Complementary roles of Theta and Alpha in Multi-Item Associations

Nicholas Ketz[1*], Tim Curran[1] and Randall C. O'Reilly[1]

1 Department of Psychology and Neuroscience, University of Colorado Boulder
December 16, 2019

* Nicholas Ketz current affiliation is: Information and Systems Sciences Lab, HRL Laboratories, Malibu, CA 90265

Acknowledgements

The authors would like to thank research assistants at Boulder, especially Chris Bird, Krystin Corby, Shaina Martis, as well as other undergraduate volunteers for help in collecting the data used in the EEG experiments. We would also like to thank faculty at Boulder, specifically McKell Carter, Anu Sharma and Erik Willcutt for helpful feedback during manuscript preparation. The authors have no known conflicts of interest with this work. This work was partially funded by NSF Research Fellowship grant #DGE-0707432. Simulation code available at *https://github.com/nickketz/deev_paper/tree/master/Docker*.

Abstract

Through the use of neural network models of the hippocampus and surrounding cortex this work proposes a framework for understanding how time frequency signatures measured at the scalp can be used to track declarative memory processes, and make quantitative predictions about how information stored in long term memory is altered by these processes. The fundamental hypothesis explored here is that neural oscillations in the theta (3-8 Hz), and alpha (8-12 Hz) frequency bands can be tied to specific neural mechanisms supporting declarative memory, and that these oscillatory signatures can be tracked in human scalp EEG recordings to predict behavioral changes in declarative memory processes. Specifically, oscillatory power in the theta band positively correlates with the how much information the hippocampus is reactivating for a given retrieval event, and that power in the alpha band positively correlates with how much information is being inhibited from being retrieved. We first explore this idea through computational simulations of thalamo-cortical-hippocampal interactions which make a series of predictions tested in a human EEG experiment using the Dependent Events paradigm. Results largely support the idea that thalamically mediated changes in alpha power during episodic memory encoding and retrieval are related to the flow of information between the neocortex and hippocampus, and that the bandwidth of this information flow tracks with theta power. Further, this model suggest the encoding and retrieval processes are inter-dependent and that attentional modulation of what gets retrieved from long term memory subsequently impacts the information that is encoded into the underlying representation.

Keywords

Episodic Memory, Brain Waves, Neural Networks, Thalamus, Hippocampus

1 Introduction

It is generally accepted that long-term memory encoding and retrieval are not isolated functions of any one brain area (Gabrieli, 1998; Prince, Daselaar, & Cabeza, 2005). However, there are many challenges in moving beyond the traditional studies focusing on individual regions (e.g., concentrating on the prominent role of the hippocampus), to obtain insight into the nature of network-level dynamics. An increasing body of work support the idea that oscillatory dynamics in multiple brain regions are associated with different aspects of memory encoding and retrieval (Nyhus & Curran, 2010; Hanslmayr, Staudigl, & Fellner, 2012; Klimesch, 2012), These studies have generally found a consistent relationship between several frequency bands and successful encoding or retrieval of experiences; namely that oscillatory power within the alpha (8 to 12Hz) and beta (13 to 30Hz) frequency bands generally decrease with the successful execution of these memory processes, while theta (3 to 8Hz) and gamma (30 to 100+Hz) increase in the regions directly involved in the memory operations. Why these relationships exist, and what they tell us about the larger functional organization of memory systems in the brain are still open questions.

One model suggests that thalamic circuits play a functional role in selectively gating communication between cortical regions by synchronizing them in various frequency bands (Sherman & Guillery, 2006; Saalmann & Kastner, 2011; N. A. Ketz, Jensen, & O'Reilly, 2015). These gating mechanisms, implemented by oscillatory synchronization, are central to system level brain function and ultimately affect cognitive behavior such as memory encoding and retrieval. Further, the particular oscillatory dynamics engaged by this selection process may provide indicators of which system-scale neural circuits are coordinating together to facilitate that particular form of memory processing.

A particular thalmo-cortical circuit of interest for memory processes involves the hippocampus and its associated subcortical network, including the anterior thalamus, and medial septum which are essential in the maintenance of theta oscillations (4 to 8 Hz) (Buzsaki, 2002). Theta oscillations have been shown to be critically involved in successful encoding and retrieval of episodic like memories in both humans and animals (Nyhus & Curran, 2010; Jones & Wilson, 2005; Fuentemilla, Barnes, Düzel, & Levine, 2014), and are believed to be primarily driven by the pacemaker cells within the medial septum. The anterior thalamus, however, provides the connectivity for the hippocampus to potentially modulate communication with the medial and orbital PFC through theta synchronization (Van der Werf, Jolles, Witter, & Uylings, 2003; Aggleton, Dumont, & Warburton, 2011). Thus, the theta circuit is important for the core functioning of the hippocampus during encoding and retrieval.

A complimentary thalamic circuit involves the pulvinar providing synchronization within the alpha (8 to 12 Hz) band between parietal and visual cortices with the medial, orbital and lateral PFC (Barbas, Henion, & Dermon, 1991; Kievit & Kuypers, 1977; Lopes da Silva, 1991). This circuit is important for bottom-up stimulus-driven encoding processes in memory (Jutras, Fries, & Buffalo, 2013), but it can also inhibit contradictory or interfering information during internally guided retrieval (Park et al., 2014). Thus, its overall contribution to memory can be mixed, and depends on the nature of the task and what kinds of memory processing is required.

Indeed, a recent spiking model of these oscillations called the Sync/deSync Model highlights the role of a synchronized hippocampus that accompanies a desynchronized neocortex during encoding and retrieval. This model in particular focuses on the role of decreased alpha power related to successful encoding of new information (Parish, Hanslmayr, & Bowman, 2018; Hanslmayr, Staresina, & Bowman, 2016). Similarly, recent empirical studies of human intracranial recordings find evidence to support the complimentary roles of hippocampal synchrony and neocortical desynchrony in successful encoding and retrieval (Griffiths et al., 2018; Staresina et al., 2016). In general these results find a decrease in neocortical alpha power and an increase in hippocampal gamma power that coincide with successful encoding and retrieval of declarative memories.

This accumulated evidence suggests a relationship between theta and alpha frequency bands and long-term memory systems as constructed by the Complimentary Learning Systems theory (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly, Bhattacharyya, Howard, & Ketz, 2014). Namely, that theta power positively tracks successful engagement (encoding/retrieval) between hippocampus and neocortex, while alpha power negatively tracks with this process as active inhibition of cortex is released in response to the exchange of information between hippocampus and neocortex. This relationship is further supported by the connectivity of the thalamus with cortex such that a specific model for how information is routed into and out of these systems can be tied to empirical measures of neural oscillations. This work seeks to test this relationship through a series of computational simulations of hippocampal-thalamo-cortical interaction, and an accompanying EEG study in humans to compare against simulation results.

In general, our results show that our model can capture broad time frequency dynamics related to alpha and theta oscillations, and that it can be used to evaluate patterns of human EEG against an underlying theoretical model. This model provides explanatory power in understanding how thamalmo-cortical attentional mechanisms can enhance the binding between multi-item associations through reactivation, as well as diminish those associations when the underlying encoding structure does not promote reactivation.

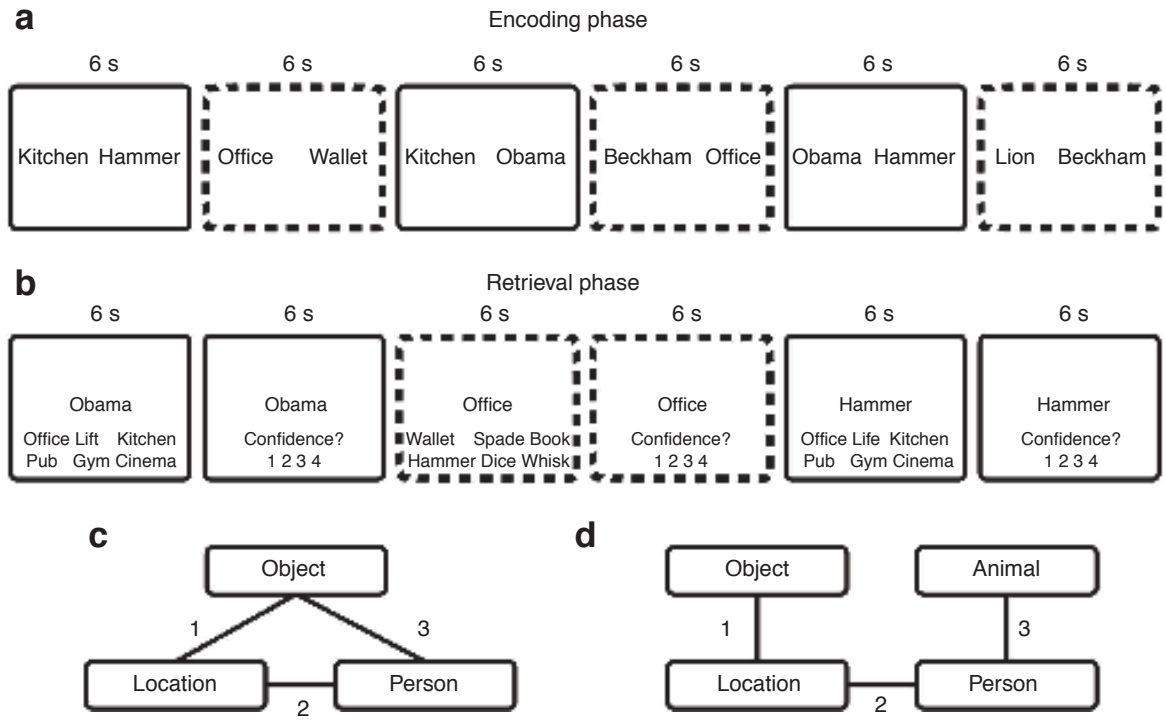

FIGURE 1: Dependent events behavioral paradigm, taken from Horner, Bisby, Bush, Lin, and Burgess (2015) (a) Encoding phase. Participants see multiple paired associates and imagine each pair 'interacting in a meaningful way as vividly as possible' for 6 s. Solid lines and dotted lines were not present, but highlight a closed-loop (solid lines) and open-loop (dotted lines) event. (b) Retrieval phase. Participants were presented with a single cue and required to retrieve one of the other elements from the same event from among five lures within 6 s. This was followed by a 1-4 confidence rating within 6 s. All 6 possible cue-target combinations are tested (c) The associative structure of closed-loop events, with example encoding order for the three pairwise associations (numbers 1-3). (d) The associative structure of open-loop events, with example encoding order for the three pairwise associations.

