons in the human entorhinal cortex. *Nat Neurosci* **22**, 2078–2086 (2019).

23. Donoghue, T. *et al.* Single neurons in the human medial temporal lobe flexibly shift representations across spatial and memory tasks. *Hippocampus* **33**, 600–615 (2023).

24. Aquino, T. G., Courellis, H., Mamelak, A. N., Rutishauser, U. & O′Doherty, J. P. Encoding of Predictive Associations in Human Prefrontal and Medial Temporal Neurons During Pavlovian Appetitive Conditioning. *J. Neurosci.* **44**, (2024).

25. Fu, Z. *et al.* The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* **376**, eabm9922 (2022).

26. Cueva, C. J. *et al.* Low-dimensional dynamics for working memory and time encoding. *Proceedings of the National Academy of Sciences* **117**, 23021–23032 (2020).

27. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat Neurosci* **21**, 102–110 (2018).

28. Vaidya, A. R., Jones, H. M., Castillo, J. & Badre, D. Neural representation of abstract task structure during generalization. *eLife* **10**, e63226 (2021).

29. Stokes, M. G. *et al.* Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* **78**, 364–375 (2013).

30.Stroud, J. P., Watanabe, K., Suzuki, T., Stokes, M. G. & Lengyel, M. Optimal information loading into working memory explains dynamic coding in the prefrontal cortex. *Proceedings of the National Academy of Sciences* **120**, e2307991120 (2023).

31.Wang, X.-J. 50 years of mnemonic persistent activity: quo vadis? *Trends in Neurosciences* **44**, 888–902 (2021).

32.Daume, J. *et al.* Control of working memory by phase–amplitude coupling of human hippocampal neurons. *Nature* 1–9 (2024) doi:10.1038/s41586-024-07309-z.

33.Kamiński, J. *et al.* Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat Neurosci* **20**, 590–601 (2017).

34.Kamiński, J., Brzezicka, A., Mamelak, A. N. & Rutishauser, U. Combined Phase-Rate Coding by Persistently Active Neurons as a Mechanism for Maintaining Multiple Items in Working Memory in Humans. *Neuron* **106**, 256-264.e3 (2020).

35.Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Current Opinion in Neurobiology* **58**, 141–147 (2019).

36.Simon, A. J. *et al.* Quantifying attention span across the lifespan. *Front Cognit* **2**, 1207428 (2023).

37.Buckley, M. J. *et al.* Dissociable Components of Rule-Guided Behavior Depend on Distinct Medial and Prefrontal Regions. *Science* **325**, 52–58 (2009).

38.Heilbronner, S. R. & Hayden, B. Y. Dorsal Anterior Cingulate Cortex: A Bottom-Up View. *Annual Review of Neuroscience* **39**, 149–170 (2016).

39.Nachev, P., Wydell, H., O'Neill, K., Husain, M. & Kennard, C. The role of the pre-supplementary motor area in the control of action. *Neuroimage* **36**, T155–T163 (2007).

40.Szczepanski, S. M. & Knight, R. T. Insights into Human Behavior from Lesions to the Prefrontal Cortex. *Neuron* **83**, 1002–1018 (2014).

41.Gläscher, J. *et al.* Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex. *Proceedings of the National Academy of Sciences* **109**, 14681–14686 (2012).

42.Umbach, G. *et al.* Time cells in the human hippocampus and entorhinal cortex support episodic memory. *Proceedings of the National Academy of Sciences* **117**, 28463–28474 (2020).

43.Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).

44.MacDonald, C. J., Lepage, K. Q., Eden, U. T. & Eichenbaum, H. Hippocampal 'time cells' bridge the gap in memory for discontiguous events. *Neuron* **71**, 737–749 (2011).

45.Bright, I. M. *et al.* A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proc Natl Acad Sci U S A* **117**, 20274–20283 (2020).

46.Howard, M. W. & Kahana, M. J. A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology* **46**, 269–299 (2002).

47.Folkerts, S., Rutishauser, U. & Howard, M. W. Human Episodic Memory Retrieval Is Accompanied by a Neural Contiguity Effect. *J. Neurosci.* **38**, 4200–4211 (2018).

48.Tyszka, J. M. & Pauli, W. M. In vivo delineation of subdivisions of the human amygdaloid complex in a high-resolution group template. *Human Brain Mapping* **37**, 3979–3998 (2016).

49.Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204–224 (2006).

50.Howard, M. W. *et al.* A Unified Mathematical Framework for Coding Time, Space, and Sequences in the Hippocampal Region. *J. Neurosci.* **34**, 4692–4707 (2014).

51.Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

52.Zhou, Y. *et al.* Transformers Can Achieve Length Generalization But Not Robustly. Preprint at https://doi.org/10.48550/arXiv.2402.09371 (2024).

53.Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661–1663 (2014).

54.Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat Neurosci* **22**, 1159–1167 (2019).

55.Wang, X.-J. Theory of the Multiregional Neocortex: Large-Scale Neural Dynamics and Distributed Cognition. *Annual Review of Neuroscience* **45**, 533–560 (2022).

56.Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H. & Wang, X.-J. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron* **88**, 419–431 (2015).

57.Gao, R., van den Brink, R. L., Pfeffer, T. & Voytek, B. Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife* **9**, e61277 (2020).

58.Dusek, J. A. & Eichenbaum, H. The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences* **94**, 7109–7114 (1997).

59.Hassabis, D., Kumaran, D., Vann, S. D. & Maguire, E. A. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences* **104**, 1726–1731 (2007).

60.Goudar, V. *et al.* Comparing rapid rule-learning strategies in humans and monkeys. *bioRxiv* 2023.01.10.523416 (2023) doi:10.1101/2023.01.10.523416.

61.Stefanacci, L., Buffalo, E. A., Schmolck, H. & Squire, L. R. Profound Amnesia After Damage to the Medial Temporal Lobe: A Neuroanatomical and Neuropsychological Profile of Patient E. P. *J. Neurosci.* **20**, 7024–7036 (2000).

62.Eichenbaum, H. On the Integration of Space, Time, and Memory. *Neuron* **95**, 1007–1018 (2017).

63.Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology* **37**, 66–74 (2016).

64.Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

65.Nieh, E. H. *et al.* Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).

66.Knudsen, E. B. & Wallis, J. D. Hippocampal neurons construct a map of an abstract value space. *Cell* **184**, 4640-4650.e10 (2021).

67.Gulli, R. A. *et al.* Context-dependent representations of objects and space in the primate hippocampus during virtual navigation. *Nat Neurosci* **23**, 103–112 (2020).

68.Goudar, V. *et al.* Comparing rapid rule-learning strategies in humans and monkeys. 2023.01.10.523416 Preprint at https://doi.org/10.1101/2023.01.10.523416 (2023).

69.Ito, T. & Murray, J. D. Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. *Nat Neurosci* **26**, 306–315 (2023).

70.Cole, M. W. *et al.* Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat Neurosci* **16**, 1348–1355 (2013).

Chapter 4

Fast and slow features of instructed human cognition

Motivation:

The human ability to specify and perform tasks on the basis of instructions correlates with simultaneous apparent benefits and drawbacks in behavior, the neural underpinnings of which have largely never been explored at the level of single neuron activity within the brain. The clear benefit is an acceleration in the rate of learning and task acquisition that humans exhibit, brought on by the ability to exactly specify a new task to be performed, and obviating the need to learn iteratively through trial and error, observation, or reinforcement. This "fast" feature of human cognition is thought to be supported by the ability to compositionally synthesize new tasks using natural language. A non-obvious drawback is the fact that switching from one task to another that has just been specified leads to a uniquely human penalty in the ability to perform that new task for a short period of time following switching. This "slow" feature of human cognition has been termed the "task switching cost" and appears to be a uniquely human phenomenon associated with engaging with new tasks in the world. In this chapter, I will develop two separate experiments, each designed to study one of these two features of instructed human cognition. Elements of each of these cognitive features can be appreciated in both experiments, but I will focus on key aspects of the psychophysical task structure, patient behavior, and neural activity that will allow us to gain some unique insights about both of these processes using the experimental and computational suite of tools I have introduced in the previous chapters of this thesis. Due to the strong overlap in methodology, I will only clarify methods and experimental details that are novel with respect to the previous chapters and experiments.

The Slow Feature: Geometry of task representations in human frontal cortical neurons is predictive of task switch costs

Introduction:

The process of switching between tasks occurs countless times throughout the day for an individual. Every instance of switching is accompanied by a cost, a decrease in task accuracy and/or speed immediately after switching that rapidly fades away[1]. Though this switch cost is reducible when preparatory time is given after instructions, an irreducible switch cost is always present the first time one engages in a task when switching from a different task. The presence of switch costs in animals is debated, being absent from some species entirely, but is a prominent aspect of human cognition[2–4]. The neural mechanisms that generate switch costs remain unknown and are hotly debated. Theories center around two possible causes: reconfiguration and lingering activity (inertia) related to the prior task[1]. Some evidence from intracranial recordings exists supporting these proposed explanations[5,6], which indicate a key role of the medial frontal cortex (MFC). However, the neurophysiological basis of switch costs remains elusive.

To arbitrate between different theories of switch costs, we recorded the activity of large populations of single neurons in the MFC of neurosurgical patients performing a task with frequent instructed switching. We find that the task context representations immediately following and far from a switch exist in orthogonal subspaces composed of non-overlapping populations of neurons.

The task representation in the latter subspace persistently encoding the previous task is predictive of switch costs.

Methods:

Experimental Design:
Subjects alternated between two possible tasks: categorization (e.g. "Is this an image of X?", where X is the target category), and memory (e.g. "Have you seen this image before?") (Fig. 4.1a). Each experiment consisted of 48 blocks of 8 trials. Task instructions were given once at the start of each block, and needed to be remembered for the ensuing 8 trials (Fig. 4.1b). All questions were yes/no questions, with subjects answering as quickly as possible. We refer to the question being answered as the context for that block, either Categorization (Cat) or Memory (Mem). Images belonged to one of two categories (fruits, faces), with some repeated ("old") and some shown the first time, resulting in 8 total possible conditions (Fig. 4.1c). A balanced number of trials of each condition were present in every block and at every trial number across blocks. Switch costs were operationalized as the excess time taken to complete the first trial after switching tasks. For each block, patients control when to proceed from the instruction screen to the first trial (Fig. 4.1a), such that they are sufficiently prepared and the behavioral cost present during Trial 1 after a switch is the irreducible switch cost.

Neural Signal Recording and Processing:
Patients with pharmacologically intractable epilepsy were implanted with Behnke-Fried electrodes [7] that allowed for recording of single-unit activity from medial frontal cortical (MFC) structures including the dorsal anterior cingulate (dACC) and pre-supplementary motor area (preSMA) (Fig. 4.1d). Unit activity from these regions was isolated using standard spike sorting techniques[8]. Spikes were counted during two time periods: baseline (-1 to 0 s prior to stimulus onset) and stimulus (0.2 to 1.2s after stimulus onset). "Trial 1" baseline spikes are recorded after a patient has read the instructions and pressed a button initiating a block, but has not yet performed the task instructed for that block.

