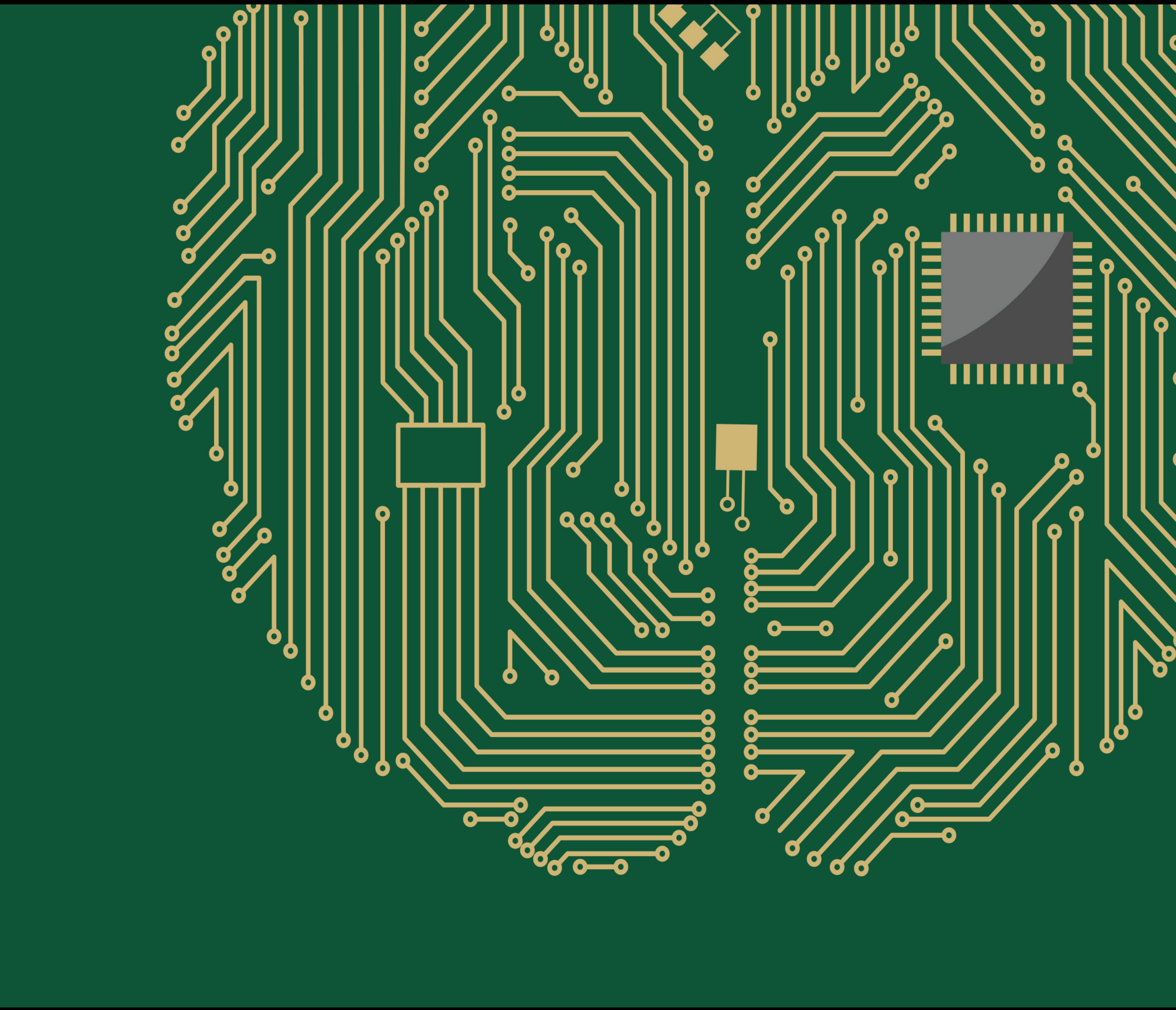
Computational Intelligence and Neuroscience

Neurocognitive Models of Sense Making

Lead Guest Editor: Giorgio Ascoli
Guest Editors: Rajan Bhattacharyya and Matthew M. Botvinick