1.1 Dependent Events Paradigm

To test this inverse relationship between theta and alpha we choose a recent hippocampal dependent multi-item association study to base our computational and empirical investigations on (Horner, Bisby, Bush, Lin, & Burgess, 2015) (see Figure 1). In this study the authors used a previously developed paradigm (see Horner and Burgess (2013, 2014)) in which they manipulated the binding of a set of overlapping paired associates. Here two types of associative stimuli were studied, closed-loop and open-loop. In the 'closed-loop' events participants studied a set of paired associates on separate trials, e.g. Person-Location, Location-Object, and Object-Person, where each category refers to the same stimulus for a given event. This created a complete set of binding between all elements. In contrast, the 'open-loop' events incorporated a new element and left some relationships unstudied, e.g. Object-Location, Location-Person, Person-Animal. Each of the associations was later tested in a cued recognition task, which showed greater inter-event Dependency for the closed-loop compared to the open-loop associates. Dependency here is operationalized as the retrieval of *all* associated pairs from a given event, or none; whereas no Dependency would imply retrieval of some of the associated pairs but not all. The goal of this metric is to capture the strength of the binding between elements within a given event, where high Dependency indicates strong associations leading to retrieval of all items within a given event, and low Dependency indicates weak associations leading to some of the items being retrieved incorrectly. In contrast, a non-associative measure of accuracy can be made using an Independent model which calculates Dependency based on the assumption that each retrieval trial is statistically independent from the next, see Section 2.4 for more details.

In this paradigm, closed-loop events show significantly greater Dependency as compared to the Inde-

pendent model while the open-loop events showed no difference compared to its Independent model, implying that closed-loop events have a stronger between element binding as compared to open-loop events. The closed-loop condition also showed more cortical reactivation of non-target elements within an event (i.e. studied elements not currently the target of the recognition trial), and that this reactivation correlated with hippocampal activity. Results from fMRI studies such as Horner et al. (2015) provide spatially specific evidence for the hippocampus driving cortex to reinstate previous studied elements, but provide no evidence relating these processes to potentially related neural oscillations. A computational model tied to the neural oscillations of this phenomena can provide an opportunity to bring clarity to this relationship by testing if patterns in theta power can be predicted by a computational model of the hippocampus. Key aspects, however, relating to the focus of attention and alpha power are required to adequately model the paired associate and cued recalled components of the Dependent Events paradigm.

1.2 Theta-Phase Model

Building off our previous connectionist model of the hippocampus (N. Ketz, Morkonda, & O'Reilly, 2013), referred to here as the Theta-Phase model, we seek to incorporate the ability to attend to particular dimensions of a given input pattern, while not attending to others. Practically, this requires that each input be described on all possible stimulus dimensions, otherwise the hippocampal model will attempt to 'recall' the missing information for those missing dimensions. For example, one valid cortical representation experienced will contain information about size and shape but never in color, while another equally valid representation will contain information about color and shape, but never size. The neural substrate supporting these representations needs to be able to represent all three dimension (and potentially many others) simultaneously and independently, but also be able to understand the inherent bias in which dimensions are likely to covary together.

A simple and ecologically valid way to reconcile this discrepancy is a mechanism that would highlight which of the stimulus dimensions are informative, or relevant given the current input. This would be very similar to a selective attention mechanism. Building from the evidence in support of an alpha mediated thalamic gating mechanism, we will explore modulation of cortical activity via thalamo-cortical interactions in the alpha band as a mechanism for the Theta-Phase model to gate or modulate specific stimulus dimensions for targeted encoding and retrieval. The behavioral and physiological consequences of this mechanism will then be explored in a long-term memory paradigm that utilizes multi-element traces, and a bottom-up attentional signal to focus on specific stimulus dimensions.

1.2.1 Attention and Alpha Oscillations

Sherman and Guillery (2006) provide the most complete framework for how the flow of information can be guided from low level sensory regions into higher level cortical regions best suited for task specific processing. The simple yet compelling idea is that cortical-cortical projections convey the contentful output from a given cortical region, while the thalamo-cortical projections contain the gating information. Sensory information is allowed to cascade from one processing region to the next through cortical-cortical projections, provided that a selective thalamo-cortical gating signal alerts/allows the next cortical region to accept (or block) the incoming cortical-cortical signal.

The other critical aspect of this bottom-up attention mechanism is the connection to alpha oscillations. Historically, alpha power has been shown to correlate with the onset of sleep, which led to the suggestion that high alpha power indicates an idle or perhaps inhibited cortical region. The mechanistic connection between alpha power and thalamic gating is captured in the firing properties of the thalamic relay cells (TRC) which show two distinct modes of firing. The first shows oscillatory firing patterns acting like a pacemaker system in the alpha frequency range (approximately 10 Hz or 100ms periodicity), and is generally referred to as bursting. This bursting mode is believed to be engage when a given cortical region is in an idle state, i.e. when no active processing is occurring in the thalamus or the corresponding

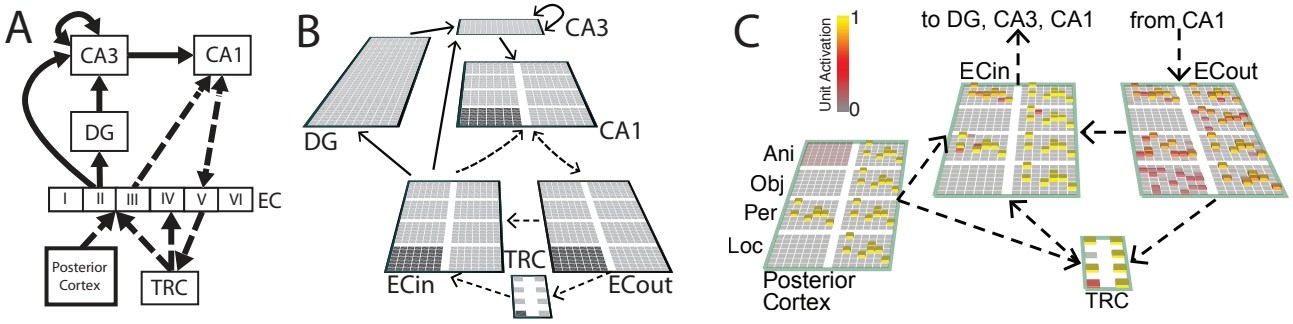

FIGURE 2: *(A)* Schematic of Theta-Phase model with thalamic gating. Posterior cortex connects with both layers 2/3 of Entorhinal cortex (EC) via cortical-cortical projects, but also to thalamic relay cells (TRC) via layer 5 bursting cells. These TRC units maintain their topography into EC layer 4 where they provide a gating signal for EC to accept or suppress the corresponding cortical-cortical projection. Topographic connectivity shown in dashed arrows, and distributed connectivity shown in solid. *(B)* Neural network showing TRC implementation as single units for a given cortical area's input and corresponding unit group in EC. This topography is kept consistent between TRC and EC unit groups. *(C)* Illustration of how TRC modulates network activity during cued-recognition. Here, a particular person (labeled Per) pattern is being used to cue recognition of an associated animal (labeled Ani) pattern, as indicated by the labels on posterior cortex unit groups. This input pattern from posterior cortex influences ECin and TRC cells releasing the thalamic suppression in unit groups where there is sufficient activation (i.e. the person and animal unit groups). The activation from ECin flows up to the hippocampus where missing information can be retrieved, and comes back to ECout through connections from CA1. ECout's activation influences both ECout and the TRC activations, using the hippocampally retrieved information to fill in the animal unit group as well as show partial recall of a non-target association of location pattern. The cortical input topography is maintained between EC and TRC where TRC activation is used to calculate the normalized gating signal shown in the corresponding EC units. Note the gating signal is not representation specific, but rather multiplied across each of the units of its target unit group.

cortex linked to that thalamic relay region. The other regime, referred to as tonic firing, shows no specific oscillatory firing pattern, and instead shows increased spontaneous firing in both the thalamus and corresponding cortical regions (Leresche, Lightowler, Soltesz, Jassik-Gerschenfeld, & Crunelli, 1991; Sherman & Guillery, 2006; Lopes da Silva, 1991). These dynamics have been studied extensively as a naturally occurring phenomena of thalamic interactions with cortex, however simulation studies have also shown the potential for these mechanisms to be used in a goal directed way (Hindriks & van Putten, 2013; Vijayan & Kopell, 2012; Lee, Whittington, & Kopell, 2013). The implications of these results are that cortical regions being driving by TRCs in bust mode will show more alpha power, and will process the incoming cortical-cortical information less veridically, while cortical regions being driven by tonic TRC firing will show a relative decrease in alpha power and process cortical-cortical information more completely. This is the foundation for the connection between alpha power and thalamic gating, such that increased alpha power is indicative of decreased cortical processing.

The detailed biology of this framework is well captured by Sherman and Guillery (2006), and relies on two critical components not yet implemented in the Theta-Phase model: 1. The thalamic relay cells (TRC) that pass the gating signal onto the next cortical region, and 2. The cortical neurons in layers 5 and 6 that send this gating signal to the next TRC, and trigger the release from inhibition for processing in associated cortex for that TRC. The following sections provides a high-level description of how these components are implemented, while supplementary info section **??** illustrates the detailed implementation into the LEABRA framework.

2 Computational Methods

In this work we start with the Theta-Phase model of the hippocampus and entorhinal cortex, as described in N. Ketz et al. (2013) and implemented in the Emergent framework (Aisa, Mingus, & O'Reilly, 2008) (see Figure 2). Briefly, this model is built upon a series of structural and functional hypotheses based on anatomical and physiological data, which have been captured in the complementary learning systems (CLS) model of the hippocampus (McClelland et al., 1995; Norman & O'Reilly, 2003). The Entorhinal Cortex (EC) in the model is assumed to be the cortical gateway to the hippocampus. This gateway feeds through the tri-synaptic pathway to the Dentate Gyrus (DG), CA3, and then to CA1. Similarly, there is a parallel connection through the mono-synaptic pathway from the EC to the CA1 (and back).

The tri-synaptic pathway connections are broadly diffuse, and support the conjunctive binding of various distributed pieces of information into an overall episodic memory representation in the CA3. The mono-synaptic pathway conversely is topologically organized, not diffuse. We capture this connectivity by organizing the simulated neurons in EC and CA1 into mutually interconnected *slots*, presumably encoding different separable elements/stimulus dimensions that loosely correspond to independent cortical regions converging on the EC (Witter, 2010). This slot architecture enables the mono-synaptic pathway to develop separable invertible pathways where a given EC input pattern can be encoded over a sparser representation in the corresponding CA1 slot, and this CA1 representation can in turn recover the full original EC slot pattern.

The learning dynamics within this model, as governed by the LEABRA algorithm, are dictated by the phases of a presumed theta oscillation, thus its short-hand name of 'Theta-Phase' used here. In short, during learning two error signals are derived within every period of a give theta oscillation. The first, during the trough of the theta oscillation, provides an error driven learning signal to the mono-synaptic pathway, learning the invertible mapping between EC and CA1 can be thought of as an encoding. The second, during the peak of the theta oscillation, provides a error driven learning signal on the tri-synaptic pathway, learning to recollect missing or incomplete information can be thought of as retrieval. To be clear, both pathways are providing an active learning signal to 'encode' information. Similarly, while the tri-synaptic pathway contributes to retrieval of associations between unit groups, the mono-synaptic pathway can contribute to retrieval within unit groups. The overall contributions of either pathway to learning in general has been illustrated in a recent study (**SchapiroEtAl17**).

Building on this model that inherently captures the hippocampal theta dynamics, we seek to add a modulatory attention signal that can capture the appropriate alpha dynamics derived from the thalamo-cortical principles described above. This following section reviews the specifics of that implementation, and the testing paradigm used to validate its performance.