Results:

Baseline context representations emerge in orthogonal subspaces following instructions.
Data recorded recorded over 56 sessions (n = 35 patients) yielded 757 well isolated neurons. Switching costs were robust for both tasks (Fig. 4.2a, each line is a session), with Trial 1 after an instruction screen on average 40% slower than the average block RT. We decoded task context from spikes counted during the baseline period and found context to be robustly decodable from activity of MFC neurons during Trials 4-8 after a switch (Fig. 4.2b, left, decoder trained on Trials 4-8). However, this decoder (henceforth steady-state subspace) did not generalize to decode activity in Trial 1. Yet, context was decodable from Trial 1 when training and testing a decoder during Trial 1 only (Fig. 4.2b, right, red). Conversely, the Trial 1 decoder failed to generalize to Trials 4-8, with context decodability in the subspace identified by this decoder (henceforth switch subspace) falling to chance after Trial 3 post-switch. These two context coding subspaces were orthogonal (Fig. 4.2c) by virtue of being largely non-overlapping populations of neurons (Fig. 4.2d,e).

Context decodabiltiy in both subspaces is predictive of task switch costs on upcoming trials.

Greater context decodabiltiy in both subspaces predicted faster RT (lower switch cost) on the upcoming trial (Fig. 4.2f). On slow trials, the context of the previous block was decodable from dACC as indicated by below-chance decoding (Fig. 4.2f, right).

Stimulus and context representations show evidence of reconfiguration during switch trials.
Representational geometry was quantified during the stimulus period by performing SD and CCGP analysis on Trials 4-8 (Stay) and Trial 1 (Switch). All three stimulus properties (context, novelty, category) were decodable on Stay trials in both dACC and preSMA (Fig. 4.3a). However, dACC alone exhibited a significant decrease in SD (Fig. 4.3a, black line) and CCGP for context (Fig. 4.3b, red) on switch trials. The mis-configuration of the dACC representation on Switch trials is visualized in Fig. 4.3c,d by performing multi-dimensional scaling (MDS) on condition-averaged neural activity from dACC alone. The systematically structured Stay trial representation (Fig. 4.3c) is contrasted with the relatively disorganized Switch trial representation (Fig. 4.3d).

Discussion:

Both the task-set inertia and reconfiguration theories are consistent with aspects of our data. Baseline and stimulus period task representations in the MFC undergo reconfiguration following switch trials, and previous-context decodability is correlated with higher switch costs (inertia). Further analysis is needed to explore switch cost prediction during the stimulus period, switch-trial response conflicts, and to clarify the effect of practice, which can reduce switch costs.

Figures:

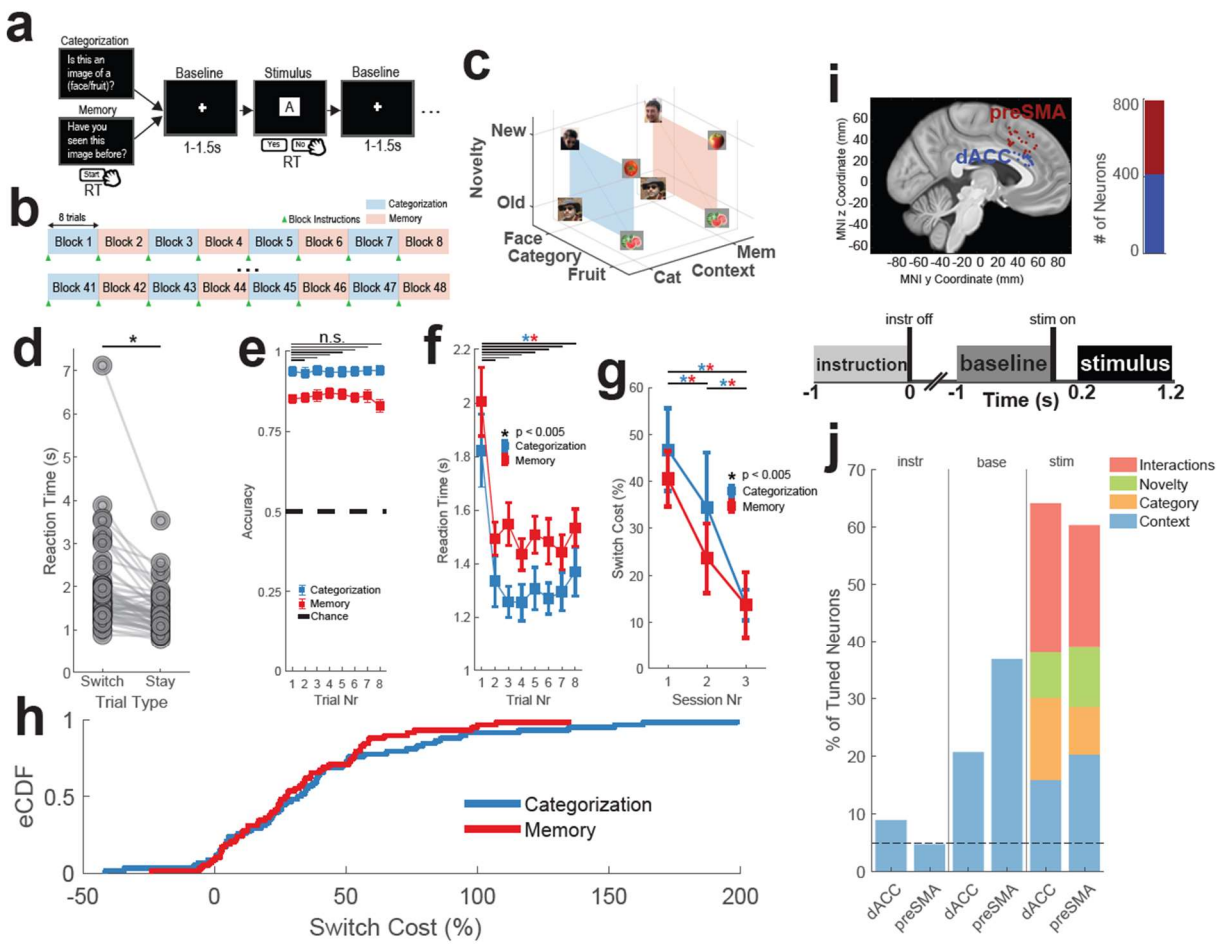

Figure 4.1. A task for studying switch costs in humans: task design, behavior, and neurons. **(a)** Illustration of trial structure. Following an instruction screen, trials consisted of a pre-stimulus baseline followed by stimulus presentation during which patients answered yes or no (left or right button press) in a speeded manner according to the current stimulus and context. No trial-level feedback was provided, and patients immediately proceeded to the next baseline. **(b)** Illustration of the block structure. Task context alternated every 8 trials, and the experiment consisted of 48 blocks (384 trials total). **(c)** Illustration of task state space structure. Stimuli from two categories (faces and fruit), that either had or had not been previously encountered by the patient in earlier trials (old and new) were presented in a balanced manner in each of the two contexts (categorization and memory). This visualization is reflective of the disentangled structure of the task variables, and does not necessarily reflect how neurons will organize their responses in neural state-space to each of these conditions. **(d)** Reaction time on switch (left) and stay (right) trials shown for all sessions and averaged over trials. * indicates $p<0.05$ with ranksum test over sessions. **(e)** Accuracy computed for different trial positions in each block separately for the categorization (blue) and memory (red) tasks. Points and error bars represent mean ± s.e.m. over sessions. n.s. indicates $p>0.05$ between Trial 1 accuracy (switch trial) and all other trials for both tasks using ranksum test over sessions. Black dashed line indicates chance performance. **(f)** Reaction time computed for different trial positions in each block separately for the categorization (blue) and memory (red) tasks. Points and error bars represent mean ± s.e.m. over sessions. * indicates $p<0.005$ between Trial 1 reaction time and all

other trials for both tasks using ranksum test over sessions. **(g)** Switch cost reported as a function of session number for categorization (blue) and memory (red). Points and error bars represent mean ± s.e.m. over sessions. * indicates $p<0.005$ using ranksum test over sessions. **(h)** Distribution of switch costs for all sessions shown for categorization (blue) and memory (red). Switch costs here are reported as % of average non-switch reaction time. **(i)** Electrode locations. Each dot corresponds to a single microwire-bundle. Locations are shown on the same hemisphere for visualization purposes only. Shown are pre-Supplementary Motor Area (preSMA, red) and dorsal Anterior Cingulate Cortex (dACC, blue). Total number of neurons recorded in each region is shown in the bar graph to the right (841 total). **(j)** Number of single units across brain areas exhibiting significant Main effects or interaction effects (n-way ANOVA with interactions, $p < 0.05$, see methods) to at least one of the principal task variables or to combinations of variables during three different 1-s time windows throughout the experiment: instruction encoding, baseline, and stimulus periods. Time windows are shown in the inset above.