Neurocognitive Models of Sense Making

Computational Intelligence and Neuroscience

Neurocognitive Models of Sense Making

Lead Guest Editor: Giorgio Ascoli
Guest Editors: Rajan Bhattacharyya and Matthew M. Botvinick

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 921695, 29 pages
http://dx.doi.org/10.1155/2013/921695

Research Article

A Functional Model of Sensemaking in a Neurocognitive Architecture

Christian Lebiere,[1] Peter Pirolli,[2] Robert Thomson,[1] Jaehyon Paik,[2] Matthew Rutledge-Taylor,[1] James Staszewski,[1] and John R. Anderson[1]

[1] *Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2] *Palo Alto Research Center, Palo Alto, CA 94304, USA*

Correspondence should be addressed to Christian Lebiere; cl@cmu.edu

Received 21 December 2012; Revised 26 April 2013; Accepted 8 July 2013

Academic Editor: Giorgio Ascoli

Sensemaking is the active process of constructing a meaningful representation (i.e., making sense) of some complex aspect of the world. In relation to intelligence analysis, sensemaking is the act of finding and interpreting relevant facts amongst the sea of incoming reports, images, and intelligence. We present a cognitive model of core information-foraging and hypothesis-updating sensemaking processes applied to complex spatial probability estimation and decision-making tasks. While the model was developed in a hybrid symbolic-statistical cognitive architecture, its correspondence to neural frameworks in terms of both structure and mechanisms provided a direct bridge between rational and neural levels of description. Compared against data from two participant groups, the model correctly predicted both the presence and degree of four biases: confirmation, anchoring and adjustment, representativeness, and probability matching. It also favorably predicted human performance in generating probability distributions across categories, assigning resources based on these distributions, and selecting relevant features given a prior probability distribution. This model provides a constrained theoretical framework describing cognitive biases as arising from three interacting factors: the structure of the task environment, the mechanisms and limitations of the cognitive architecture, and the use of strategies to adapt to the dual constraints of cognition and the environment.

1. Introduction

We present a computational cognitive model, developed in the ACT-R architecture [1, 2], of several core information-foraging and hypothesis-updating processes involved in a complex sensemaking task. Sensemaking [3–6] is a concept that has been used to define a class of activities and tasks in which there is an active seeking and processing of information to achieve understanding about some state of affairs in the world. Complex tasks in intelligence analysis and situation awareness have frequently been cited as examples of sensemaking [3–5]. Sensemaking, as in *to make sense*, implies an active process to construct a meaningful and functional representation of some aspects of the world. A variety of theories and perspectives on sensemaking have been developed in psychology [3, 4], human-computer interaction [6], information and library science [7], and in organizational science [8]. In this paper we present a cognitive model of basic sensemaking processes for an intelligence analysis task.

A major concern in the intelligence community is the impact of cognitive biases on the accuracy of analyses [9]. Two prominent biases are confirmation bias, in which an analyst disproportionately considers information that supports the current hypothesis, and anchoring bias, in which an initial judgment is insufficiently revised in the face of new evidence. In the task used in this paper, sensemaking is instantiated in terms of estimation of probability distributions over hypothesis space. Rational Bayesian optima are defined over those distributions, with cognitive biases defined as deviations from those optima. In this framework, confirmation bias can then be defined as a distribution "peakier" than the Bayesian optimum, whereas anchoring bias is a flatter-than-rational distribution reflecting an insufficient adjustment from the original uniform prior. We present simulation results that

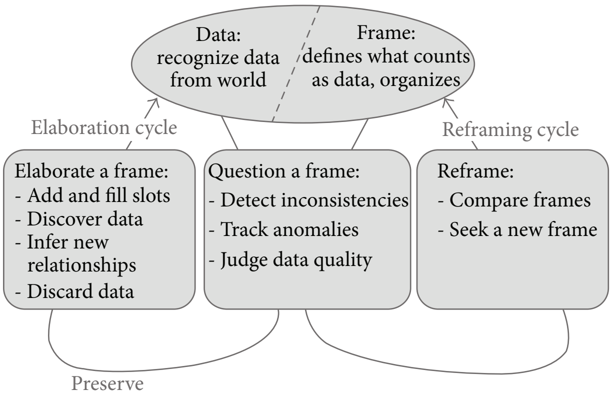


Figure 1: The Data-Frame model of sensemaking. Image reproduced by Klein et al. [4].

exhibit several cognitive biases, including confirmation bias, anchoring and adjustment, probability matching, and base-rate neglect. Those biases are not engineered in the model but rather result from the interaction of the structure and statistics of the task, the structure and mechanisms of our cognitive architecture, and the strategies that we select to perform the former using the latter.