2.1 Thalamic Modulation as Alpha

The effective implementation of the selective attention mechanism into the Theta-Phase model is done through a separate layer that reflects the thalamic relay cell (TRC) activity, as well as an attentional modulation value calculated from this TRC activity that scales the net input of downstream units. This can be seen implemented in the Theta-Phase model in Figure 2. Here connectivity between the posterior cortical inputs and Entorhinal Cortex (EC) also goes through a 'TRC' layer corresponding to the thalamic relay cells that then supply the attentional modulation signal. Each of the posterior cortical regions maintains topographic input into EC and the corresponding TRC, which is manifest in the model as separate unit groups as shown in Figure 2B. Each of these unit groups has a single TRC unit that calculates its modulatory signal as determined by the activation in the corresponding posterior cortical region and EC unit group. Finally, a single modulatory scalar value is calculated for each of the TRCs and sent to its corresponding EC unit group either inhibiting activation in that unit group or allowing normal processing to occur. For the purposes of this work the posterior cortex

is not simulated, but rather assumed to provide the constructed input patterns into the separate EC unit groups.

The detailed description of this modulator mechanism's implementation is described in supplementary info section **??**, however a coarse description and its functional implications are described here. The attentional modulation is effectively a normalized scaling factor based on the net input to a given TRC unit. A given TRC unit receives inputs with fixed weighted connections (i.e. no learning signal adjusts these weights), and its activation is used to calculate a single value that multiplicitively scales the net input of each of the units it is connected to. This attentional scaling value is simply the TRC activation normalized to range from some pre-determined minimum to a maximum of 1. This pre-determined minimum scaling factor can be thought of as the most inhibitory influence possible, and was loosely fit for these simulations to a value of 0.25 to match the anticipated accuracies across conditions. It should be noted this minimum scaling value did not vary across trials or conditions and therefore is unable to bias reactivation.

As can be seen in Figure 2C, the TRC influences the retrieval of information from the hippocampus by modulating which unit groups in EC are allowed to become active. If a sufficient input signal from posterior cortex, or a sufficient amount of retrieved information from the hippocampus pushes activity within a EC unit group to provide enough input to the corresponding TRC unit, it starts a positive feedback loop of patterns that are well learned. During cued recognition, shown in Figure 2C, the target unit group (animal in this case) from posterior cortex is given a uniform, low-level of activity sufficient to start this positive feedback process without actually providing any information about the ultimate pattern of activity. Retrieved information from the hippocampus can more easily fill in the specifics of this pattern because of the category of the target associate is provided by the recognition candidates (see Figure 1b). This is in contrast to non-target associates, which in this case is in the location unit group. Retrieved non-target information is being shown influencing ECout, and thereby starting to influence the corresponding TRC unit. If this retrieved pattern continues to cohere it may provide enough influence to allow ECin to reflect the same pattern and further strength the retrieval process by input that partial pattern back to the hippocampus.

The mechanism used here can be mapped more generally in Figure **??**

2.2 Model Validation

The Dependent Events paradigm (see Figure 1) was chosen to test the performance of the combined Theta-Phase model with attentional gating. This paradigm ideally captures behavioral components related to Complimentary Learning Systems ideas of pattern completion and separation, as well as components relating to selective attention over multi-element memory traces. These two aspects allow for the testing of the computational model to match known behavioral performance and relate the findings to predictions in the EEG time frequency domain. First the model with augmented attentional mechanisms is evaluated in detail within the dependent events paradigm. These results are then contrasted with results from the original Theta-Phase model without any attentional modulation.

Basing the testing procedure off Horner et al. (2015), the paradigm was implemented as close to the behavioral version as possible. The model learned a series of multi-element 'events', see Figure 1. Each event consisted of three or four elements (locations, people, objects and animals), plus 4 contextual elements that were consistent across presentations. Events were built up over three separate encoding trials (interleaved with encoding trials of other events). Each trial consisted of the presentation of one of the three possible pairwise associations from a given event. This paradigm allows for the building of 'events' with different associative structures of overlapping pairs: 'closed-loop', in which all event elements were presented paired with all other elements of the event; or 'open-loop', in which elements of an event were presented as a chain of overlapping pairs. Here, the term 'event' groups the set of overlapping associations encoded across separate paired associate trials.

The model's input patterns, described in more detail Section **??**, were created with a base vocabulary

of 100 7x7 unit group patterns with 7 units active and a minimum Hamming distance of 10 between each pattern. The composite input patterns used to train the model were generated from this set of vocabulary patterns, with 8 unit groups in each input pattern corresponding to 8 hypothetical posterior cortical regions. These input patterns were conceptualized to have 4 unit groups that each corresponded to one of the various element types of the paired associates used in the Dependent Events paradigm. For each event, one unit group had a unique pattern of activity that corresponded to 'location' representations, another for 'people', another for 'objects', and another for 'animals'. These patterns were consistent across presentations of a given event, i.e. the same unit group patterns were used across the 3 encoding trials, and 6 retrieval trials for each event. In this way selective attention during encoding and retrieval can be guided to these specific stimulus dimensions of the input pattern, as shown in Figure 2C where two element types are being presented and the corresponding attentional modulation signals are being shown inhibiting the non-attended dimensions. In addition to these 4 unit groups corresponding to the different event element types, 4 incidental unit groups also maintained the same representation across event presentations. This was conceptualized as other contextual elements of the event participants imagined, and likely aided in the model's ability to recall the associations (although simulations without these contextual elements were not explored).

2.3 Testing Procedure

The encoding phase was split into three mini-blocks. One pairwise association for each of the 36 events was presented during each of the three mini-blocks. Presentation order within each mini-block was randomized. For the closed-loop events, the order of presentation across the three mini-blocks for a specific event was either: (1) location-object/animal, location-person and person-object/animal, (2) person-object/animal, location-person and location-object/animal, (3) location-person, location-object/ animal, person-object/animal or (4) location-person, person-object/animal and location-object/animal. For open-loop events, it was either: (1) location-object, location-person, person-animal, (2) person-animal, location-person, location-object, (3) location-person, location-object, person-animal or (4) location-person, person-animal, location-object. A total of 8 random weight initializations were trained, where 2 runs on each of the 4 random ordering were used to make up the pool of results. The model was trained on 2 epochs (random permutations of all the trials within a mini-block) for each mini-block before starting training on the next mini-block. Thus, for the first and second mini-blocks, the closed-loop and open-loop conditions were identical aside from stimuli order, presenting a single pairwise association and then a second overlapping pairwise association. In contrast, the third mini-block either formed an all-to-all associative structure in the closed-loop condition or an associative chain in the open-loop condition as shown in Figure 1.

At retrieval, the model was tested on each encoded pairwise association for each event in both directions. In this way, each event was tested across six separate retrieval trials. In the behavioral paradigm each retrieval trial was a six-forced choice recognition test with a particular cue element type and potential 6 targets all from the same element type (for example, cue location, retrieve the associated person among five people from other events). However, in the model, this was captured by providing an activation of 0.1429 to all units in the target element dimension, providing a similar net activation compared to the complete patterns (i.e. $49 * 0.1429 = 7$), and thus releasing the thalamic suppression to allow processing in the target dimension. Note that a single cue pattern, and non-informative target dimension activation were presented on all retrieval trials with the task being to reactivate the original paired associate. This reactivation success was measured by the binary Name Error metric described in Section **??**, i.e. either correct reactivation or not for a given cue-target pair. Thus, as with encoding, the closed-loop and open-loop conditions were exactly matched in terms of stimuli and task demands, differing only in the potential occurrence of incidental reactivation (that is, different levels of pattern completion).

Two parameters were loosely fit to match retrieval performance from previous behavioral results of the Dependent Events paradigm (Horner et al., 2015; Horner & Burgess, 2014). The first was the number of training repetitions (referred to as 'epochs') for each mini-block of encoding. The second was the

Cue_{A_i}	Hit_{C_i}	Miss_{C_i}
Hit_{B_i}	$H_{A_iB_i} \wedge H_{A_iC_i}$	$H_{A_iB_i} \wedge M_{A_iC_i}$
Miss_{B_i}	$M_{A_iB_i} \wedge H_{A_iC_i}$	$M_{A_iB_i} \wedge M_{A_iC_i}$

Targ_{A_i}	Hit_{C_i}	Miss_{C_i}
Hit_{B_i}	$H_{B_iA_i} \wedge H_{C_iA_i}$	$H_{B_iA_i} \wedge M_{C_iA_i}$
Miss_{B_i}	$M_{B_iA_i} \wedge H_{C_iA_i}$	$M_{B_iA_i} \wedge M_{C_iA_i}$

TABLE 1: Contingency tables for a given event i focused on element A. $H_{A_iB_i}$ implies a correct retrieval with cue element A and target element B from event i, while $M_{A_iB_i}$ implies an incorrect retrieval with the same cue and target, $\wedge$ implies a logical and.

Cue_A	Hit_C	Miss_C
Hit_B	$P_{AB}P_{AC}$	$P_{AB}(1-P_{AC})$
Miss_B	$(1-P_{AB})P_{AC}$	$(1-P_{AB})(1-P_{AC})$

Targ_A	Hit_C	Miss_C
Hit_B	$P_{BA}P_{CA}$	$P_{BA}(1-P_{CA})$
Miss_B	$(1-P_{BA})P_{CA}$	$(1-P_{BA})(1-P_{CA})$

TABLE 2: Independent model contingency table focused on element A across all events. P_{AB} implies percent correct across all retrieval trials where element category A is the cue and element category B is the target.

minimum amount of attentional modulation, i.e. when the thalamic suppression is fully engaged how much activation is allow to get through, see Section 2.1 and Appendix **??** for more details. By coarsely comparing results from Name Error determined retrieval accuracy in the network, shown in Figure 4, to the known behavioral results from Horner et al. (2015), Horner and Burgess (2014), the two parameters were fit (2 epochs of training per mini-block, and a minimum activation of 0.25) to loosely match network performance with human behavior.

2.4 Measure of Event Dependency

Based on Horner et al. (2015), dependency between event elements was assessed by constructing contingency tables for successfully retrieving, on separate trials, the two associated elements (for example, person and object) when cued by the remaining element (for example, location), as well as for retrieving a particular target element (for example, location) when cued by the other two elements (for example, person and object), see Table 1 for example contingency tables. Averaging across events within conditions, this resulted in six 2x2 tables per model run per condition with the on-diagonal elements reflecting the Dependency for a given cue or target element. These were then averaged across the 4 contingency tables with shared associations across both open and closed loop conditions (i.e. location and object cue and target tables) to get a single measure of dependency per condition.

To accurately compare condition differences in Dependency between open and closed-loop events its necessary to control for accuracy differences. For example, if all paired associates are correctly retrieved it's unclear if there is inherent Dependency between event elements or if each association was memorized independently. To address this, an 'Independent' model that reflects the expected Dependency irrespective of the underlying event structure can be calculated to compare with the Dependency inherent in the event structures. For a given model random weight initialization, the mean proportion of correct retrievals for a given element type, say B, (over all events) when cued by a given element type, A , can be denoted as P_{AB}. This accuracy can be used to create contingency tables for an Independent model where the probability of correctly or incorrectly retrieving two associations is simply the product of the mean probability for each association (i.e., the proportion of events for which both B and C were correctly retrieved when cued by A would be $P_{AB}P_{AC}$, while the proportion where B was correctly retrieved but C was not would be $P_{AB}(1-P_{AC})$). A example contingency tables for the Independent model are show in Table 2.