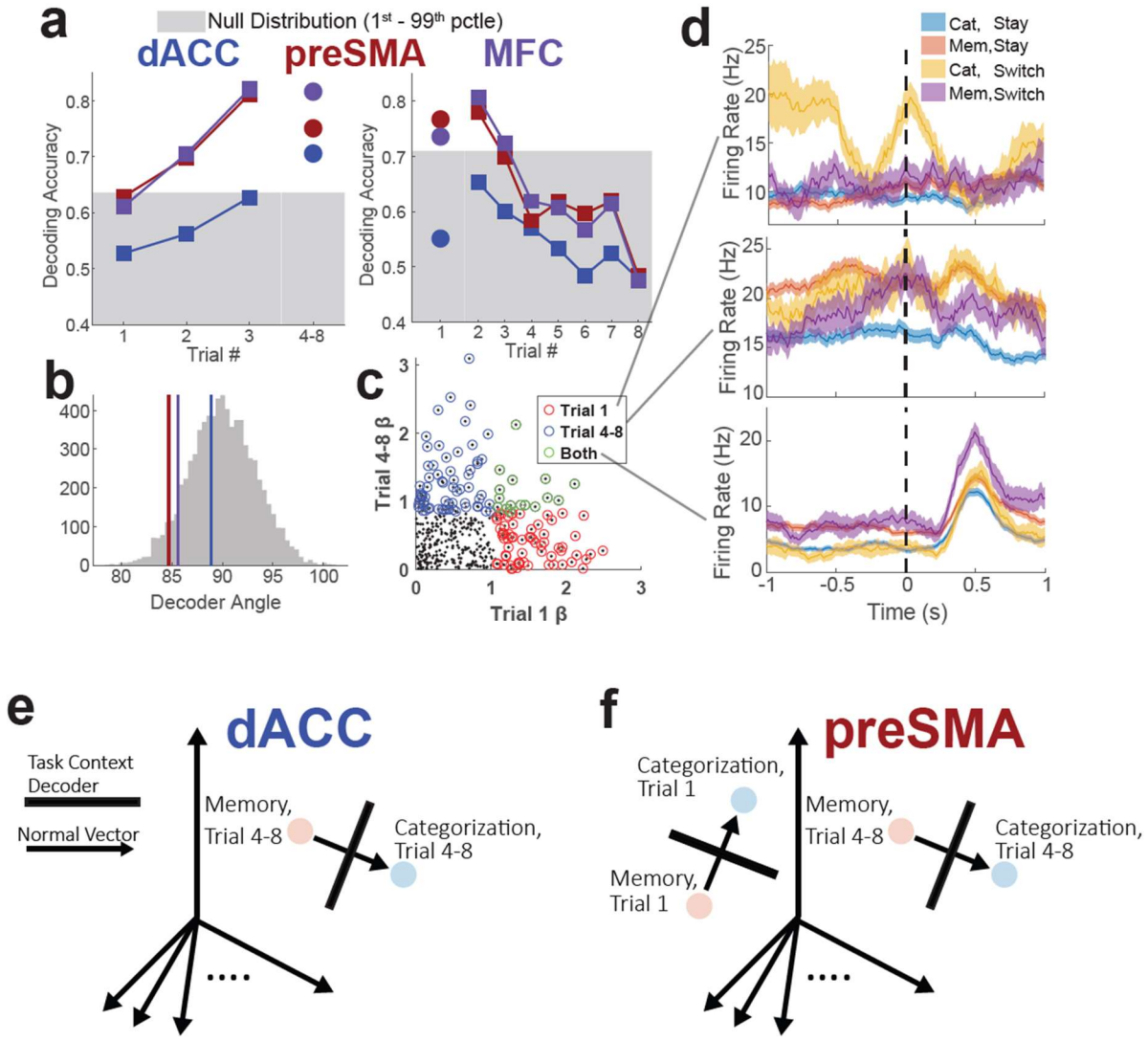

Figure 4.2. Baseline representations of context in preSMA and dACC.
(a) Baseline context decoder trained on Trials 4-8 (right) and on Trial 1 (left) after task switching. Circles indicate cross-validated training performance and squares indicate generalization performance to held-out trials. 99th pctle of shuffle null distribution shown in gray. Decoder performance is color coded according to region, with dACC in blue, preSMA in red, and MFC (both dACC and preSMA) in purple. **(b)** Angle between context coding vectors computed from Trial 1 and Trial 4-8 decoders. Gray histogram indicates shuffle null. **(c)** Scatter plot of single-neuron importance index (β) for Trial 1 and Trial 4-8 decoders in the preSMA only. Each black dot corresponds to one neuron. Neurons in the top 20% for Trial 1 decoder (red), Trial 4-8 decoder (blue), or both (green) are circled. **(e)** Example PSTHs of neurons contributing to each of the context decoders. Conventions for PSTH plotting are identical to those used in previous chapters.

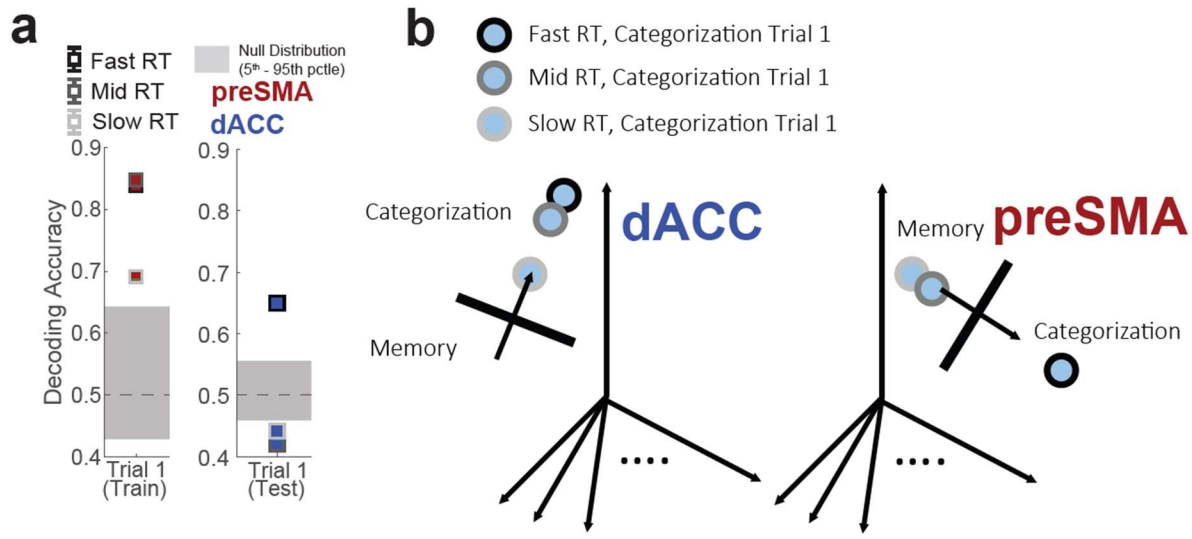

Figure 4.3. Baseline context representations correlate with the degree of switch cost. **(a)** Correlation of baseline context representations and Trial 1 reaction time (switch cost) for the preSMA context decoder trained on Trial 1 (left) and the dACC context decoder trained on Trials 4-8 (right). Fast, medium, and slow switch trial decoding performance is highlighted with a black, gray, and white outline respectively. Decoder performance for the preSMA is cross-validated. Gray rectangles indicate 5^{th}-95^{th} pctle of shuffle null distribution. Performance below 0.5 indicates significant decoding of the other context (e.g. neurons significantly representing the memory task while the patient performs the categorization task). **(b)** Schematic illustrating the decoder-based findings shown in **(a).** Horizontal black lines represent context-decoder hyperplanes in neural state space, and arrows represent normal vectors. The schematic illustrates the neural state representation of the Categorization task during the trial 1 baseline as it correlates with the degree of switch cost (white, gray, black, for slow, mid, fast) on the upcoming trial.

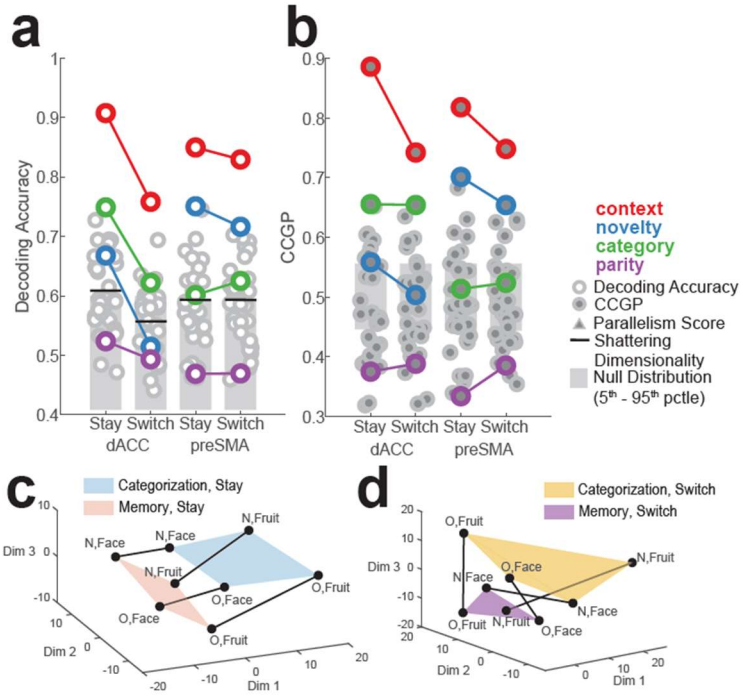

Figure 4.4. Stimulus period context and stimulus representations. (a) Reduction in decodability of task-relevant dichotomies and reduction of shattering dimensionality on switch trials compared to stay trials in the dACC (left). This effect is absent from the preSMA (right). **(b)** CCGP of context representation significantly reduced on switch trials in dACC. Dimensionality reduction of dACC neural responses using MDS during stay **(c)** and switch **(d)** trials. Plotting conventions used here are identical to those used for geometric measure plots in Chapter 2.

Bibliography:

1. Monsell, S. Task-set control and task switching. in *The handbook of attention* 139–172 (Boston Review, Cambridge, MA, US, 2015). doi:10.1093/med:psych/9780198528883.003.0002.
2. Caselli, L. & Chelazzi, L. Does the Macaque Monkey Provide a Good Model for Studying Human Executive Control? A Comparative Behavioral Study of Task Switching. *PLoS One* **6**, e21489 (2011).
3. O'Donoghue, E. & Wasserman, E. A. Pigeons proficiently switch among four tasks without cost. *Journal of Experimental Psychology: Animal Learning and Cognition* **47**, 150–162 (2021).
4. Stoet, G. & Snyder, L. H. Executive control and task-switching in monkeys. *Neuropsychologia* **41**, 1357–1364 (2003).
5. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by human medial-frontal cortex. *Science* **368**, eaba3313 (2020).
6. Weber, J. *et al.* Subspace partitioning in the human prefrontal cortex resolves cognitive interference. *Proceedings of the National Academy of Sciences* **120**, e2220523120 (2023).
7. Fried, I. *et al.* Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. *J Neurosurg* **91**, 697–705 (1999).
8. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204–224 (2006).
9. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).
10. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *Journal of Neuroscience Methods* **154**, 204–224 (2006).
11. Fried, I. *et al.* Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients: Technical note. *Journal of Neurosurgery* **91**, 697–705 (1999).
12. Courellis, H. S. *et al.* Abstract representations emerge in human hippocampal neurons during inference behavior. 2023.11.10.566490 Preprint at https://doi.org/10.1101/2023.11.10.566490 (2023).

The Fast Feature: Task representations in frontal cortical neurons inherit the compositional structure of natural language and facilitate zero-shot generalization.

Introduction:

Humans are the only species on planet earth that can learn arbitrarily complex, novel tasks in (nearly) constant time. Given certain constraints on the nature of the task, namely that it is specified using natural language that is comprehensible to the receiver, and is composed of elements, either states to recognize or actions to perform, that are meaningful and familiar to the receiver through past experience, the time cost incurred grows with the length of the message, in words, as codified by natural language, and the time it takes the transmitter to state it[1]. That is to say, the speed at which humans learn new instructed tasks is strongly supported by the compositional structure of language, and the usage of a basis set of states and actions that can be recombined to specify a task that an individual has never before performed, but can immediately execute accurately on the first try. In the parlance of machine learning, the novel, linguistically specified task lies outside of the training distribution of the receiver, but the receiver can nonetheless generalize to this novel data distribution zero-shot (i.e. without first observing any examples), as opposed to one- or few- shot learning where one or several example solutions to the new task are first provided. Thus, humans have the ability to perform "compositional generalization" with tasks, that is, to synthesize novel tasks compositionally through language and to immediately perform them with high accuracy.