Figure 1 presents the Data/Frame theory of sensemaking [3]. The Data/Frame theory assumes that meaningful mental representations called *frames* define what counts as *data* and how those data are structured for mental processing [4]. A similar conceptual model was employed in Pirolli and Card [5] to perform a cognitive task analysis of intelligence analysis [10–12]. Frames can be expressed in a variety of forms including stories, maps, organizational diagrams, or scripts. Whereas frames define and shape data, new data can evoke changes to frames. In this framework, sensemaking can involve elaboration of a frame (e.g., filling in details), questioning a frame (e.g., due to the detection of anomalies), or reframing (e.g., rejecting a frame and replacing it with another). The Data/Frame theory proposes that backward-looking processes are involved in forming mental models that explain past events, and forward-looking mental simulations are involved in predicting how future events will unfold. We describe how frames can be represented in a cognitive architecture and how the architectural mechanisms can implement general sensemaking processes. We then demonstrate how the dynamics of sensemaking processes in a cognitive architecture can give rise to cognitive biases in an emergent way.

The structure of this paper is as follows. Section 2 defines the AHA (Abducting Hotspots of Activity) experiment consisting of a suite of six sensemaking tasks of increasing complexity. Section 3 outlines our cognitive modeling approach to sensemaking: it describes the ACT-R architecture, how it is used to prototype neural models, which cognitive functions compose the model, and how four cognitive biases can be accounted for by the model. Section 4 presents the measures used to assess the cognitive biases in the AHA framework and then compares human and model results. Section 5 presents a test of the model's generalization on a data set that was unavailable at the time of initial model development. Finally, Section 6 summarizes our account of cognitive biases centered around the mechanisms and limitations of cognitive architectures, the heuristic that these mechanisms use to adapt to the structure of the task, and their interaction with the task environment.

2. The Task Environment

The AHA experiment consists of a series of six tasks developed as part of the IARPA (Intelligence Advanced Research Projects Activity), ICArUS (Integrated Cognitive-neuroscience Architectures for the Understanding of Sensemaking) program, whose goal is to drive the development of integrated neurocognitive models of heuristic and biases in decision-making in the context of intelligence analysis. The AHA tasks can be subdivided into two classes: the first focusing on learning the statistical patterns of events located on a map-like layout and generating probability distributions of category membership based on the spatial location and frequency of these events (Tasks 1–3) and the second requiring the application of probabilistic decision rules about different features displayed on similar map-like layouts in order to generate and revise probability distributions of category membership (Tasks 4–6).

The AHA tasks simulate the analysis of artificial geospatial data presented in a manner consistent with and informed by current intelligence doctrine (Geospatial Intelligence Basic Doctrine; http://www.fas.org/irp/agency/nga/doctrine.pdf). The tasks involve the presentation of multiple features consistent with intelligence data, which are presented in a GIS (Geographic Information System) display not unlike Google maps (https://maps.google.com). These features include

HUMINT: information collected by human sources such as detecting the location of events,

IMINT: information collected from imagery of buildings, roads, and terrain elements,