3 Behavioral and EEG Methods

3.1 Participants

49 University of Colorado undergraduates participated in the experiment and received payment of $15 per hour (ages 18-29, mean=19.76; 30 male, 19 female). All participants were right handed, had normal or corrected-to-normal vision, and all but 5 were native english speakers. Informed consent was obtained from each participant, and the study conformed to the Human Research Committee guidelines. Two participants were removed due to equipment malfunction, nineteen were removed due excessively noisy EEG data such that they had less then 20 trials in any experimental condition. This excessive number of participants removed based on low trial counts is likely due to two factors: first is the paradigm requires participants to move their eyes, causing more EEG artifacts then would be expected otherwise, and second is the sub-selection of trials by correctness and confidence as discussed in Section **??**.

3.2 Materials

Experimental stimuli, originally adapted from Horner et al. (2015), were 252 individual words split into 4 categories with 72 locations, 72 people, 54 objects, and 54 animals. 72 individual 'events' were generated for each participant, half of which were closed-loop events consisting of three elements (e.g. location-person-object), and half were open-loop events consisting of 4 elements. Half of the closed-loop events consisted of location-person-object triplets and half consisted of location-person-animal triplets. These 72 events were equally split into two blocks balancing across all event types, and the subset of words used to generate events in block 1 was the same across participants. An example set of events for block 1 is shown in Appendix **??**.

Stimuli were presented on a 17-in flat-panel display with a resolution of 1024 × 768 (60 Hz frame rate) placed approximately 1 m in front of the participants. Words were presented in 30 point Geneva font. All portions of the display not occupied by stimuli or text were filled with white pixels. The experiment was programmed in MATLAB (versions R2012b and R2014a; The MathWorks, Inc., Natick, MA) and was presented using Psychtoolbox (Brainard, 1997). Mixed effect statistical analysis was done using R 3.2.1, and the lme4 package version 1.11-11 (Bates, Mächler, Bolker, & Walker, 2014). Significance values for mixed effect models were estimated using the Satterthwaite approximation for degrees of freedom, using the lmerTest package version 1.0.

3.3 Design

The experiment consisted of 2 blocks of 2 experimental phases (as shown in Figure 1): paired associate encoding, and cued recognition. Each session, including application of the electrode net and performing the task, lasted approximately 2 hours. Stimuli were subdivided into 2 blocks such that there were 18 unique open-loop events per block and 18 closed-loop events per block. Each of the open-loop events consisted of a unique location-person-object-animal quadruplet of words, and half of the closed-loop events consisted of a unique location-person-object triplet of words and the other consisted of a unique location-person-animal triplet of words.

Each of these events were presented in pairwise association during the paired associate encoding phase. The encoding phase was split into three mini-blocks such that one pairwise association for each of the 36 events was presented during each mini-block. Presentation order of events within each mini-block was randomized. Presentation of association pairings across mini-blocks for a specific event was done in 1 of 4 ways dictated by participant number (i.e. participant number *mod* 4 determined which order to use): for closed-loop events the ordering was: (1) location-object/animal, location-person and person-object/animal, (2) person-object/animal, location-person and location-object/animal, (3) location-person, location-object/ animal, person-object/animal or (4) location-person, person-object/animal

and location-object/animal. For open-loop events, it was: (1) location-object, location-person, person-animal, (2) person-animal, location-person, location-object, (3) location-person, location-object, person-animal or (4) location-person, person-animal, location-object. Note, that for the first and second mini-blocks, the closed-loop and open-loop conditions were identical, presenting a single pairwise association and then a second overlapping pairwise association. The third encoding trial either formed an all-to-all associative structure in the closed-loop condition or an associative chain in the open-loop condition, as illustrated in Figure 1.

3.4 Procedure

Experimental procedures for the Dependent Events paradigm are based on Horner et al. (2015), as well as the neural network simulations. This paradigm was modified as follows to accommodate EEG data collection.

An electrode net was applied to each participant's head, and the session began with the experimenter reading the instructions including details of EEG data collection as well as the tasks to be performed. Participants were made aware of all task details, including the memory test, at the start of the experiment. After completing the instructions, EEG data collection began the first of 2 blocks, and participants performed 2 separate tasks in each block, see Figure 3.

Each paired associate encoding trial started with a fixation cross which lasted a random duration between 0.5 and 1.5 seconds, uniformly sampled. Then the current paired associate appeared on the screen for 6 seconds with the two words vertical stacked and the fixation cross between them. Vertical location for the words was randomly determined. Participants were instructed to visualize the two elements interacting with each other, and that their memory for the pairing of the elements would be later tested. Encoding was broken down into 3 mini-blocks with a single paired associate from each of the 36 events shown once. The ordering of paired associates across these presentations is described in the Design section above. There were no breaks between mini-blocks, however participants were given blink breaks every 35 s, at the end of any on-going trial. These blink breaks lasted 2 s (with the option of extending an additional 2 s with a button press). After all 108 paired associates were studied participants were allowed to take a short break if desired, and then continued on to the current block's cued recognition task.

The start of each cued recognition trial showed a fixation cross with 0.5 to 1.5 s uniform random duration, followed by the presentation of the cue word above an array of 6 target words (stacked 3 horizontally by 2 vertically, numerically label from left to right and top to bottom), and the fixation cross between them. The cue word and one of the target words were randomly drawn from a particular event, and the other 5 target words were drawn from the same element type as the correct target word, but randomly from other events. Locations of the target words in the array were determined randomly. Participants were instructed to press the numeric key corresponding to the target word that was previously studied with the cue word. After a button press was a made, a rectangle appeared on-screen for 0.5 s, highlighting the selected item. Participants were not made aware of the underlying event structure, or the open and closed-loop conditions.

3.5 Electrophysiological recordings

A 128-channel HydroCel Geodesic Sensor Net (GSN 200, v. 2.1; Tucker, 1993) was used to measure the EEG at the scalp using a central vertex reference (Cz) with a sampling rate of 250 Hz and a low-pass hardware filter at 100 Hz. The net was connected to a DC-coupled, high-input impedance amplifier (200 MΩ, Net AmpsTM; Electrical Geodesics, Inc., Eugene, OR) and recordings were made using the Net Station application. The electrodes were adjusted until impedance measurements were less than 40 kΩ.

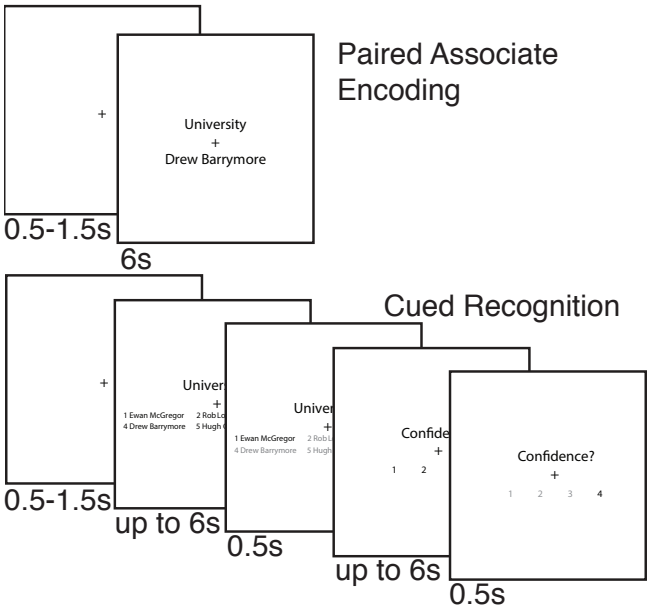

FIGURE 3: Example stimuli and timing for Dependent Events paradigm adapted for EEG. All trials had a uniform random inter-trial-interval (ITI) between 0.5s and 1.5s. Timing was response driving for both cued recognition and confidence judgements. After 6s a red warning message appeared directing the participant to make their selection. Feedback was giving by turning grey the choice that was not selected during cued recognition and confidence judgements.

All EEG data was taken from retrieval trials where all processing steps and analyses were done in MATLAB using in-house scripts (mat-mvm, n.d.) and the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011). A high-pass filter at 0.1 Hz, low-pass filter at 100 Hz, and a notch filter from 59–61 Hz were applied to the data. ICA artifact removal targeting blinks, and movement artifacts was done on the continuous (i.e. pre-epoched) EEG data after removal of large artifacts through coarse visual inspection. A semi-automated approach to artifact estimation and detection was done using the SASICA plugin ported for FieldTrip data structures, and methods described in Chaumon, Bishop, and Busch (2015). After ICA artifact removal, each cued recognition trial was epoched in two ways: stimulus locked (SL) and response locked (RL). SL epochs were 5000 ms segments, 1 s before the onset of the cue and target array, and 4 s after. RL epochs were also 5000 ms segements, 4 s before target response was made and 1 s after. Fieldtrips's automatic artifact detection was used to reject particularly noisy epochs as determined by a z-score threshold of 20, as well as those that exceed an amplitude of ± 200 μV or a maximum change in amplitude of 200 μV.

On average across subjects and conditions (i.e. both stimulus and response locked open and closed-loop trials) 37 ± 4(standard deviation) percent of trials were removed due to potential artifacts; there were no significant differences between conditions within epoch type in the number of removed trials. After artifact rejection the data were referenced to the average of all channels, after which wavelet based time frequency decompositions were done.

3.5.1 Spectral Analysis

The spectral decomposition was performed using a set of 71 Morlet wavelets that were equally spaced in 0.67 Hz intervals from 3 to 50 Hz. Each wavelet had a width that was 4 times the period of its center frequency. The power, i.e., the squared magnitude of each complex coefficient, was then computed for every 40 ms time bin within the extracted time window. In total, the spectral decomposition transformed each of the time bins of a trial into the power values of 71 frequency bands for each of the 128 electrodes. For each stimulus locked (SL) trial, power estimates after the response time were nulled out and were not figured into condition averages. Similarly, for each response locked (RL) trial

power estimates before the onset of the stimulus were also nulled out. Each trial, both SL and RL, were then baseline corrected by z-scoring the full trial length based on a a given trial's pre-stimulus mean and standard deviation from time window between -300 ms to -100 ms relative to cue-target array onset.

EEG results were analyzed according to the difference between open and closed-loop conditions. To achieve this, *time* × *electrode* clusters were determined across subjects for each a-priori defined frequency band using a cluster based permutation test (Maris & Oostenveld, 2007). Frequency data were first averaged over a the defined frequency bands: 3-8 Hz for theta, 8-12 Hz for alpha and 12-30 Hz beta. These data were then analyzed across all electrodes except the 4 surrounding the eyes, and across the time window of 500 to 3000 ms relative to the cue-target array onset in the stimulus locked epochs, and -3000 ms to -500 ms relative to the button press for a given trial in response locked epochs. Clustering was done by performing a t-test comparing open and closed-loop conditions within each time/electrode bin across subjects for each of the target frequency ranges (theta, alpha and beta). These t-values were then thresholded at $p < .05$, and clustered together based on spatial and temporal adjacency. Cluster significance was calculated using a monte-carlo style permutation test of the summed t-values within a given cluster. Each observed cluster was subject to 500 random permutations of condition labels where its significance was estimated by the proportion of random permutations which yielded clusters that had a summed t-value as large or larger than the observed cluster.