As humans, we might take such capabilities for granted given how integrated they are with our everyday lives. However, compared to a non-human primate, which might take many thousands of trials over the course of several months to learn a psychophysical task, a human can acquire that same task with a 60-second instruction screen. Modern state-of-the art reinforcement learning based networks can be trained to play video-games with super-human performance, but again our sample-efficiency exceeds these systems by many orders of magnitude provided we have a one to two minute explanation from a friend before we hop into a game on our Atari64. Furthermore, after having learned to play a game, a brief set of instructions, such as "Now try to get the lowest score you can without dying!" or "Try to end each level on a score that is divisible by 10!" can systematically and radically alter our behavior without needing to re-learn how to play the game from scratch[2].

The ability to compose novel rules and tasks to accelerate learning and constrain behavior must have some underlying neurophysiological substrate in the human brain. The debate regarding how systematic and rule-like the systematic compositionality exhibited by human thought and language dates back to Fodor and Pylyshyn, and earlier[1,3]. Furthermore, novel evidence suggests that certain kinds of neural networks trained under specific meta-learning frameworks and compositional objective functions can indeed perform human-like compositional generalization[4]. It has been speculated that the ability to generate compositionally structured behavior is supported by underlying neural representations that they themselves are also compositional[5–7]. The argument proceeds as follows: states and actions that adopt a vector-representation through the activity of a population of neurons (either biological or artificial) can be composed through vector addition such that the representation of a completely novel set of instructions or rules can be accurately decoded by linear downstream readouts because it contains the vector-additive sum of all of the individual attributes for which decoders already exist. The behavioral advantage of such disentangled, or abstract, representations in human neurons was expounded in Chapter 2, albeit for a combination of stimulus and latent context variables, the latter of which was specified through instruction in some cases, but that needed to be learned experientially in every session. Similar abstract representations have been observed at the level of single-neurons and neural populations in other regions in the frontal,

temporal, and parietal lobes under various cognitive task demands, and in each case the compositional structure of the stimulus, task context, and action representations create a situation where the neural response to a specific set of held-out task states is predictable provided one can train a linear decoder on the remaining conditions[5,8–11]. Representations of variables implemented in this manner are quantifiably abstract, and having multiple, jointly abstract variables encoded in neural state space is the hallmark of a compositional representation. The seeming ubiquity of such representations at the single-neuron level in the human brain in a variety of non-linguistically specified experimental settings (e.g. stimulus identity, familiarity, error and conflict, etc…) invites the question of whether such representations would arise for natural-language based rules that are compositionally specified by construction, and if the emergence of such representations correlates with the speed at which humans can generalize to novel tasks.

In the field of machine learning, the utility of learning low-dimensional, disentangled representations of high-dimensional input signals for both discriminative and generative tasks has successful history[12]. Furthermore, it has long been appreciated that standard feedforward neural networks, recurrent networks[13], and transformer-based models[14] trained large amounts of natural language data under various objectives all form internal representations that are structured in such a way as to reflect semantic relationships through systematic geometric relationships in vector-embeddings. Recurrent neural networks trained to perform many cognitive tasks simultaneously also learn internal representations that rely on modules of neurons that are compositionally activated in service of new tasks that share elements with previously learned tasks[15].

To investigate the neurophysiological basis of instruction-based zero-shot generalization in humans, and to probe for the presence, study the format, and determine the behavioral relevance of compositional task representations in the human brain, we recorded the activity of populations of neurons in the brains of awake, behaving epilepsy patients who explored a large task space defined by combinatorically-specified task rules. We find that neurons in the frontal cortex form a compositional representation of the instructed task rules that mimics the compositional structure of the natural language prompts. We demonstrate that individual task rules with hierarchically nested compositional structure can be simultaneous encoded in a jointly abstract format as patients generalize to novel task rule combinations. Furthermore, by leveraging a recording opportunity in a single bilingual patient, we develop evidence that the same task rule representation can be induced in a manner that is invariant to the language used to specify the rules.

Methods:

Experimental Design:

A task involving compositionally specified task contexts was constructed using the framework of Boolean operations performed on the Target category membership of pairs of images. The task context, i.e. the combination of rules enacted at any given time, was determined by three task rules that uniquely specified one of 16 possible task contexts. The location of each rule on the instruction screen, name of the rule, and possible values of the rule are as follows: Top Row, Target Rule, (ALIVE, FLY). Middle Row, Boolean Rule, (AND, NAND, OR, NOR). Bottom Row, Motor Rule, LEFT, RIGHT). The total number of Instructions were provided once at the beginning of each block (Fig. 4.5a, Left) Patients responded to begin each block. Trials (Fig. 4.5a, Right) consisted of a pre-stimulus baseline (baseline) followed by the first stimulus presentation (stim 1, 1s), a brief, jittered delay (1-1.5s), and then the second stimulus (stim 2, RT) at which point the patients provided a response. Trial-level feedback was provided after another brief delay. Blocks consisted of 6 trials,

and task contexts were presented once for every traversal of task space, which was a sequence of 16 blocks. Each experiment consisted of 4 full passes of task space (Fig. 4.5b). Thus, patients completed 384 trials during a standard session of this experiment. An illustration of the task state space structure is shown in Fig. 4.5c. Individual points represent unique task contexts specified by combinations of Target, Boolean, and Motor rules, unlike similar schematics in previous chapters where points corresponded to trial states specified by both context-level and stimulus-level variables. The top right corner of the red square (ALIVE, AND, RIGHT) corresponds to the task context shown on the instruction screen in Fig. 4.5a. Each of the two stimuli presented in a trial was drawn from one of four semantic categories: planes, birds, cars, and humans. Target rule memberships were defined such that humans and birds are ALIVE, and birds and planes can FLY (Fig. 4.5d). Boolean task rules were also codified using number counting and equality/inequality comparisons for ease of learnability for the patients. Note, however, that the relational structure of the Boolean operators in their native task construction space (Fig. 4.5e) differs from the relational structure realized through number counting (Fig. 4.5f). A second variant of the task with Boolean task rules codified using natural language (Fig. 4.5g) was also administered in some cases.

For one patient, in order to increase the accessibility of the task, a reduced variant was constructed that consisted of 160 trials. This variant featured only the Boolean and Target rules, with block lengths increased to 20 trials, and had a single pass through task space instead of four. These reductions were deemed necessary in order to accommodate the patient, and reduce completion time of the task so that back-to-back sessions could be run continuously. Data from this patient were excluded from all major analyses apart from Fig. 4.10.

Neural Signal Recording and Processing:

Patients with pharmacologically intractable epilepsy were implanted with Behnke-Fried electrodes [16] that allowed for recording of single-unit activity frontal and temporal lobe structures including the ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate (dACC), pre-supplementary motor area (preSMA), hippocampus, amygdala, and ventral temporal cortex (VTC). Unit activity from these regions was isolated using standard spike sorting techniques[17]. Spikes were counted during five time periods: baseline (-1 to 0 s prior to stimulus 1 onset), stim 1 (0 to 1s after stimulus 1 onset), delay (-0.75s to 0s prior to stimulus 2 onset), stim 2 (0 to 1s after stimulus 2 onset), and response (-1s to 0s prior to patient response).

Results:

Patients can perform zero-shot generalization in a compositionally-constructed task space.

The full experiment was completed in 15 sessions (n = 7 patients), with an additional 3 sessions of the limited variant of the task being collected from a single patient. The following results all pertain to the full experiment sessions unless otherwise specified. Patients performed significantly above chance throughout all blocks of the task, with an average performance of 91.1% ± 0.5% (Fig. 4.6a, mean ± s.e.m. over blocks). Patients also exhibited a learning effect in their reaction time (Fig. 4.6b), which decreased significantly from 0.92 ± 0.42 on the first block to -0.15 ± 0.08 on the 16th block (mean ± s.e.m. over sessions) with z-scored reaction times within session ($p<0.05$ Ranksum between blocks). Absolute reaction times were on average 1.77s ± 0.20s (mean ± s.e.m. over sessions). No significant trend in accuracy was detected as a function of block number over the experiment on average ($p=0.4$, Linear Model, accuracy vs block nr). However, a significant effect was present in

the reaction time ($p=1.4x10^{-5}$, Linear Model, reaction time vs block nr), indicating that patients did exhibit some amount of trial-level learning as they increased experience with the task. However, that learning was purely related to speed of task execution and not ability to accurately perform the task, as no trends in accuracy were present.

In order to specifically quantify the zero-shot generalization performance of the patients, we also performed analysis on blocks 5-16 in isolation. Despite using different stimuli between the tutorial and full task, the tutorial provides experience for the first four task contexts (blocks), thus obviating claims related to zero-practice generalization for these blocks in particular. We compare the accuracy and reaction time of blocks 5-16, henceforth generalization blocks, to the remainder of the experiment (blocks 17-64, henceforth repeated blocks), which constitutes 3 repeated traversals of the entire task space. Average task performance during generalization blocks did not significantly differ from task performance during repetition blocks (92.6%±0.9% vs 91.2%±0.5%, p=0.16 Ranksum over sessions). However, average z-scored reaction time did significantly differ between generalization and repetition blocks (0.06±0.03 vs -0.06±0.02, p=0.0068 Ranksum over sessions), with reaction time on generalization blocks being slower.

Patients also exhibited a significant switch cost in their reaction time (Fig. 4.6c, p=0.0040, Linear Model, reaction time vs trial nr) and not in their accuracy (Fig. 4.6d, p=0.87, Linear Model, accuracy vs trial nr), consistent with the switch-cost behavior exhibited in the experiments shown in previous sections.

Together, we take these findings to indicate that, while patients may be slower in their responses early in the experiment and early in trials following a switch, consistent with standard cognitive effects related to task practice and switch costs, they are able to perform zero-shot generalization in our compositional task with high accuracy.

Single neurons in the frontal and temporal lobe exhibit mixed responses to task variables.

Neural recordings were performed during the 15 sessions, yielding 1020 well isolated neurons across all brain areas including vmPFC (148), dACC (117), preSMA (186), amygdala (227), hippocampus (184), and VTC (158). For the purposes of subsequent geometric analyses, neurons from vmPFC, dACC, and preSMA are grouped together under the label “frontal cortex” (henceforth FC), with locations of microelectrode recordings indicated in Fig. 4.7a. Thus, the total number of FC neurons participating in all analyses henceforth is 451. Medial temporal lobe (MTL) will be used to refer to neurons from the hippocampus and amygdala (411 neurons). Univariate analyses performed on spike counts in each of the 5 time periods described earlier indicate that between 40 and 50% of FC neurons exhibit significant Main effects or interaction effects (3-way ANOVA with interactions, $p < 0.05$ for any term) to at least one of the task rule variables (Target, Boolean, Motor) or to combinations of those variables. A unit is linearly tuned if it has at least one significant main effect, and non-linearly tuned if it has at least one significant interaction term in the ANOVA model. Example neurons linearly tuned to Target (Fig. 4.7d), Motor (Fig. 4.7e), and Boolean (Fig. 4.7f-h) are shown. This tuning contrasts to classical stimulus tuning, such as visual category tuning while stimuli are being presented or during the inter-stimulus delay period (example hippocampal visual category neuron shown in Fig. 4.7c).