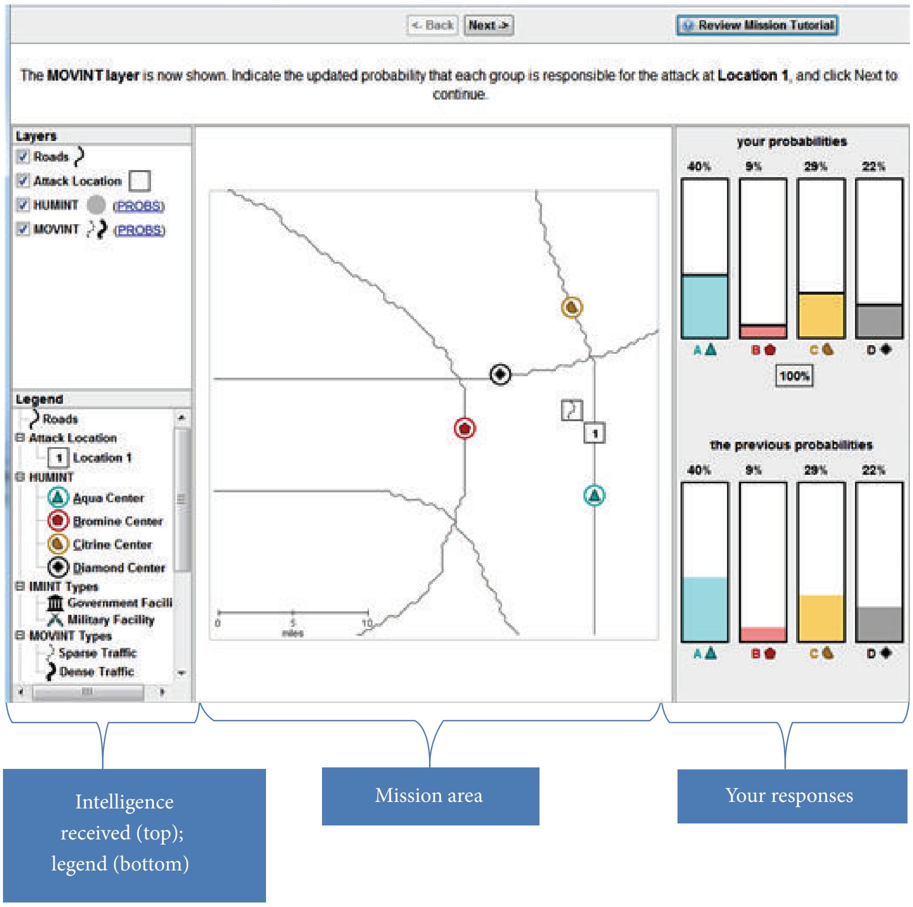
MOVINT: analysis of moving objects such as traffic density,

SIGINT: analysis of signals and communications,

SOCINT: analysis of social customs and attitudes of people, communities, and culture.

The display (see Figure 2) includes access to the mission tutorial and instructions (the top-right corner), a legend to understand the symbols on the map (the left pane), the map (the center pane), and participants' current and past responses (the right pane).

For Tasks 1–3, the flow of an average trial proceeds according to the following general outline. First, participants perceive a series of events (SIGACTs; SIGnals of ACTivity) labeled according to which category the event belonged. Categories were both color- and shape-coded, with the appropriate label {Aqua, Bromine, Citrine, or Diamond} listed in the legend. After perceiving the series of events, a probe event is displayed (represented as a "?" on the display). Participants were asked to generate a center of activity (e.g., prototype)



The **MOVINT layer** is now shown. Indicate the updated probability that each group is responsible for the attack at **Location 1**, and click Next to continue.

Figure 2: The image is a sample of the display in Task 4. To the left is a legend explaining all the symbols on the map (center). To the right are the probability distributions for the four event categories (both for the current and prior layer of information). The panel across the top provides step-by-step instructions for participants.

for each category's events, reflect on how strongly they believed the probe belonged to each category, and generate a probability estimate for each category (summed to 100% across all groups) using the sliders or by incrementing the counters presented on the right side of the Task interface. As an aid, the interface automatically normalized the total probability such that the total probability summed across each category equaled 100%. Participants were not provided feedback at this step. Scoring was determined by comparing participants distributions to an optimal Bayesian solution (see Section 4 for a detailed description of how the probability estimate scores are calculated). Using these scores it was possible to determine certain biases. For instance, participants' probability estimates that exhibited lower entropy than an optimal Bayes model would be considered to exhibit a confirmation bias, while probability estimates having higher entropy than an optimal Bayes model would be considered to exhibit an anchoring bias.