3.5.2 Representation Similarity Analysis

The Representation Similarity Analysis (RSA), was motivated by the non-target reactivation seen in Horner et al. (2015) and was based on the methods originally developed by Kriegeskorte, Mur, and Bandettini (2008). The underlying structure used in the RSA was based on individual retrieval trials within a given event and was done in two ways. The first was the standard correlation approach where a similarity matrix was constructed by correlating the full $time \times frequency \times electrode$ matrix, splayed out to a single vector, pairwise with each retrieval trial within a given event. Ideally this results in a 6x6 matrix with all 6 retrieval trials per event correlated with each other, however some retrieval trials were removed due to the presence of artifacts and therefore all events with more then 3 intact retrieval trials were used in the RSA. The off diagonal R-values in this matrix were then converted to z-values using the inverse hyperbolic tangent, and averaged to achieve a single 'similarity' measure across retrieval trials for a given event. These were then averaged across condition with the prediction that more non-target reactivation would manifest in a higher similarity value averaged across events within a given condition.

In the second approach we adapted the RSA method to our data by using the pairwise variance between trials within a given event to determine the 'dissimilarity' of a given set of retrieval trials, i.e. the larger the variance across retrieval trials from a given event the more dissimilar the set of retrieval trails are. Specifically, the variance across all 6 retrieval trials within a given event was calculated for each significant $time \times frequency \times electrode$ bin. For example, in subject 1 the power estimate at time bin 0.4s, frequency bin 5 Hz, and electrode 128, is extracted for the 6 retrieval attempts in event 1. The variance is then calculated over these 6 values, and put into a 3 dimensional array with the same dimensions as the original data. Repeating this process for all $time \times frequency \times electrode$ bins yields a *dissimilarity* measure for each bin in the full 3 dimensional array for a specific event. This dissimilarity array was then averaged across *a-prior* defined frequency bands (i.e. 3-8 Hz for theta, 8-12 Hz for alpha and 12-30 Hz beta), then averaged within conditions (closed-loop and open-loop) ultimately yielding a *time* × *electrode* matrix of dissimilarity per condition for each subject. This matrix was then analyzed using the same non-parametric cluster statistics described above to determine significant differences between conditions clustered in time and space for each frequency band.

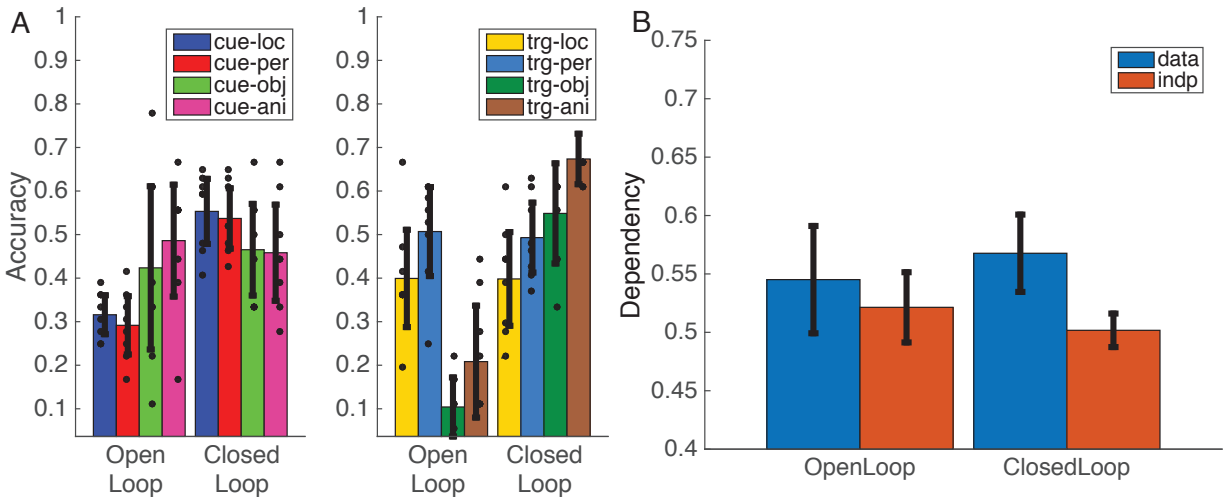

FIGURE 4: *(A)* Name Error assessed cued retrieval accuracy for all cue-targ and targ-cue combinations, averaged across network runs, within conditions. Black dots show individual network runs with unique random weight initialization balanced across encoding orders. Error bars show 95% confidence intervals. *(B)* Dependency assuming underlying event structure (labeled 'data') and assuming all cue-targ trial types are independent (labeled 'indp'), broken down by condition. Horizontal axis shows individual network runs. Dash lines show mean across network runs, and error bars show 95% confidence intervals.

4 Results

4.1 Computational Model Results

First results are presented for the new proposed model with the augmented thalamic attentional mechanisms, cursorily referred to as the attention augmented model. Then a comparison of these results with the original Theta-Phase model which has no attentional modulation is made to illustrate the difference in performance with the additional attentional modulation.

Overall performance in the attention augmented model was good for both closed and open-loop events. Accuracy of cued recall in the network was well above chance (1 of out 100 possible unit group patterns used to determine Name Error, or more conservatively 1 out 36 possible patterns used in the training) in all cue-target types for both conditions, shown in Figure 4A. This assures that the model was learning the paired associates patterns sufficiently in both conditions, and performing as expected. Closed-loop events did show a higher average accuracy compared to open-loop events across network runs ($\mu = 0.14 \pm 0.06$ 95% confidence), with the largest difference being when the object or animal elements were the retrieval target, similar to results from Horner et al. (2015).

The first critical comparison with behavior was the Dependency difference between open and closed-loop events, with the anticipation of closed-loop events having greater within event Dependency as compared to open-loop events. The second was the non-target reactivation difference with the anticipation that closed-loop events would shown greater non-target reactivation as compared to open-loop events.

4.2 Condition Differences in Dependency

As shown in Figure 4B, Dependency was significantly greater than the Independent model for closed-loop events (paired difference within model runs: $\mu = 0.07 \pm 0.04$ 95% confidence, $t(7) = 3.76, p < 0.01$), and not different for open-loop events ($\mu = 0.02 \pm 0.03, t(7) = 1.82, p = 0.1$). The interaction between condition differences and Independent model was also significant ($t(7) = 2.84, p < 0.05$), with the difference between closed-loop Dependency and its Independent model being greater than the difference

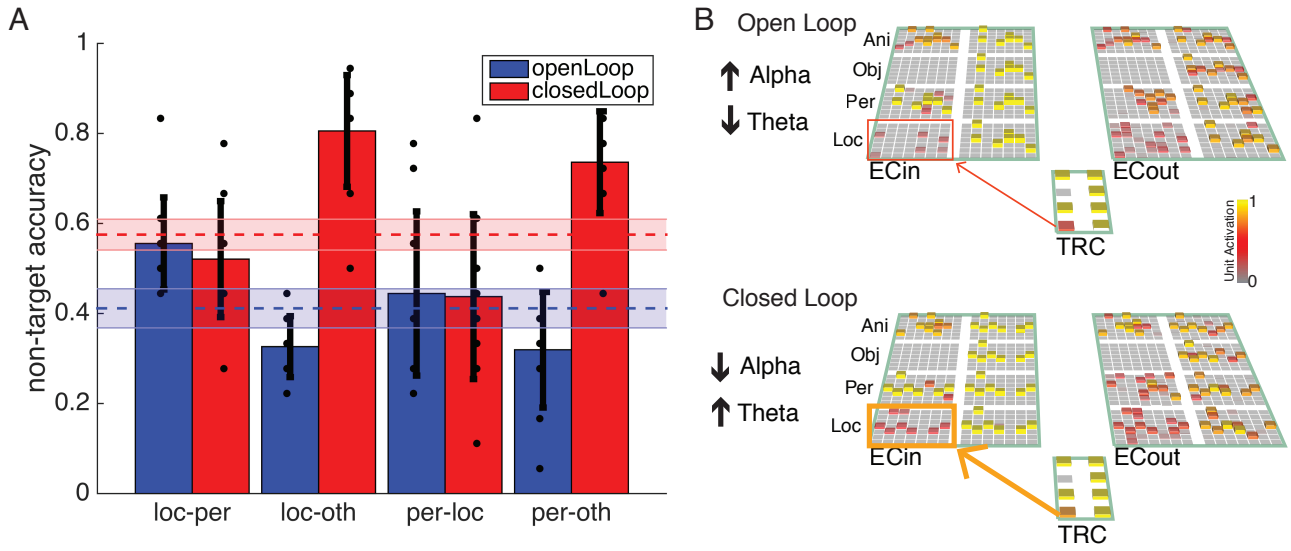

FIGURE 5: *(A)* Name Error accuracy for non-target reactivation in open-loop and closed-loop events. The horizontal axis labels correspond to the cue and target respectively, and the non-target element is the one not listed in the label. For example 'loc-per' indicates a location cue and person target trial, and therefore the non-target element is either object or animal depending on the encoding pairs for the given event. The 'oth' label corresponds to either the object or animal element again depending on the corresponding encoding pairs for the given event. Note, non-target accuracy was not assessed for object or animal elements as cues, as open-loop events have no corresponding non-target element in these cases, and would be a biased comparison with closed-loop events which do. Black dots indicate individual network runs with different training orders (see section 2.3), red bars are closed-loop and blue bars are open-loop, dash lines are means across all cue-targ trial types, and error bars are 95% confidence intervals. *(B)* Example cued retrieval trials from open-loop (top) and closed-loop (bottom) events. Both show a retrieval trial with a person cue (unit group labeled Per), and a animal target (Ani), and a location non-target (Loc). TRC shows bottom-up (person and animal unit groups) as well as hippocampal driven (Location unit group) modulation of thalamic attention. Predicted pattern of oscillatory power for corresponding open and closed loop trials, shown where the higher the TRC activity the lower the expected Alpha power; conversely the lower the higher the hippocampally induced EC activity the higher the Theta power.

between open-loop Dependency and its Independent model.

4.3 Condition Differences in Non-Target Reactivation

As shown in Figure 5A, successful non-target reactivation, as assessed through Name Error, was significantly greater in the closed-loop condition compared to the open-loop. Interestingly this difference was most pronounced in the retrieval trials in which the 'other' (or object/animal) stimulus dimension was the target of the cued retrieval. This implies that in open-loop trials, the network was less likely to successfully reactivate the location or person stimulus elements when attempting to retrieve the animal/object elements. This can be seen in Figure 5B where cued retrieval is shown for an example open-loop event on top and a closed-loop event on bottom. Both examples are showing a retrieval trial where the cue element is person (unit group 2, numbering from lower left proceeding up), and the target element is animal (unit group 4), and the non-target element is location (unit group 1). The relatively lower TRC activation in the non-target element for the open-loop retrieval compared to the closed-loop quantitatively shows the reduced non-target reactivation. Similarly, the qualitative activity in the closed-loop EC in and out non-target unit group is more complete and closer to the correct pattern as compared to the open-loop activity.