Category-tuned neurons in the MTL are conditionally modulated by the Target rule.

In addition to task rule tuning in the frontal cortex, a large proportion of MTL neurons exhibit image category tuning during the stim 1 (27.7%), delay (12.1%), and stim 2 (29.8%) periods, consistent with many previous single-neuron studies recording from these regions (1x4 ANOVA for image

category, $p<0.05$ for significant neurons). However, a new class of neuron that modulates its categorical stimulus response as a function of the currently instantiated Target rule was also detected (2-way ANOVA for image category and Target rule, $p<0.05$ interaction effect for significant neurons). These neurons were also present throughout stim 1 (13.2%) and stim 2 (11.3%) periods. Such neurons had previously not been observed in migrating categorization-rule tasks with changing semantic target categories, but prominently feature in the MTL of patients performing this task. Examples of such neurons are shown in Fig. 4.8. It is noted that target-conditional category responses are observed even for neurons preferring categories whose Target membership does not change between the two Target levels. For example, Fig. 4.8a shows a bird-preferring neuron modulated by Target = FLY and Target = ALIVE despite the category not crossing the Target membership boundary. This example is contrasted with Fig. 4.8c which shows a face-preferring neuron modulated by Target, and Target membership of this category does change between ALIVE (yes) and FLY (no).

Frontal cortical neurons form a compositional representation of Boolean and Target rules.

Initial decoding analyses were performed to establish the presence of task rule information at the level of the neural population in FC. Decoding of unique task contexts (1/16) was performed, and demonstrated above-chance decodability during all experimental time periods (Fig. 4.9a), ranging from 15%-21% (chance = 6.25%). To quantify the representational format of the Boolean rule, we first perform a geometric analysis over the three possible dichotomies, corresponding the target number, equality, and SAT-0/1 as described in the methods. All three dichotomies were significantly decodable (Fig. 4.9b, left), and the CCGP (Fig. 4.9b, middle) and parallelism score (Fig. 4.9b, right) together suggest that the Boolean task rule representation is organized in a 2-dimensional configuration around the Equality and SAT variables, as evidenced by the significantly elevated parallelism and CCGP. In that case, target number would exhibit negative parallelism/below chance CCGP, and this is indeed what the analysis reveals. Given this organization of the Boolean task rule, we can next collapse this 2-dimensional space into a single dimension by marginalizing over one of the two significant dichotomies (selecting the Equality axis to retain as described in the methods), thus creating a binary Boolean task rule variable that can be incorporated into a full 35-balanced dichotomy geometric analysis during all time periods when combined with the Target rule and the Motor rule.

Performing the full balanced dichotomy analysis in Boolean-Target-Motor task rule space reveals significant decodability of the Boolean rule (Fig. 4.9c, green) and the Target rule (Fig. 4.9c, blue) simultaneously, most prominently during the stim 2 period. High parity decodability (Fig. 4.9c, purple) during this period is a signature of non-linear distortions in the representation. However, CCGP (Fig. 4.9d) and Parallelism score (Fig. 4.9e) analysis reveal that the Boolean and Target rules are in an abstract format. That is, they exist simultaneously in the FC neural state space, and are disentangled in a manner consistent with a linearly compositional representation of task rules. A representation of the Motor task rule is absent from this population of neurons.

Taken together, the analysis of Boolean task rule structure in Fig. 4.9b and the balanced dichotomy analysis in Fig. 4.9c-e indicate that the representation of Boolean task rules is 2-dimensional, and is disentangled form a third, 1-dimensional, simultaneous representation of the Target task rule during the stim 2 period. These findings are summarized in the schematic shown in Fig. 4.9f, which clarifies the relational structure of the different task contexts.

The geometry of the task representation is invariant to language in one bilingual patient.

In a single patient who was raised in a home speaking both English and Spanish, the opportunity arose to perform back-to-back sessions of this experiment in each of the two languages in which this patient was proficient. The compositional structure of all task rules, including the relational structure of the Boolean operators codified as number counting rules, is preserved between English and Spanish, allowing for direct comparison of neural representations when identical tasks with approximately identical linguistic structure (e.g. Fig. 4.10a) across the two languages. For technical reasons (see methods, discussion), a smaller number of trials and task contexts were provided to this patient in the two sessions. Nevertheless, an analysis of the Boolean task rule geometry during the baseline period (Fig. 4.10b) revealed that the relational structure of the Boolean task rule was preserved, being a 2-dimensional space organized around the target number and SAT-0/1 variables in both the English and Spanish variants of this task. Thus, the same relational task structure emerges in the frontal cortical task context representation of a bilingual patient independently of the language used to specify the tasks.

Discussion:

Thus, we have preliminary evidence suggesting that neurons in the human frontal cortex generate a compositional representation of novel task rules throughout the process of instructed learning, inheriting the structure of the language prompts that are used to specify them. The compositional representation in FC demonstrated here departs sharply from those studied in the previous chapter since no part of the representation relied on dimensions generated by stimuli that were available in the sensory input stream of the patient. Task rules were presented once during the encoding screen of each block, and the subsequent compositional representation of task context that formed relied purely on the internally-maintained persistent representation of those rules by neurons across many intervening trial phases until the next instruction block. It should also be noted that the task context representation was not purely disentangled around the linguistic structure of the natural language prompt. In particular, during the stim 2 period, the emergence of the contingency representation[18] (SAT-0/1) as an organizing variable for the Boolean rule instead of the target number reflects a structured representation of task states according to a state-predictive model of the environment rather than a representation strictly structured around linguistic input. The presence of such organization during both the stim 2 period (Fig. 4.9b) and the baseline period (Fig. 4.10). Indicates that this organization persists regardless of whether the patient is in the process of making a decision, and provides evidence against the idea that a rehearsed phonological loop of the task rule alone is structuring the FC task representation.

A panoply of questions remain related to the specific representational geometry adopted by the FC and MTL in this experiment. First, and foremost, a much larger corpus of data needs to be collected. Currently, grouping neurons across all FC regions creates a neural state space that is assuredly not available to any reasonable downstream readout in the brain. Segregated contribution to this compositional task context representation by neurons in different regions of FC remains to be clarified.

A second, pressing line of inquiry relates specifically to the task context representation formed as patients are generalizing to truly novel task contexts during generalization blocks (5-16). The vast majority of the trials contributing to the representation geometry analyses performed here were repeated blocks, and thus the geometries we describe could only be emerging after a significant amount of exposure to the task (e.g. in the second half), and could not contribute to the

ability of a patient to zero-shot generalize at the beginning of the experiment. That behavior could be subserved by a different representation that is masked by the overwhelming amount of repeated block data that is contributed to the analysis. More careful time-resolved analysis across blocks during the generalization blocks and comparison to repeated blocks is needed to clarify this point.

A third line of inquiry relates to the natural language parametrization of the Boolean task. Several sessions of this variant have been collected and were included in the above analysis, but only for the balanced dichotomy analysis over all three task rules. These sessions were included since the process of marginalizing the Boolean task rules over the "Equality" axis in the number counting parametrization (Fig. 4.5f) is equivalent to marginalizing over the "Negation" axis in the natural language parametrization (Fig. 4.5g). Of course, the one component of task construction space that this collapse does not take into account is the fact that the Boolean task rules seem to be organized around the contingency representation, which groups the diagonals together in the number counting parametrization but not in the natural language variant (something I have realized just now as I am writing this). Collapsing the contingency representation (Parity dichotomy in boolean operator space) injects signal on both sides of the categorization boundary in the collapsed condition, and is likely problematic for the full geometric analysis. More data needs to be collected on the natural language variant and the Boolean rule geometry needs to be clarified therein before I try to draw more conclusions on the collapsed analysis of the full task context representation. It is possible that in order to do the alignment across Boolean variants I need to reparametrize one of the spaces around the contingency dichotomy before marginalizing one of the two dimensions out.

A fourth line of inquiry corresponds to the dynamics exhibited by the task context representation across trial phases. Task-context related signals are clearly present throughout all time periods in the experiment (Fig. 4.9a), and signatures of representation re-coding across time periods is already evident in Fig. 4.9c-e, where the representation transiently reorganizes around a different variable that forces the target category rule to exhibit below-chance CCGP and parallelism, followed by rapid re-emergence during the stim 2 period. Characterizing these step-wise transformations in the context code, the emergence of the contingency representation, and how these representations correlate trial-by-trial with accuracy and reaction time of patients can provide critical insights into the sequence of representations the frontal cortex adopts during sequential decision making, moving us in the direction of beginning to unravel the algorithms employed by the human brain to solve general, arbitrarily complex tasks.

Another exciting line of inquiry relates to the presence of conditional category responses in MTL neurons. Conditional category responses in the hippocampus imply presence of non-linearly distorted population codes for stimulus and context in the MTL. This prospect represents a departure from the strongly disentangled compositional rule code present in the FC, and requires a more extensive geometric analysis performed in stimulus x context space as opposed to simply performing analysis on task context representations while implicitly (through single-trial resampling) or explicitly (through condition averaging) removing stimulus information as was done for the geometric analysis of task context we performed in the FC here. The prior absence of such conjunctive context/stimulus codes could have several potential explanations, including the lack of repeated sampling of target categories[9,19], unlike the current experiment where the same Target rule is revisited many times, and that target categories in prior experiments were the image categories themselves (e.g. the target category might have been "person" or "plane") rather than a set of high-level attributes that need to be extracted and may or may not cause the image category to cross the target classification boundary. The increased task complexity with respect to categorizing the image category may be necessary to encourage the formation of nonlinear

stimulus x target representations that could facilitate solving this migrating target-classification problem. Further exploration of these neurons, and the neurons in VTC which do exhibit some amount of context-modulation as demonstrated in Chapter 2, constitutes an immediate next step in the study of stimulus representations in the human brain and how they interact with task context to generate decisions.

The final line of inquiry considered here relates to the bilingual patient and the generalizability of those findings. While exciting, they must be interpreted with caution due to the pathology unique to this patient. This patient had significant portions of both frontal lobes resected in previous surgeries, and exhibited very long instruction encoding times and response times as a result. The reduction in trial count, removal of the Motor rule, and extension of stimulus presentation times to 2s were all necessary measures to facilitate the completion of this task by the patient. Nevertheless, this patient was able to complete hundreds of trials of this experiment with high accuracy across both English and Spanish variants. While I believe it unlikely that the baseline context representation adopting the same geometry in both languages to be a feature of this patient's pathology, repeated testing in at least one more brain is needed to increase confidence in the validity of these findings. The situation is complicated further by the fact that, while constant at the level of the population, it appears that individual FC neurons shift their tuning to task variables across the two flanking sessions (data not shown). As an internal control, MTL and VTC neurons do not change stimulus tuning upon restarting the experiment with a different language. Additional control experiments that involved flanking recording sessions recorded in another patient suggest, albeit weakly, that this phenomenon was not unique to the bilingual patient, and that single neurons tuned to persistently-encoded task variables might reorganize their tuning whenever a new experiment is commenced, even if that experiment is identical to the experiment that was just completed. How and if the neurons do indeed change their tuning over task rules while preserving population-level geometry is yet another question that requires more data to be answered. In short, additional neurons are required to get to the bottom of all this nonsense (perhaps on the part of the researcher).