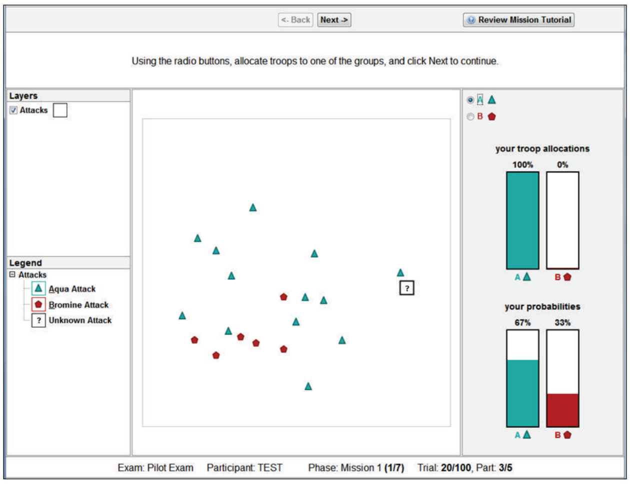
After finalizing their probability estimates, participants were then asked to allocate resources (using the same right-side interface as probability estimates) to each category with the goal of maximizing their resource allocation score, which was the amount of resources allocated to the correct category. Participants would receive feedback only on their resource allocation score. For Tasks 1–3, the resource allocation response was a forced-choice decision to allocate 100% of their resources to a single category. If that category produced the probe event, then the resource allocation score was 100 out of 100 for choosing the correct category, otherwise 0 out of 100 for choosing an incorrect category. Following this feedback, the next trial commenced.

For Tasks 4–6, the flow of an average trial was structurally different as intelligence "features," governed by probabilistic decision rules (see Table 1), were presented sequentially as separate layers of information on the display. These Tasks required reasoning based on rules concerning the relation of observed evidence to the likelihood of an unknown event belonging to each of four different categories. Participants updated their beliefs (i.e., likelihoods) after each layer of information (i.e., feature) was presented, based on the probabilistic decision rules described in Table 1.

For instance, in Task 4, after determining the center of activity for each category (similar in mechanism to Tasks 1–3) and reporting an initial probability estimate, the SOCINT (SOCial INTelligence) layer would be presented by displaying color-coded regions on the display representing each category's boundary. After reviewing the information presented by the SOCINT layer, participants were required

Table 1: Rules for inferring category likelihoods based on knowledge of category centroid location and an observed feature.

Features	Rules
HUMINT	If an unknown event occurs, then the likelihood of the event belonging to a given category decreases as the distance from the category centroid increases.
IMINT	If an unknown event occurs, then the event is four times more likely to occur on a *Government* versus *Military* building if it is from category A or B. If an unknown event occurs, then the event is four times more likely to occur on a *Military* versus *Government* building if it is from category C or D.
MOVINT	If an unknown event occurs, the event is four times more likely to occur in *dense* versus *sparse* traffic if it is from category A or C. If an unknown event occurs, the event is four times more likely to occur in *sparse* versus *dense* traffic if it is from category B or D.
SIGINT	If SIGINT on a category reports *chatter*, then the likelihood of an event by that category is seven times as likely as an event by each other category. If SIGINT on a category reports *silence*, then the likelihood of an event by that category is one-third as likely as an event by each other category.
SOCINT	If an unknown event occurs, then the likelihood of the event belonging to a given category is twice as likely if it is within that category's boundary (represented as a colored region on the display).



Using the radio buttons, allocate troops to one of the groups, and click Next to continue.

Exam: Pilot Exam Participant: TEST Phase: Mission 1 **(1/7)** Trial: **20/100**, Part: **3/5**

Figure 3: Sample output from Task 1. Participants must generate the likelihood that a probe event (denoted by the "?") was produced by each category and then perform a forced-choice resource allocation to maximize their trial score. Likelihoods are based on the distance from each category's centroid and the frequency of events. For instance, Aqua has a higher likelihood because its centroid is closer to the probe and it has a higher frequency (i.e., more events) than Bromine.

to update their likelihoods based on this information and the corresponding probabilistic decision rule.