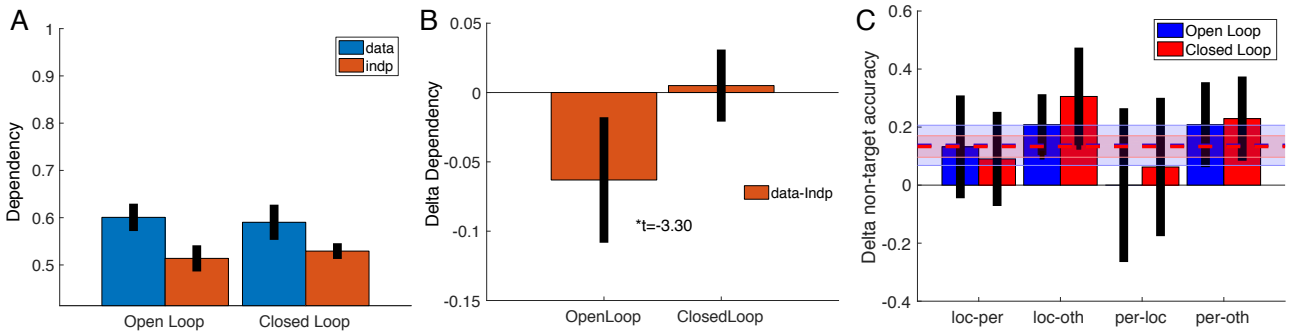

FIGURE 6: *(A)* Dependency results from original Theta-Phase model. Here an increase in the difference between data and independent measures of dependency within open-loop trials can be seen compared to augmented model results shown in Figure 5A. *(B)* Delta-Dependency: Difference in data and independent measures of dependency when comparing the augmented model with thalamic attention to the original Theta-Phase model. Here the original model shows a greater difference in dependency in the open-loop stimuli as compared to the augmented model. *(C)* Comparison of Non-target accuracy between the attention augmented model and the original Theta-Phase model. The attention augmented model shows an increase Theta-Phase model shows lower accuracy in open-loop animal and object (oth) target stimuli manifest in higher differences in dependency as compared to the thalamic attention model (Figure 5A). This increase in non-target accuracy illustrates to attention model's ability to represent the association structure of open and closed-loop events in a similar fashion to human subjects. All error bars are 95% confidence intervals.

4.4 Comparison with Original Theta-Phase Model

Results comparing the augmented model with the additional attentional mechanisms to the original Theta Phase model are shown here to illustrate the most direct and meaningful differences in performance. As shown in Figure 6, the original Theta-Phase model showed a greater difference between data and independent measures of Dependency within the open-loop stimuli. This is most directly illustrated in Figure 6B, where the difference in data vs. independent measures of Dependency are contrasted as the augmented model minus the Theta-Phase. Here a significant decrease can be seen ($t(14) = 3.30$, two sample t-test), which is driven by original Theta-Phase model learning both open and closed-loop stimuli in a similar as illustrated by similar levels of Dependency between the condition types.

Interestingly, both models show no difference in accuracy across cue and target pairings. However, non-target accuracy for animal and object targets (i.e. 'loc-oth' and 'per-oth' in Figure 6C) showed a significant increase when comparing the augmented model to the Theta-Phase for both open-loop (two sample t-test, location cue: $t(14) = 4.69, p < 0.001$, person cue: $t(14) = 3.36, p = 0.004$) and closed-loop (location cue: $t(14) = 4.29, p < 0.001$, person cue: $t(14) = 3.73, p = 0.002$). Although this non-target increase in accuracy does not directly impact measures of dependency as it can only be inferred through co-activation from particular cue-target trials, it does suggest the mechanism working to manifest the change in dependency measures shown between the original Theta-Phase and augmented models. Here the Theta-Phase model co-activates unit groups that are not intended as part of the association structure, decreasing non-target accuracy relative to the augmented model. This over active association in the Theta-Phase model leads to higher differences in data vs. independent measures of Dependency but misrepresents implicit association structure that human subjects are able to capture.

4.5 Behavioral Results

Behavioral results, as seen in Figure 7, show a significant difference in accuracy between closed-loop and open-loop conditions (closed-loop minus open-loop: $\mu = 0.05 \pm 0.03$ 95% confidence, $t(27) =$

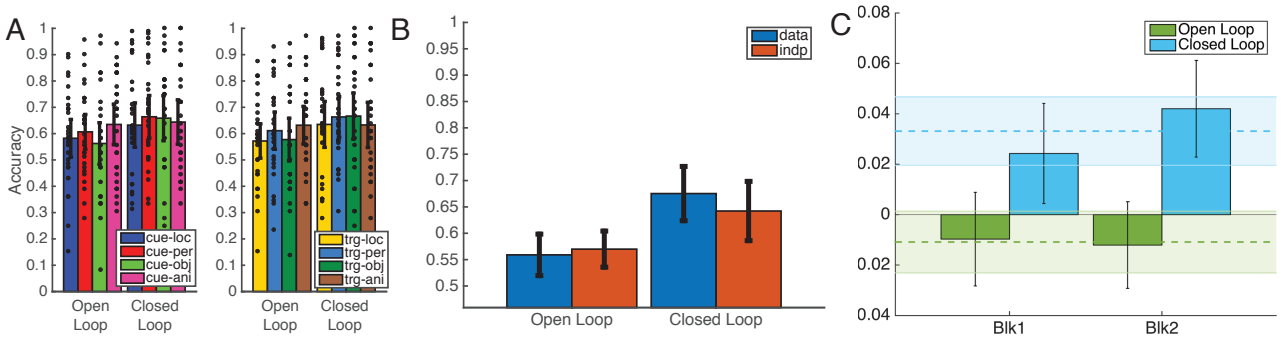

FIGURE 7: *(A)* Cued recognition accuracy for all cue-targ and targ-cue combinations, averaged across participants, within conditions. Black dots show individuals' performance. Error bars show 95% confidence intervals. *(B)* Average Dependency over subjects for both open-loop and closed-loop conditions. Blue bars show Dependency assuming underlying event structure referred to as 'data', and red bars are assuming all cue-targ trial types are independent referred to as 'indp'. Error bars show 95% confidence intervals. *(C)* Dependency difference as a function of block. Dash lines show mean across blocks, and error bars show 95% confidence intervals.

$3.92, p < 0.001$), similar to Horner et al. (2015) and the neural network simulation results. Retrieval confidence was also significantly higher in the closed-loop trials as compared with open-loop (closed-loop confidence minus open-loop confidence: $\mu = 0.21 \pm 0.1 t(27) = 4.34, p < 0.001$). Similarly, average retrieval confidence was highly correlated with accuracy across subjects ($r = 0.83, p < 0.001$).

Critically, there was significant difference between Dependency and the Independent model in the closed-loop condition ($\mu = 0.03 \pm 0.01, t(27) = 4.91, p < 0.001$), and not in the open-loop condition ($\mu = -0.01 \pm 0.01, t(27) = -1.78, p = 0.09$), and there was also a significant interaction in Dependency by condition, with the difference between closed-loop Dependency and its Independent model being greater then the difference between the open-loop Dependency and its Independent model ($\mu = 0.04 \pm 0.02, t(27) = 4.93, p < 0.001$). There were significant differences for accuracy between blocks for closed-loop events (block 1 minus block 2: $\mu = -0.10 \pm 0.07, t(27) = -2.74, p < 0.05$) and Dependency ($\mu = -0.07 \pm 0.06, t(27) = -2.68, p < 0.05$). Interestingly, these block effects are in the opposite direction as would be predicted from proactive interference, where memory performance progressively gets worse over the blocks. Similarly, the difference between closed-loop Dependency and its Independent model was significant when considered within each block separately (block 1: $\mu = 0.03 \pm 0.02, t(27) = 2.51, p < 0.05$; block 2: $\mu = 0.04 \pm 0.02, t(27) = 4.49, p < 0.001$), however there was no significant interaction in this difference across blocks ($\mu = -0.02 \pm 0.03, t(27) = -1.32, p = 0.2$).

4.6 EEG Results

4.7 Closed vs. Open Loop EEG Clusters

Condition differences between open and closed-loop EEG time frequency was assessed using the cluster based approach described in Maris and Oostenveld (2007), with a-priori defined frequency bands in the theta (3 to 8 Hz), alpha (8 to 12 Hz), and beta (12 to 30Hz) bands, and temporal windows between 500 and 3000 ms relative to cue-target onset (referred to as Stimulus Locked), and -3000 to -500 ms relative to button press (referred to as Response Locked). Critically, open and closed-loop trials for this analysis were sub-selected to only include trials where the target item was correctly identified, this yielded Stimulus Locked condition trial counts: closed-loop $\mu = 81 \pm 19$ 95% confidence interval over subjects, open-loop $\mu = 72 \pm 15$; and Response Locked condition trial counts: closed-loop $\mu = 88 \pm 18$, and open-loop $\mu = 79 \pm 15$. This subselection highlights the fact that any differences are not driven by accuracy between the conditions (however very similar effects were also found using all trials).

Two clusters were found meeting the $p < 0.05$ permutation significance level, shown in Figure 8. The

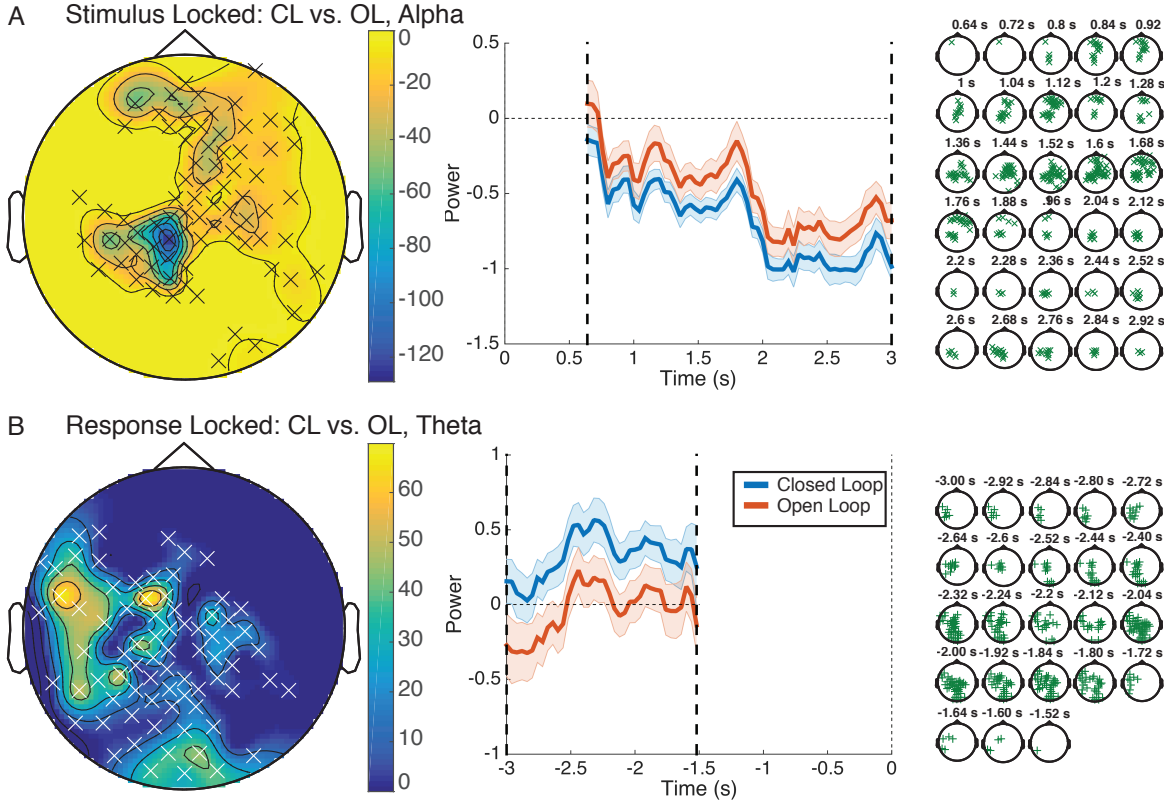

FIGURE 8: *(A)* Cluster of alpha power showing significant difference between closed vs. closed-loop events, i.e here open-loop events showed more power as compared to closed-loop events for the highlighted *time* × *electrode* bins within the alpha band. Cluster significant between 640 ms to 3000 ms relative to onset of cue-target array. Far right plot show spatial location of significant electrodes over time. *(B)* Cluster of theta power showing significant difference between closed vs. closed-loop events. Here closed-loop events showed more power compared to open-loop events in the highlighted *time* × *electrode* bins within the alpha band. These were the only two clusters to meet the $p < 0.05$ criteria within theta, alpha, and beta frequency bands for both stimulus locked and response locked epochs. Topography on the left shows summed t-values over time with significance determined by permutation test. Middle plots show the average power within the significant electrodes with closed-loop shown in blue and open-loop shown in red. The significant time window is bounded by the large dashed lines. Far right plot shows spatial location of cluster electrodes over time.