Nevertheless, the analysis and findings presented here constitute a first step in uncovering the logic by which the brain organizes its representation of arbitrarily complex tasks, specified by natural language, that enable zero-shot generalization.

Figures:

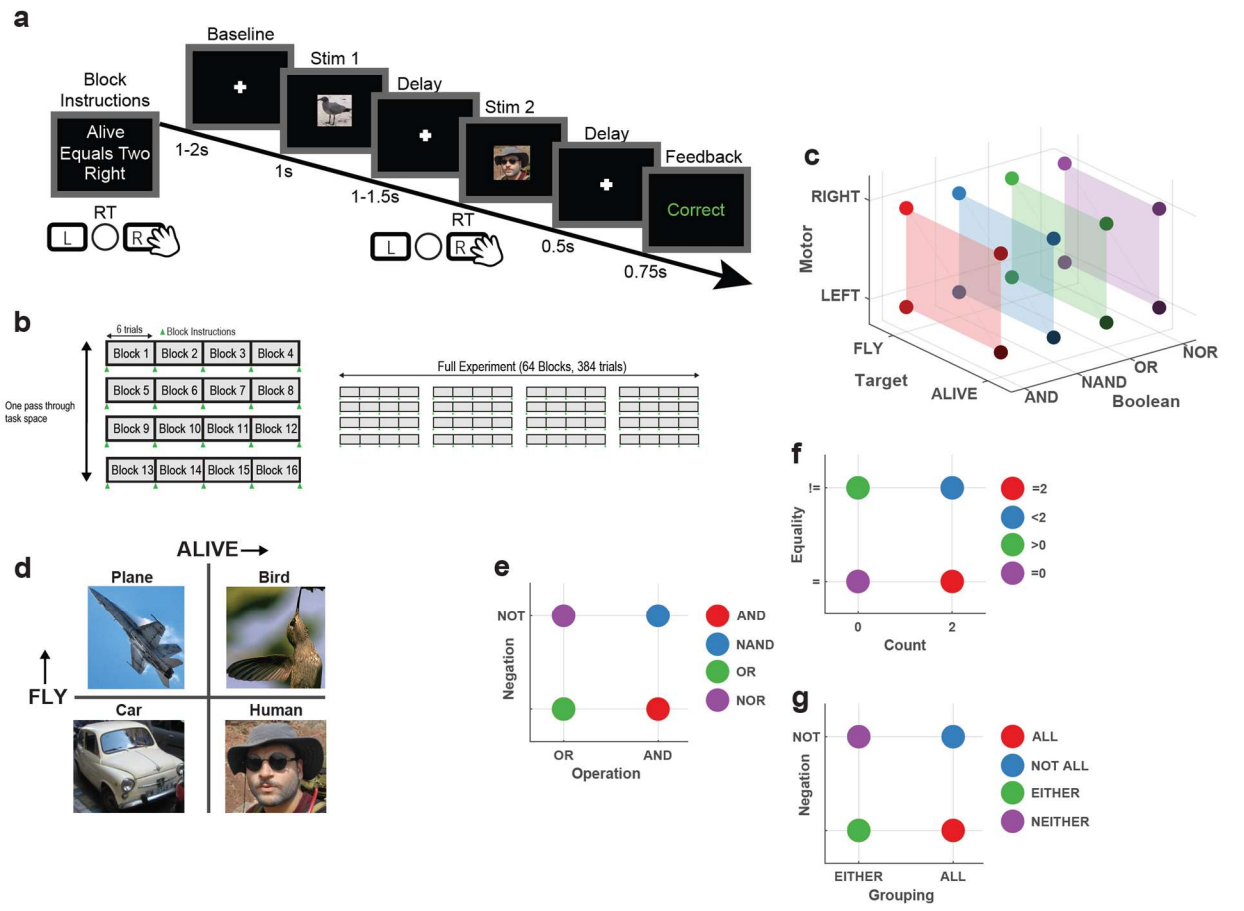

Figure 4.5. Design of a compositionally structured task that induces a large task space. **(a)** Illustration of trial structure. Instruction screens consisted of three task rules that uniquely specified one of 16 possible task contexts (Top Row/Target Rule/2 Levels, Middle Row/Boolean Rule/4 Levels, Bottom Row/Motor Rule/2 Levels). Patients responded to begin each block. Trials consisted of a pre-stimulus baseline (baseline) followed by the first stimulus presentation (stim 1, 1s), a brief, jittered delay (1-1.5s), and then the second stimulus (stim 2, RT) at which point the patients provided a response. Trial-level feedback was provided after another brief delay. **(b)** Illustration of the block structure. Blocks consisted of 6 trials, and task contexts were presented once for every traversal of task space, which was a sequence of 16 blocks. Each experiment consisted of 4 full passes of task space **(c)** Illustration of task state space structure. Individual points here represent unique task contexts specified by combinations of Target, Boolean, and Motor rules, unlike similar schematics in previous chapters where points corresponded to trial states specified by both context-level and stimulus-level variables. The top right corner of the red square (ALIVE, AND, RIGHT) corresponds to the task context shown on the instruction screen in **(a)**. Linear ordering of rules along the Boolean dimension is arbitrary, for visualization purposes, and does not represent an expected relational structure in neural state space. **(d)** Schematic of image categories encountered throughout the task and their Target rule membership. **(e-g)** Mapping of Boolean rule to operations that are easily comprehensible by patients. **(e)** Ground truth relational structure of Boolean operators shown in a fictitious rule-construction space. **(f)** Boolean operations can be mapped onto number counting

and equality/inequality evaluations that are readily comprehensible, but with a relational structure that differs from the ground-truth rule construction space shown in **(e)**. Note: green and purple (OR, >0 and NOR, =0 respectively) have switched positions in **(f)** compared to **(e)**. **(g)** Another realization of the same Boolean task rules using natural language prompts. Note that the relational structure of this task construction space is identical to **(e)**.

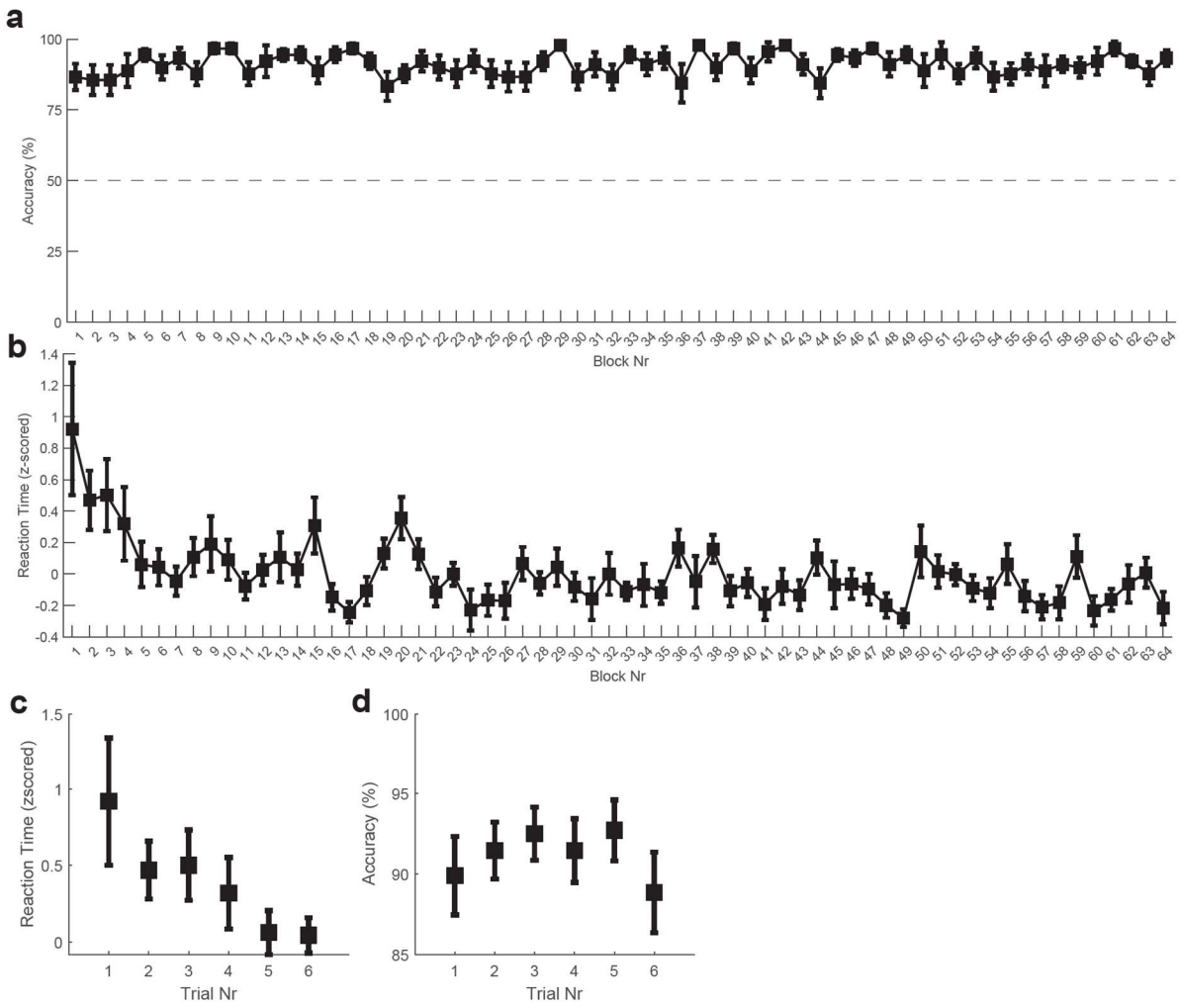

Figure 4.6. Patient performance on compositional task.
(a) Task accuracy is averaged over all trials-in-block and shown across all blocks. Points and error bars correspond to mean and s.e.m. performed over sessions. Data from 7 patients and 15 sessions is shown. Horizontal dashed line indicates chance performance (50%). **(b)** Same as **(a)**, but for trial-level reaction time z-scored within-session then averaged across sessions. Reaction time **(c)** and accuracy **(d)** also shown as a function of trial-in-block. Values are averaged for a given trial across all blocks, then mean and s.e.m. are reported across sessions.