When all the layers have been presented (two layers in Task 4, five layers in Task 5, and four layers in Task 6), participants were required to generate a resource allocation. In these Tasks, the resource allocation response was produced using the same interface as probability estimates. For instance, assuming that resources were allocated such that $\{A = 40\%, B = 30\%, C = 20\%, D = 10\%\}$ and if the probe belonged to category A (i.e., that A was the "ground truth"), then the participant would receive a score of 40 out of 100, whereas if the probe instead belonged to category B, they would score 30 points. The resource allocation score provided the means of measuring the probability matching bias. The optimal solution (assuming one could correctly predict the right category with over 25% accuracy) would be to always allocate 100% of one's resources to the category with the highest probability. Allocating anything less than that could be considered an instance of probability matching.

Finally, participants were not allowed to use any assistive device (e.g., pen, paper, calculator, or other external devices), as the intent of the Task was to measure how well participants were able to make rapid probability estimates without any external aids.

2.1. Task 1. In Task 1, participants predicted the likelihood that a probe event belonged to either of two categories {Aqua or Bromine}. Categories were defined by a dispersion value around a centroid location (e.g., central tendency), with individual events produced probabilistically by sampling in a Gaussian window using a similar function as seen in prototype distortion methodologies from dot pattern categorization studies [13]. The interface was presented spatially on a computer screen (see Figure 3) in a 100 × 100 grid pattern (representing 30 square miles; grid not shown).

Participants were instructed to learn about each category's tendencies according to three features: the category's center of activity (i.e., centroid), the dispersion associated with each category's events, and the frequency of events for each category. Using these three features, participants determined the likelihood that the probe event belonged to each category.

A trial consisted of 10 events, with 9 events presented sequentially at various locations about the interface, with participants required to click "next" after perceiving each event. The 10th event was the probe event, which was presented as a "?" on the interface. Each participant completed 10 trials, with events accumulating across trials such that 100 events were present on the interface by the end of the task.

After perceiving the probe event, participants were instructed to generate likelihoods that the probe event belonged to each category based on all the events that they have seen not just the recent events from the current trial. These likelihoods were expressed on a scale from 1 to 99% for each category and summing to 100% across both categories. If necessary, the interface would automatically normalize all likelihoods into probabilities summing to 100%.

Finally, participants entered a forced-choice resource allocation response, analogous to a measure of certainty. Resource allocation was a forced-choice decision to allocate 100% of their resources to a single category. If that category produced the probe event, then the participant would receive feedback that was either 100 out of 100 for choosing the

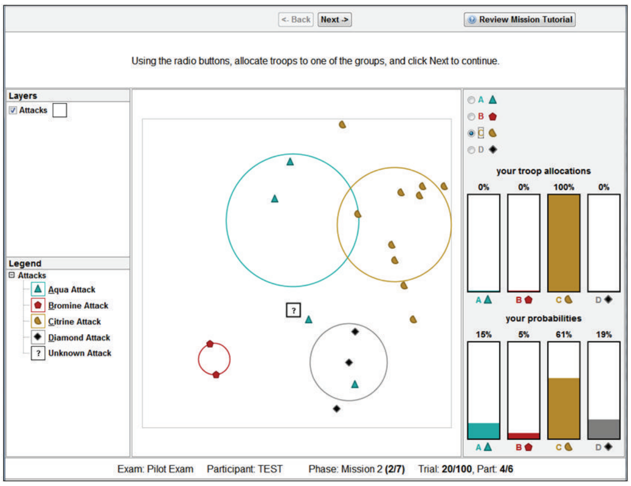


Figure 4: Sample output from Task 2. Participants must generate the likelihood that a probe event (denoted by the "?") was produced by each category and then do a forced-choice resource allocation to maximize their trial score. In addition, participants had to draw a 2-to-1 boundary for each category whose boundary encapsulates 2/3 of that category's events and whose center represents the center of activity for that category. Likelihoods are based on the distance from each category's centroid and the frequency of events. For instance, Citrine has the highest likelihood because it has a higher frequency than the other categories, while Diamond has a marginally higher likelihood than Aqua and Bromine because it has the closest distance.