first was found in the alpha band from 640 to 3000 ms Stimulus Locked, and showed a greater decrease in baseline corrected closed-loop power as compared to open-loop. The majority of the t-value mass was in the central posterior electrodes and spanned towards right-central anterior. The second significant cluster was found in the theta band from -3000 to -1520 ms Response Locked, and showed a greater increase in baseline corrected closed-loop power as compared to open-loop events. Its bulk of t-value mass was in left anterior electrodes, and spanned toward central posterior electrodes.

4.8 Representation Similarity Analysis

In an attempt to test for non-target reactivation a Representation Similarity Analysis was done across retrieval attempts within a given event. The interpretation of this analysis is that events with more non-target reactivation across the various cued recognition retrieval attempts would show more similarity in their EEG time frequency topography. The first approach, see Section 3.5.1 for a more complete description, tried to maintain as much alignment with standard RSA as possible, by doing correlations using the full frequency spectrum (3 to 50 Hz), and the full time window (500 ms to 3000 ms for Stimulus Locked epochs, and -3000 ms to -500 ms Response Locked). Averaging pairwise correlations across all retrieval attempts within a given event yielded a single measure of similarity for each event, which were then averaged into open and closed-loop conditions. No significant differences were found between open and closed-loop events for both stimulus locked and response locked epochs. It is likely the full $time \times frequency \times electrode$ space is non-optimal for detecting these similarity differences, and in an attempt to narrow down this space a second approach was adopted to take advantage of the clustering methods used in the power analysis.

This second approach measured the variance across the 6 retrieval attempts within an event for each $time \times frequency \times electrode$ bin to get a measure of dissimilarity within a given event. Then averaged over theta, alpha and beta frequency bands, and averaged over events within open-loop and closed-loop condition yields a single $time \times electrode$ matrix of dissimilarity for each condition. Using the clustering method described in Section 3.5.1, a permutation test was done looking for the $time \times electrode$ clusters for each frequency band that showed a significant difference between the open and closed-loop conditions. This yielded two significant clusters from the Stimulus Locked data at the $p < 0.05$ level as determined through a permutation test, see Figure 9. The first was in the theta band, showing more dissimilairty in the open-loop condition compared to the closed-loop (permutation significance $p < 0.05$). This cluster spanned from 2320 ms to 3000 ms, and the main bulk of the t-value mass was in the central posterior electrodes and spanned up to central anterior electrodes. The second was a alpha cluster, also showing more dissimilarity in the open-loop condition compared to the closed-loop. This cluster spanned from 2560 ms to 3000 ms, and the bulk of the t-value mass started in right anterior electrodes and spread to virtually all electrodes.

5 Discussion

Building from the Theta-Phase model (N. Ketz et al., 2013), a cortical-thalamic mechanism supporting perceptual inhibition in service of long-term memory encoding was purposed. This encoding mechanism in fact works through the successful retrieval of previously encoded information by providing an attentional filter on the retrieval processes; enhancing the patterns associated with the current stimulus that are most strongly retrieved and inhibiting those less strongly retrieved. This enhancement is based on a more general cortical-thalamic attention mechanism that can be used to provide top-down attentional modulation for any cortical region. Here the top-down signal is coming from hippocampally retrieved associated information relative to a provided cue in a cued recognition task.

This comprehensive model of the hippocampus and surrounding cortex was then used to explore a multi-element memory paradigm shown to differentially engage the hippocampus, referred to as the Dependent Events paradigm. Here, the mechanisms relating hippocampal reactivation to theta oscillations and attentional filtering to alpha oscillations provided explanatory power in understanding why

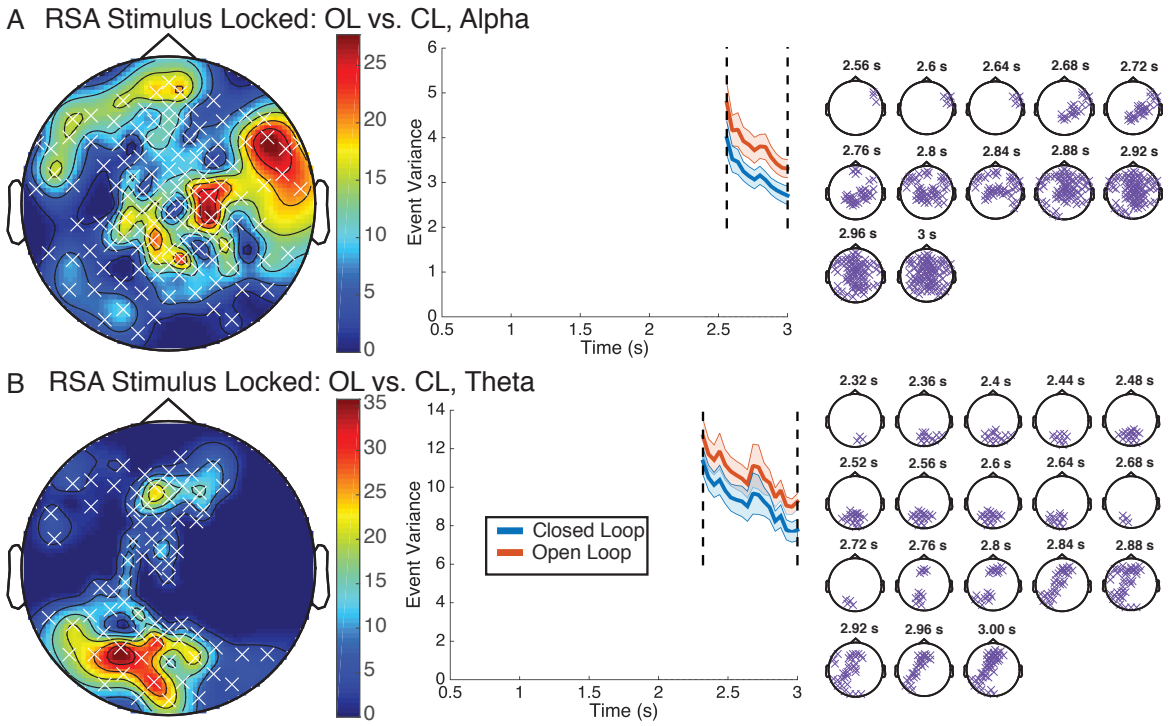

FIGURE 9: *(A)* Cluster of representation variance difference between open-loop and closed-loop events in the alpha band. Here open-loop events showed more within event variance as compared to closed-loop events for the highlighted *time* × *electrode* bins within the alpha (8 to 12 Hz) band. Cluster significant from 2560 ms to 3000 ms relative to onset of cue-target array, and in the highlighted electrodes. Right plot shows spatial distribution of significant electrodes over time. *(B)* Similar cluster for representation variance difference found in theta band. Again, open-loop events showed more within event variance as compared to closed-loop events for the highlighted *time* × *electrode* bins, this time in the theta (3 to 8 Hz) band. Cluster significant from 2320 to 3000 ms relative to onset of cue-target array. These were the only two clusters to meet the $p < 0.05$ criteria within theta, alpha, and beta frequency bands for both stimulus locked and response locked representation similarity analyses. Topography on the left shows summed t-values over time with significance determined by permutation test. Middle plot shows the average power within the significant electrodes with closed-loop shown in blue and open-loop shown in red. The significant time window is bounded by the large dashed lines. Far right plot shows spatial location of cluster electrodes over time.

different association structures lead to more pattern completion, i.e. closed-loop events, as compared to others, i.e. open-loop events. The results showed that repeated reactivation of information incidental to the current task, i.e. in the closed-loop events, strengthen the underlying association structure of those memories, while memory traces which didn't experience this reactivation, i.e. open-loop events, showed more attentional filtering of cortical processing. This filtering blocked reactivation of weaker associations, and instead strengthened only the cue-target associations. The implications of this attentional selection process revealed a potential mechanism to explain the behavioral differences in the Dependent Events paradigm, and more generally provides a model of how retrieval processes can both strengthen are degrade associations depending on task goals as well as memory strength.

These simulations were followed up with empirical studies testing the behavioral and EEG time frequency predictions derived from the simulations. Those results support the predictions of the computational model. Specifically, it was found that theta oscillations increased for successful retrieval attempts which showed more reactivation of incidental, or 'non-target', information as compared to retrieval attempts which did not. Similarly, alpha power showed a greater decrease when retrieval reactivated more non-target information. This reactivation process manifests as a behavioral change in 'Dependency' between pairs of stimuli that shared an underlying association structure, referred to as 'events'. Here Dependency is measured as an increased probability of correctly retrieving all (or none) of the associated items, and the computational model predicted that more reactivation leads to greater Dependency. This reactivation was also independently tested through a Representation Similarity Analysis (RSA). Here, more similarity between retrieval attempts within a given event was shown to track with Dependency, and this increased similarity manifested itself in the theta and alpha frequency bands.

5.1 Model of Thamlo-Cortical-Hippocampal Interactions

What are the mechanisms in the model which support the differences witnessed between open and closed-loop association structures? The main cause of these condition differences likely stem from the model's ability to reactivate the non-target elements more so in the closed-loop compared to the open-loop condition. This is best illustrated in the comparison between the original Theta-Phase model with the attention augmented model shown in Figure 6.

In the attention augmented model the increased reactivation initially comes from the fact that each of the different event elements are studied together in the closed-loop events, allowing the network to be more likely to later reactivate those associations. In contrast, because the open-loop events don't have the fully overlapping association structure that the closed loop events do, a given cue in the open-loop retrieval trial is only likely to reactivate its previously studied association which would be the target element of a given retrieval trial, and unable to reactivate the non-target element. This is in contrast to the original Theta-Phase model which more freely associates in open-loop events because there is no thalamic modulation to down-regulate its natural tendencies to create associations between unit-groups. This leads to an increase in open-loop differences in data vs independent measures of dependency, but it comes at the cost of decreased non-target accuracy ultimately reflecting a misrepresentation of the 'true' underlying association structure of the open-loop events.