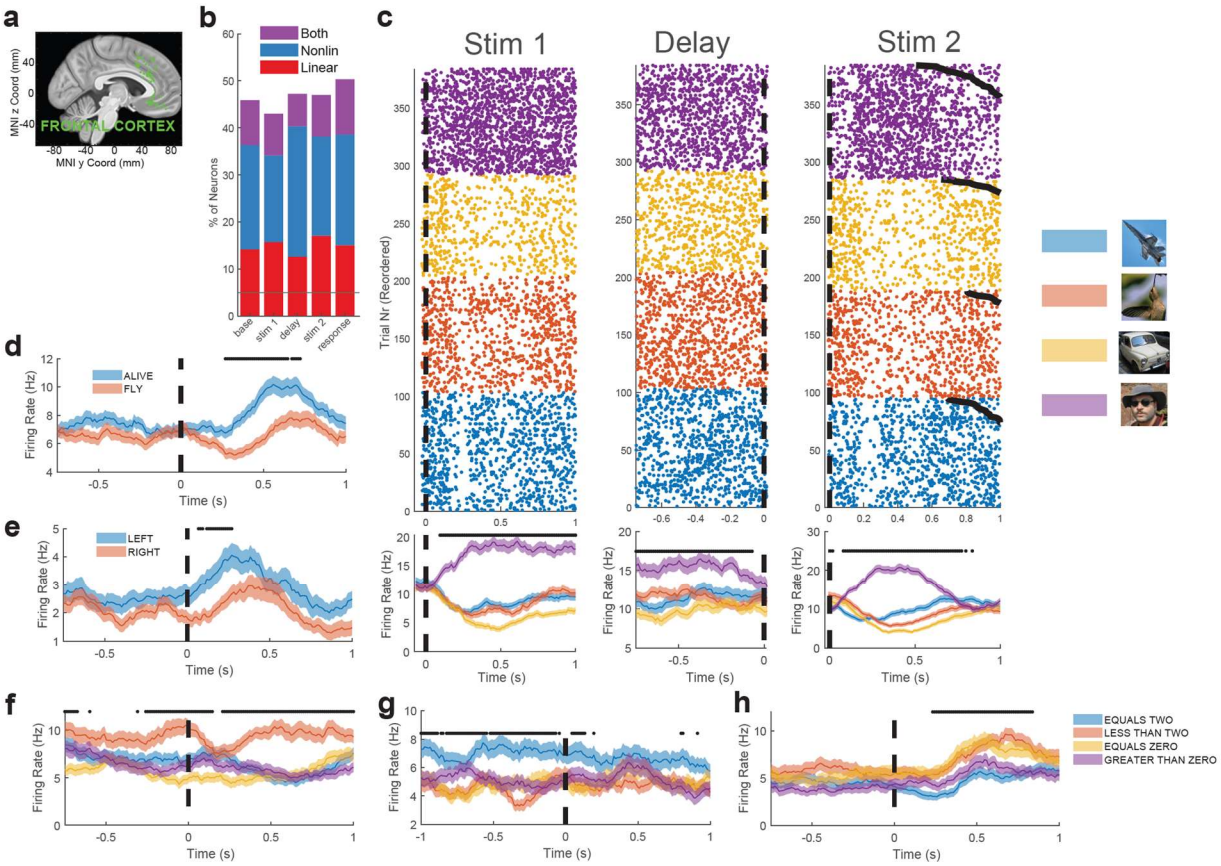

Figure 4.7. Single neuron responses to task rules and stimuli.
(a) Electrode locations. Each dot corresponds to a single microwire-bundle. Locations are shown on the same hemisphere for visualization purposes only. All electrodes from the preSMA, dACC, and vmPFC are colored green and collectively labeled as "Frontal Cortex". **(b)** Fraction of single units in frontal cortex exhibiting significant Main effects or interaction effects (3-way ANOVA with interactions, $p < 0.05$) to at least one of the task rule variables (Boolean, Target, Motor Rule) or to combinations of variables. A unit is linearly tuned if it has at least one significant main effect, and non-linearly tuned if it has at least one significant interaction term in the ANOVA model. Horizontal dashed line indicates chance (5%). **(c)** Example rasters and PSTHs for a single category-tuned neuron with human faces as its preferred category. Stim 1 (left) and Delay (middle) column trials are organized according to the category of Stim 1. Stim 2 (right) trials are organized according to the category of Stim 2. Stim 1 is aligned to stim 1 onset. Delay and stim 2 plots are both aligned to stim 2 onset. All plotting conventions are identical to previous rasters and PSTHs. **(d-f)** PSTHs of other example neurons exhibiting task rule tuning during different trial periods. **(d)** Neuron tuned to Target rule during stim 2. **(e)** Neuron tuned to Motor rule during stim 2. **(f)** Neuron tuned to Boolean rule during stim 2. **(g)** Neuron tuned to Boolean rule during stim 1. **(h)** Neuron tuned to Boolean rule during stim 2.

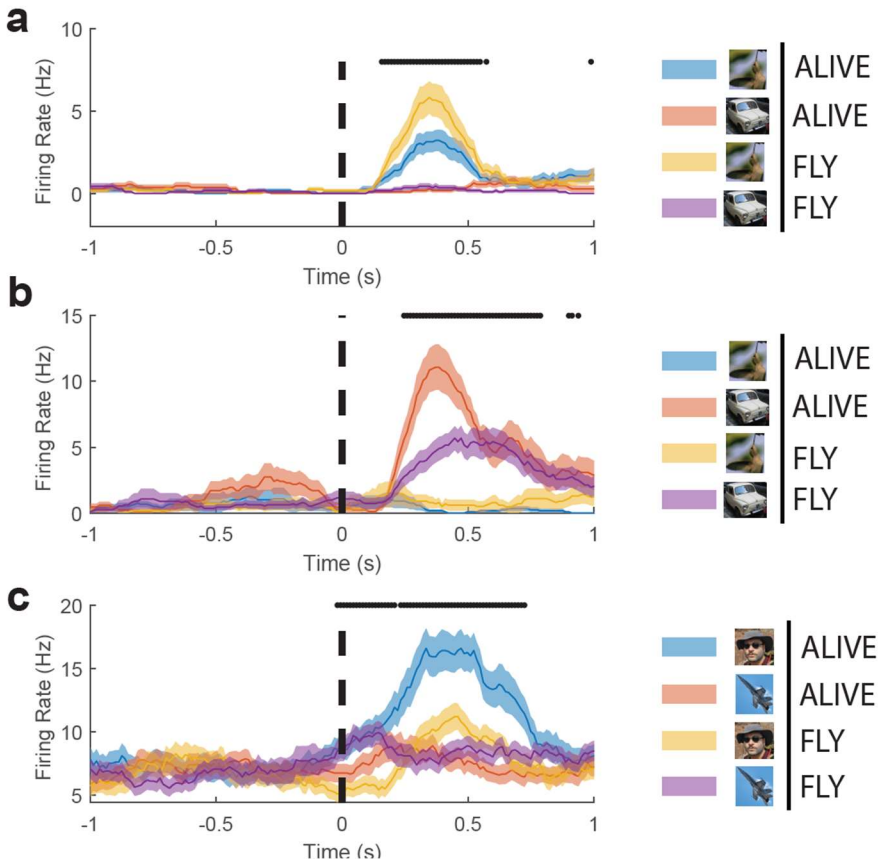

Figure 4.8. Example hippocampal neurons exhibiting conditional category responses. PSTHs of example hippocampal neurons exhibiting image category responses that are modulated by the current target rule in the task context. **(a)** The neuron is selectively responsive to birds, and is differentially modulated by categorizing birds as ALIVE (blue) or FLY (yellow). **(b)** Another neuron that differentially responds to cars under the same target rule manipulation: ALIVE (orange) and FLY (purple). Note: for **(a)** and **(b)**, the image category does not cross the target rule classification boundary (i.e. birds are ALIVE and can FLY, cars are not ALIVE and cannot FLY). **(c)** A third neuron that differentially responds to human faces conditioned on the target rule (ALIVE, blue and FLY, yellow). In this case, the image category does cross the target rule classification boundary (i.e. humans are ALIVE, but cannot FLY). All PSTHs shown here are from the stim 2 presentation trial period. All plotting conventions are identical to previous PSTHs.