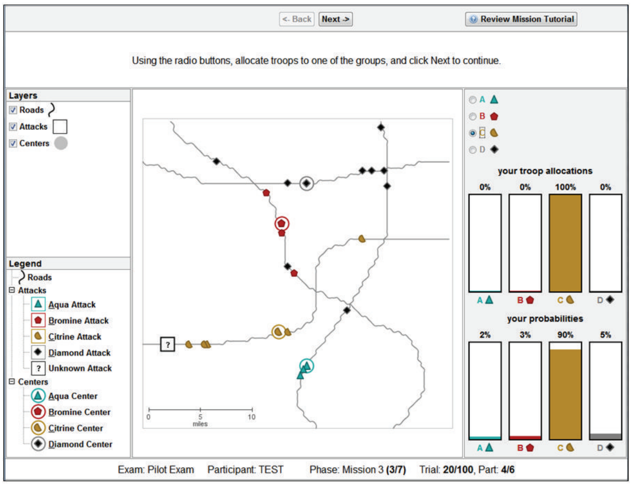


Figure 5: Sample output from Task 3. Participants must generate the likelihood that a probe event (denoted by the "?") was produced by each category and then do a forced-choice resource allocation to maximize their trial score. Likelihoods are based on the road distance from each category's centroid and the frequency of events. For instance, Citrine has the highest likelihood because it is the closest category.

correct category or 0 out of 100 for choosing an incorrect category. Following this feedback, the next trial commenced.

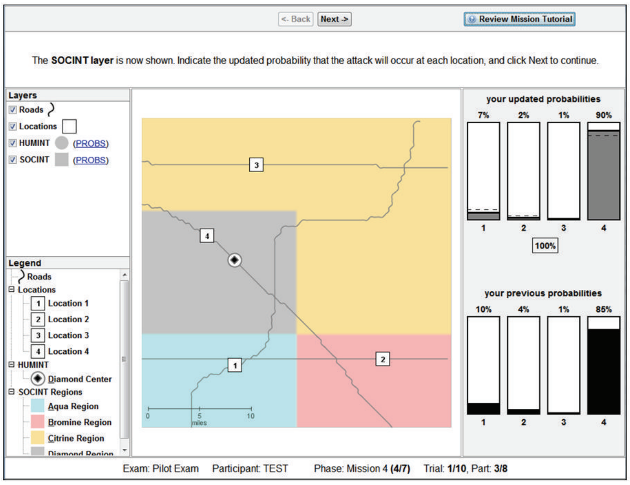
2.2. Task 2. In Task 2, participants predicted the likelihood that a probe event belonged to either of four categories {Aqua, Bromine, Citrine, or Diamond}. The interface and procedure were similar to Task 1, with the following differences. A trial consisted of 20 events, with 19 events presented sequentially at various locations about the interface. The 20th event was the probe event, which was presented as a "?" on the interface. Each participant completed 5 trials, with events accumulating across trials such that 100 events were present on the interface by the end of the task. Participants were further required to estimate each category's centroid and dispersion by drawing a circle for each category representing a 2-to-1 boundary with 2/3 of the category's events inside the circle and 1/3 outside (see Figure 4). Participants clicked with the mouse to set the centroid and dragged out with the mouse to capture the 2-to-1 boundary, releasing the mouse to set the position. It was possible to adjust both the position and dispersion for each category after their initial set. Estimating category centroids and dispersion preceded generating likelihoods.

Finally, participants entered a similar forced-choice resource allocation response as in Task 1. Resource allocation was a forced-choice decision to allocate 100% of their resources to a single category. If that category produced the probe event, then the participant would receive feedback that was either 100 out of 100 for choosing the correct category or 0 of out 100 for choosing an incorrect category. Following this feedback, the next trial commenced.

2.3. Task 3. In Task 3, participants predicted the likelihood that a probe event belonged to either of four categories similar to Task 2, with the following differences. Instead of the interface instantiating a blank grid, it displayed a network of roads. Events were only placed along roads, and participants were instructed to estimate distance along roads rather than "as the crow flies." In addition, participants no longer had to draw the 2-to-1 boundaries but instead only identify the location of the category centroid.