By incorporating this thalamically mediated attentional modulation mechanism into the Theta-Phase model a new set of data can be captured by the simulations, and the implicit biases of human memory can be more accurately reflected. Further, a new handle on EEG time frequency signatures and the underlying thalamic mechanisms is in place for further testing. Due to the motivating biology supporting this computational model, several hypothesis can now be generated regarding the relationship of neural oscillations and the Dependent events paradigm, or pattern completion and bottom-up attentional modulation more generally. First, due to the greater Dependency and greater non-target reactivation witnessed in closed-loop events we can expect greater cortical theta power in the EEG time frequency signature. This is due to the hippocampus driving cortex into pattern completion, as witnessed in the model at Entorhinal Cortex (EC) but assumed to also propagate out to distributed

cortical regions that ultimately feed into EC. The second major hypothesis is that cortical alpha power should inversely scale with the amount of attentional modulation used in the network, such that when the network's attentional modulation is high, EEG alpha power should be low.

In general this work provides another perspective on the same mechanisms of the Sync/deSync model (Parish et al., 2018). Here, we propose a specific thalamic mechanism responsible for the alpha dynamics witnessed in declarative memory processes and implement it in a rate coding neural network model. This model, however, does not directly manifest alpha dynamics that can be directly measured, it only indirectly models them through the thalamo-coritical-hippocampal interactions. Parish et al. (2018), use a spiking model to ellicit measures that are more directly related to the oscillatory dynamics that both their and our models are interested in. Similarly, varying levels of detailed thalamo-cortical models provide solid ground for exploring the attention mechanisms proposed here (Hindriks & van Putten, 2013; Becker, Knock, Ritter, & Jirsa, 2015). Future work can try and bridge this gap by including the spiking dynamics of the Sync/deSync model with the systems level mechanisms of our model.

5.2 Behavioral and EEG Support

The computational simulations explored in this work suggests that the hippocampus is driving cortex during retrieval more so in the closed-loop events compare to the open-loop, and the EEG time frequency clusters and Representation Similarity Analysis (RSA) witnessed in this study suggest this reactivation manifests in the theta and alpha bands. Specifically, the closed-loop trials showed a greater increase in theta power from a fixation baseline as compared to the open-loop trials. Conversely, in the alpha band there was a greater decrease in baseline corrected power for the closed-loop compared to the open-loop trials.

The RSA results also support this relationship to alpha and theta power. The clusters shown in Figure 9 illustrate that the open-loop condition shows more variance in EEG power over retrieval attempts within a given event as compared with the closed-loop condition, and that this variance is specific to the theta and alpha bands in the scalp topographies shown. This reinforces the interpretation that closed-loop retrieval attempts are more likely to reactivate all elements within a given event, where the hippocampus drives cortex in theta band and turns off thalamo-cortical suppression in the alpha band.

Its interesting to note that the alpha cluster was found in the Stimulus Locked epochs and the theta cluster was found in Response Locked. This fits with the interpretation that alpha is more directly related to perceptual information while theta is related to the reactivation of previous experiences. Here the difference in alpha power is locked to the input stimulus and sustains over the extent of the trial, suggesting that even before a strong retrieval has occurred some differences between closed and closed-loop events are present. Conversely, theta differences are locked to the response, suggesting that once a sufficient amount of information has been retrieved a decision can be made. This decision process is presumable happening in the latter half (i.e. -1500 to 0 ms) of the Response Locked epochs.

These results provide new insight into the Dependent Events paradigm, and helps illustrate the temporal dynamics that was not previously possible in the fMRI studies. Specifically, the response locked theta effects and stimulus locked alpha effects both support the mechanism proposed by the computational model, and provide extra insight into the dynamics support the Dependent Events paradigm.

6 Conclusions

In this work we have proposed a model of thalamo-coritical-hippocampal interactions that makes specific hypothesis about the role and relative magnitudes of neural oscillations during declarative memory processes. Within the Dependent Events paradigm, simulations show that this model captures broad EEG time frequency dynamics related to alpha and theta oscillations, and makes mechanistic

predictions about how these oscillations relate to the underlying neural processes and their implications for cognitive level behavior. An EEG study using the same paradigm replicated previous studies' behavioral results and showed similar patterns of oscillatory dynamics as predicted by the neural network model.

References

Aggleton, J. P., Dumont, J. R., & Warburton, E. C. (2011, June). Unraveling the contributions of the diencephalon to recognition memory: a review. *Learning & memory*, *18*, 384–400.

Aisa, B., Mingus, B., & O'Reilly, R. (2008, October). The emergent neural modeling system. *Neural Networks*, *21*(8), 1146–1152.

Barbas, H., Henion, T. H., & Dermon, C. R. (1991, November). Diverse thalamic projections to the prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, *313*, 65–94.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014, June). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [stat]*. arXiv: 1406.5823 [stat]

Becker, R., Knock, S., Ritter, P., & Jirsa, V. (2015, September). Relating Alpha Power and Phase to Population Firing and Hemodynamic Activity Using a Thalamo-cortical Neural Mass Model. *PLOS Computational Biology*, *11*(9), e1004352. doi:10.1371/journal.pcbi.1004352

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Buzsaki, G. (2002). Theta oscillations in the hippocampus. *Neuron*, *33*(3), 325–340.

Chaumon, M., Bishop, D. V. M., & Busch, N. A. (2015, July). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*, *250*, 47–63. doi:10.1016/j.jneumeth.2015.02.025

Fuentemilla, L., Barnes, G. R., Düzel, E., & Levine, B. (2014, August). Theta oscillations orchestrate medial temporal lobe and neocortex in remembering autobiographical memories. *NeuroImage*, *85*(2), 730–737.

Gabrieli, J. D. (1998, January). Cognitive neuroscience of human memory. *Annual Review of Psychology*, *49*, 87–115.

Griffiths, B. J., Michelmann, S., Roux, F., Chelvarajah, R., Rollings, D. T., Sawlani, V., ... Staresina, B. (2018). Hippocampal synchrony and neocortical desynchrony cooperate to encode and retrieve episodic memories. *bioRxiv*, 305698.

Hanslmayr, S., Staresina, B. P., & Bowman, H. (2016). Oscillations and episodic memory: addressing the synchronization/desynchronization conundrum. *Trends in neurosciences*, *39*(1), 16–25.

Hanslmayr, S., Staudigl, T., & Fellner, M.-C. (2012). Oscillatory power decreases and long-term memory: the information via desynchronization hypothesis. *Frontiers in human neuroscience*, *6*.

Hindriks, R. & van Putten, M. J. (2013, April). Thalamo-cortical mechanisms underlying changes in amplitude and frequency of human alpha oscillations. *NeuroImage*, *70*, 150–63.

Horner, A. J., Bisby, J. A., Bush, D., Lin, W.-J., & Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, *6*, 7462. doi:10.1038/ncomms8462

Horner, A. J. & Burgess, N. (2013, November). The associative structure of memory for multi-element events. *Journal of Experimental Psychology. General*, *142*(4), 1370–1383. doi:10.1037/a0033626

Horner, A. J. & Burgess, N. (2014, May). Pattern completion in multielement event engrams. *Current biology: CB*, *24*(9), 988–992. doi:10.1016/j.cub.2014.03.012

Jones, M. W. & Wilson, M. A. (2005, December). Theta rhythms coordinate hippocampal-prefrontal interactions in a spatial memory task. *PLoS biology*, *3*.

Jutras, M. J., Fries, P., & Buffalo, E. A. (2013, August). Oscillatory activity in the monkey hippocampus during visual exploration and memory formation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 13144–9.

Ketz, N. A., Jensen, O., & O'Reilly, R. C. (2015, January). Thalamic pathways underlying prefrontal cortex-medial temporal lobe oscillatory interactions. *Trends in neurosciences*, *38*, 3–12.

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013, June). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, *9*, e1003067.

Kievit, J. & Kuypers, H. G. (1977, September). Organization of the thalamo-cortical connexions to the frontal lobe in the rhesus monkey. *Experimental brain research*, *29*, 299–322.

Klimesch, W. (2012, December). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, *16*, 606–17.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4).

Lee, J. H., Whittington, M. A., & Kopell, N. J. (2013). Top-down beta rhythms support selective attention via interlaminar interaction: a model. *PLoS Computational Biology*, *9*.

Leresche, N., Lightowler, S., Soltesz, I., Jassik-Gerschenfeld, D., & Crunelli, V. (1991, September). Low-frequency oscillatory activities intrinsic to rat and cat thalamocortical cells. *The Journal of physiology*, *441*, 155–74.

Lopes da Silva, F. (1991, August). Neural mechanisms underlying brain waves: From neural membranes to networks. *Electroencephalography and Clinical Neurophysiology*, *79*(2), 81–93.

Maris, E. & Oostenveld, R. (2007, August). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–90.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995, August). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, *102*(3), 419–457.

Norman, K. A. & O'Reilly, R. C. (2003, November). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Nyhus, E. & Curran, T. (2010, June). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience and biobehavioral reviews*, *34*(7), 1023–1035.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014, December). Complementary Learning Systems. *Cognitive Science*, *38*(6), 1229–1248.

Parish, G., Hanslmayr, S., & Bowman, H. (2018). The Sync/deSync model: how a synchronized hippocampus and a desynchronized neocortex code memories. *Journal of Neuroscience*, *38*(14), 3428–3440.

Park, H., Lee, D. S., Kang, E., Kang, H., Hahm, J., Kim, J. S., ... Jensen, O. (2014, February). Blocking of irrelevant memories by posterior alpha activity boosts memory encoding. *Human brain mapping*.

Prince, S. E., Daselaar, S. M., & Cabeza, R. (2005, February). Neural correlates of relational memory: successful encoding and retrieval of semantic and perceptual associations. *The Journal of neuroscience*, *25*, 1203–1210.

Saalmann, Y. B. & Kastner, S. (2011, July). Cognitive and perceptual functions of the visual thalamus. *Neuron*, *71*(2), 209–223.

Sherman, S. & Guillery, R. (2006). *Exploring the Thalamus and Its Role in Cortical Function*. Cambridge, MA: MIT Press.

Staresina, B. P., Michelmann, S., Bonnefond, M., Jensen, O., Axmacher, N., & Fell, J. (2016). Hippocampal pattern completion is linked to gamma power increases and alpha power decreases during recollection. *Elife*, *5*, e17397.

Van der Werf, Y. D., Jolles, J., Witter, M. P., & Uylings, H. B. M. (2003, October). Contributions of thalamic nuclei to declarative memory functioning. *Cortex*, *39*, 1047–1062.

Vijayan, S. & Kopell, N. J. (2012, June). Thalamic model of awake alpha oscillations and implications for stimulus processing. *Proceedings of the National Academy of Sciences*, *109*(45), 18553–18558. doi:10.1073/pnas.1215385109

Witter, M. P. (2010). Hippocampal Microcircuits. In V. Cutsuridis, B. Graham, S. Cobb, & I. Vida (Eds.), *Connectivity of the Hippocampus* (Vol. 5, pp. 5–26). Springer Series in Computational Neuroscience. Medical-Technical Research Centre, Norwegian University of Science and Technology Kavli Institute for Systems Neuroscience and Centre for the Biology of Memory Trondheim Norway.