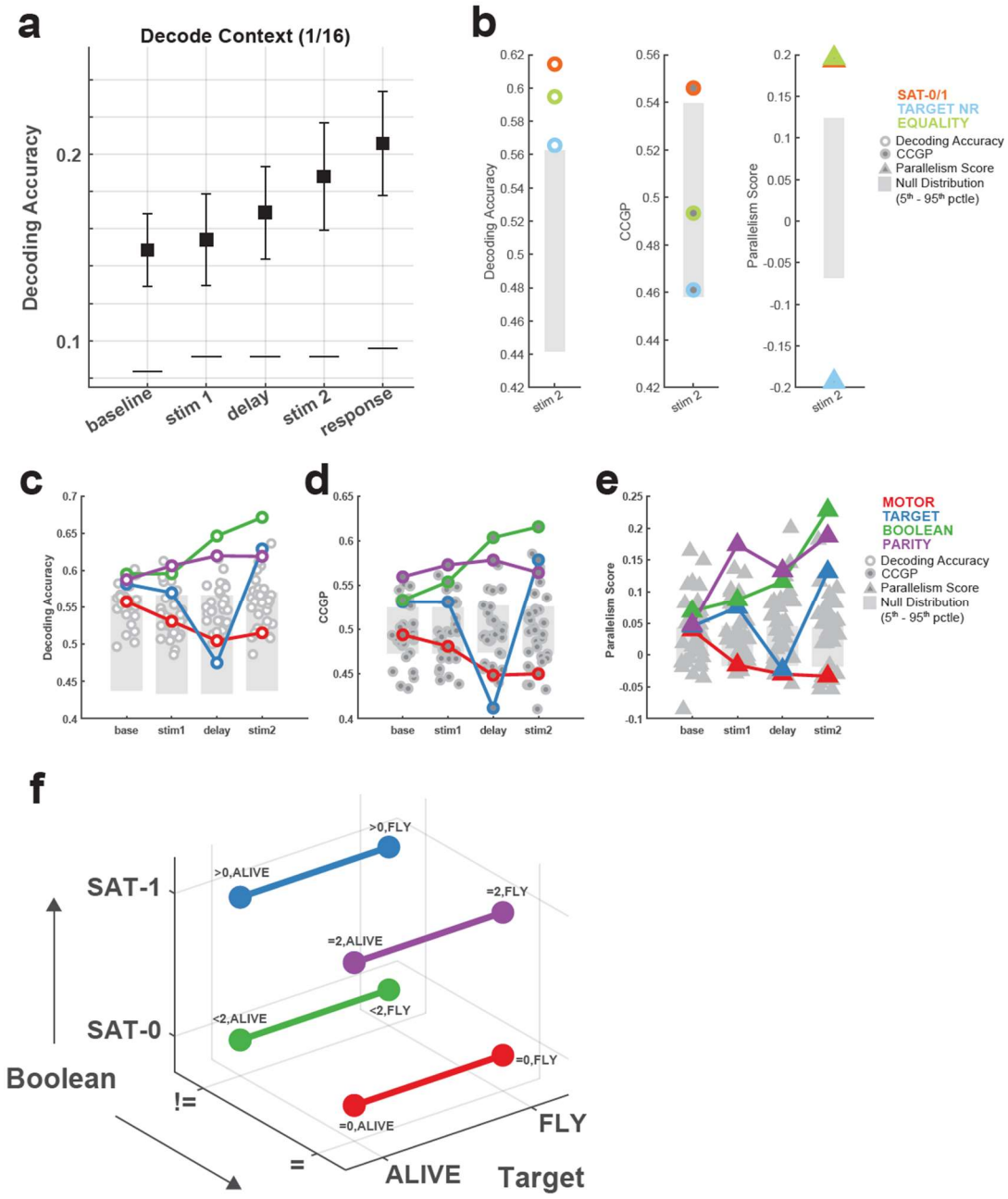

Figure 4.9. The geometry of frontal cortical task rule representations is compositional.
(a) Decoding of current task context. Points and error bars indicate mean and s.e.m. over single-trial bootstrap resampling per-context as described in the methods. Horizontal black lines indicate 95th pctle of shuffle null distribution. Chance performance is 0.0625 (1/16 possible task contexts). Decoding accuracy is reported for different trial time periods indicated on the x-axis. **(b)** Geometry of Boolean task rules analyzed using Boolean rule dichotomies only during the stim 2 presentation period. Decoding accuracy (left), CCGP (middle), and parallelism score (right) are computed for all three such dichotomies using frontal cortical neurons. Dichotomies are color coded by meaning. Gray shading indicates 5th-95th percentile of shuffle-null distribution. **(c-e)** Balanced dichotomy analysis performed over all task rules, with decoding accuracy **(c)**, CCGP **(d)**, and parallelism score

(e) shown for all trial periods before the response in trial. All plotting conventions identical to those used in Chapter 2. Note: Target rule (Alive/Fly, blue) and Boolean task rule (green) are simultaneously encoded and disentangled in the stim 2 period. **(f)** Synthesis of geometric findings from **(b-e).** Target category rule (Alive/Fly) and Boolean task rules are disentangled by analysis in **(c-e)**, and the Boolean task rules are organized in a 2-dimensional space with SAT and Equality/Inequality disentangled organizing variables within the Boolean task rule subspace.

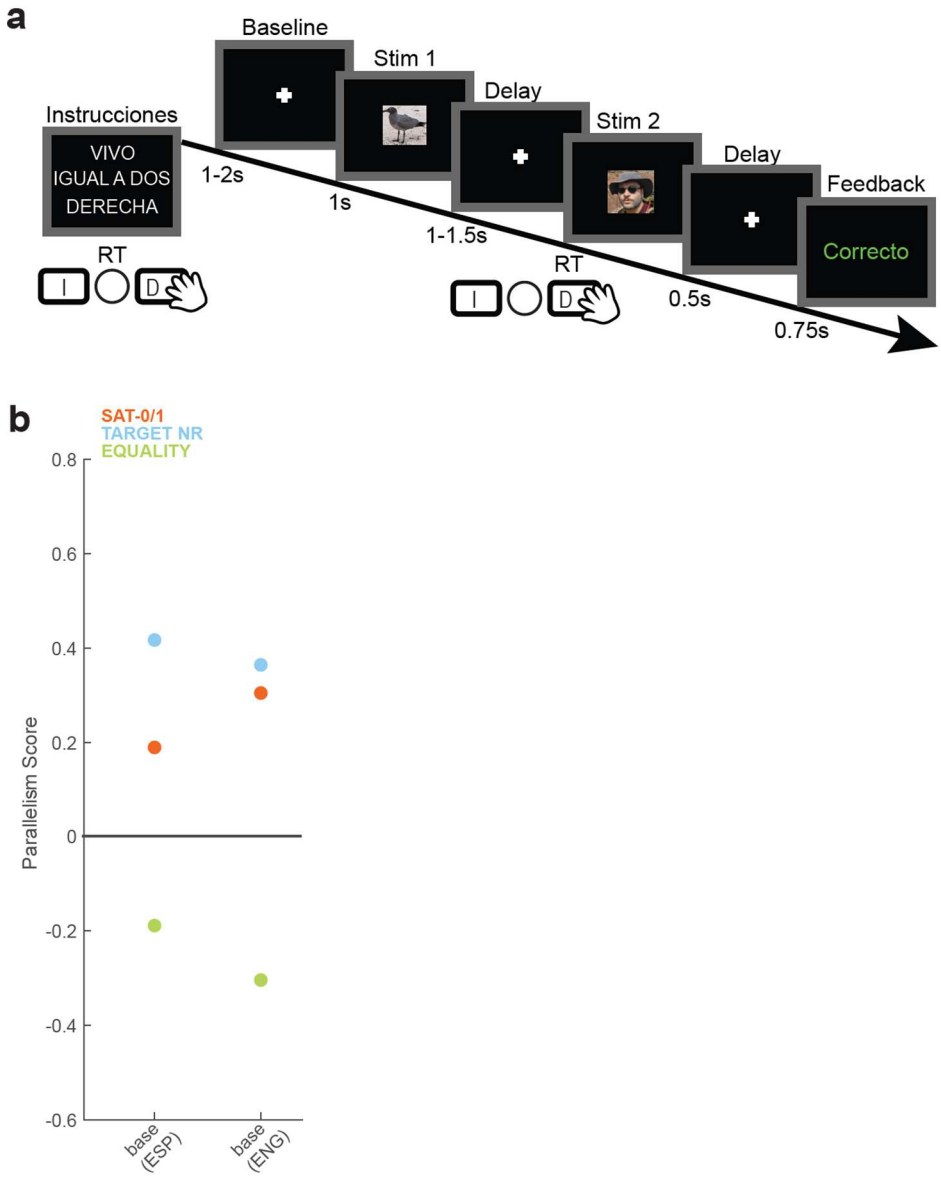

Figure 4.10. Geometry of rule representation is language-invariant in one bilingual patient. **(a)** Identical task structure to that administered in English, but with all text (including instructions and verbal facilitation from experimenter) provided in Spanish for one bilingual patient. Back-to-back recording of the Spanish and English variant were performed, ensuring retention of the same population of neurons to facilitate comparison. **(b)** Parallelism score computed for Boolean task rule dichotomies. Parallelism for each dichotomy (each point) during the baseline period is shown for the Spanish session (left) and English session (right). Horizontal black line marks 0 parallelism, with dichotomies above the line indicating disentangling of those variables.

Bibliography:

1.Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**, 3–71 (1988).

2.Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences* **40**, e253 (2017).

3.Johnson, K. & Journal of Philosophy, Inc. On the Systematicity of Language and Thought: *Journal of Philosophy* **101**, 111–139 (2004).

4.Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).

5.Courellis, H. S. *et al.* Abstract representations emerge in human hippocampal neurons during inference behavior. 2023.11.10.566490 Preprint at https://doi.org/10.1101/2023.11.10.566490 (2023).

6.Ito, T. & Murray, J. D. Multitask representations in the human cortex transform along a sensory-to-motor hierarchy. *Nat Neurosci* **26**, 306–315 (2023).

7.Ito, T. *et al.* Compositional generalization through abstract representations in human and artificial neural networks. Preprint at https://doi.org/10.48550/arXiv.2209.07431 (2022).

8.Fu, Z. *et al.* The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* **376**, eabm9922 (2022).

9.Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by human medial-frontal cortex. *Science* **368**, eaba3313 (2020).

10.Guan, C. *et al.* Decoding and geometry of ten finger movements in human posterior parietal cortex and motor cortex. *J. Neural Eng.* **20**, 036020 (2023).

11.Zhang, C. Y. *et al.* Partially Mixed Selectivity in Human Posterior Parietal Association Cortex. *Neuron* **95**, 697-708.e4 (2017).

12.Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]* (2014).

13.Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at https://doi.org/10.48550/arXiv.1301.3781 (2013).

14.Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).

15.Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci* **22**, 297–306 (2019).

16.Fried, I. *et al.* Cerebral microdialysis combined with single-neuron and electroencephalographic recording in neurosurgical patients. Technical note. *J Neurosurg* **91**, 697–705 (1999).

17.Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J Neurosci Methods* **154**, 204–224 (2006).

18.Ehrlich, D. B. & Murray, J. D. Geometry of neural computation unifies working memory and planning. *Proceedings of the National Academy of Sciences* **119**, e2115610119 (2022).

19.Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat Neurosci* **18**, 1041–1050 (2015).

Chapter 5

Parting Thoughts

As those before me have stated, it is quite common that, after a decade or more of walking along a straight line, a traveler will have grown to absolutely despise the straightness of that line. The purpose of this discussion is to summarily complain about the line's straightness.

First and foremost, I believe that discovery science performed to deepen understanding of neural computation in the human brain as it relates to behavior will continue to be of utmost importance in the coming years. The fund of knowledge available on the inner workings of the brain, let alone the suite of tools to productively intervene to restore or augment function is laughably poor. The engineering approach of principles-based construction of physical solutions to problems only works in the context of an abundance of robust, rigorously tested principles that accurately describe and predict, at some spatial and temporal scale of description, the behavior of complex systems. We are nowhere close to this level of understanding for the human brain, particularly in the realm of complex cognition.

So, if we are not close, have I at least succeeded, through this work, in bringing us a bit closer? Let us take stock of our results over the last several chapters of content in concise format.
Abstraction – Codifying abstract variables in the environment explicitly in one's hippocampal representation enables the use of those variables to rapidly update behavior. This can be achieved through experience or through instruction.
Checkerboard – Disentangling task representations from the passage of time facilitates temporally extended, persistent behavior. Which specific regions do this temporally-disentangled representing might vary as a function of task demands.
Rapid Switch – Neurons not being properly configured and exhibiting inertia after reading instructions can predict one's switch cost with 1.59 bits of precision, despite having arbitrary amounts of time to prepare.
Compositional – The compositional structure of arbitrary language-specified task rules is inherited by frontal cortical neurons during zero-shot generalization, though there appear to be limits to this structure when the geometry of language conflicts with the geometry of state-action relationships.

Many watts of electrical power, neurons in the brains of patients as well as my own, US dollars, and of course years of my life were sacrificed to reduce uncertainty about the world such that the above sentences could accurately be placed into print. However, the fact of the matter is that after all of this experimentation and analysis, this information can still not be productively used to improve the life of any one of the patients who were so generous to lend me their time and their brains. This fact is hardly surprising considering our understanding of human brain computation in complex, cognitive, multi-task environments is quite literally only a few years old. The time scale on which this knowledge will begin to yield tangible benefit to patients on a scale beyond one-of experimental implants should probably still be measured in decades, and there are many limitations which need to be overcome both in practice and in principle before that goal is realized. A veritable lifetime of effort is needed to realize that goal, and that is, in fact, the plan.