A trial consisted of 20 events, with 19 events presented sequentially at various locations about the interface. The 20th event was the probe event, which was presented as a "?" on the interface. Each participant completed 5 trials, with events accumulating across trials such that 100 events were present on the interface by the end of the task. Participants were further required to estimate each category's centroid by placing a circle for each category (see Figure 5). Participants clicked with the mouse to set the centroid. It was possible to adjust the position for each category after the initial set. Based on the requirement to judge road distance, Task 3 (and Task 4) involved additional visual problem solving strategies (e.g., spatial path planning and curve tracing) [14].

Finally, participants entered a similar forced-choice resource allocation response as in Task 2. Resource allocation was a forced-choice decision to allocate 100% of their resources to a single category. If that category produced the probe event, then the participant would receive feedback that was either 100 out of 100 for choosing the correct category or 0 out of 100 for choosing an incorrect category. Following this feedback, the next trial commenced.



The **SOCINT layer** is now shown. Indicate the updated probability that the attack will occur at each location, and click Next to continue.

Exam: Pilot Exam Participant: TEST Phase: Mission 4 **(4/7)** Trial: **1/10**, Part: **3/8**

Figure 6: Sample output from Task 4. Participants must generate the likelihood that a probe event (denoted by the Diamond) was produced by each category (1–4), first by the HUMINT layer (distance from category centroids to probe event) and then by the SOCINT layer (likelihoods are doubled for the category in whose region the probe event falls). Finally, participants allocate resources to maximize their trial score. For instance, category 4 has the highest likelihood because it is the closest category and the probe falls within its boundary.

2.4. Task 4. Beginning with Task 4, instead of gathering information from a sequence of events, participants instead generated and updated likelihoods after being presented with a number of features as separate layers of information. These features were governed by probabilistic decision rules [15] described previously in Table 1. In Task 4, two features were presented to participants in a fixed order. The first layer was HUMINT (HUMan INTelligence), which revealed the location of the category centroid for each category. The second layer was SOCINT (SOCial INTelligence), which revealed color-coded regions on the display representing each category's boundary (see Figure 6). If a probe event occurred in a given category's boundary, then the probability that the probe belonged to that category was twice as high as the event belonging to any of the other categories.

Participants were instructed that the feature layers provided "clues" revealing intelligence data (called INTs) and the probabilistic decisions rules (called PROBs rules) provided means to interpret INTs. Participants were instructed to refer to the PROBs handbook (based on Table 1; see the appendix for the complete handbook), which was accessible by clicking on the particular layer in the legend on the left side of the display or by reviewing the mission tutorial in the top-right corner. They were further instructed that each feature layer was independent of other layers.

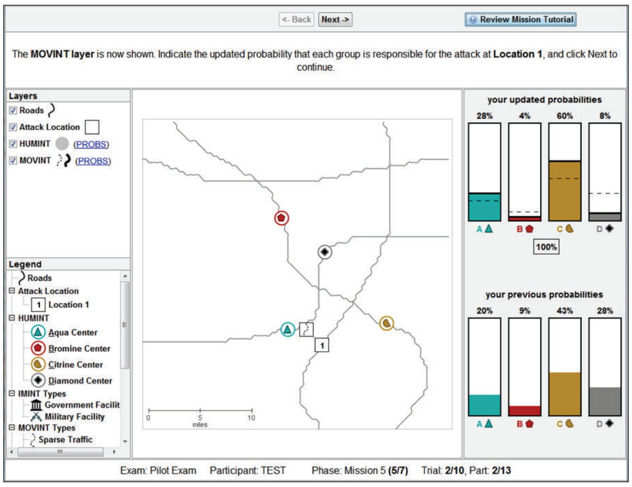
The same simulated geospatial display from Task 3 was used in Task 4; however, instead of a trial consisting of a series of individual events, a trial instead consisted of reasoning from category centroids to a probe event by updating likelihoods after each new feature was revealed. A trial consisted of two features presented in sequence (HUMINT and SOCINT, resp.). The HUMINT layer revealed the centroids for each category along with the probe event. Participants reported likelihoods for each category {1, 2, 3, or 4} based on the road distance between the probe and each category's centroid. Similar to previous tasks, likelihoods were automatically normalized to a probability distribution (i.e., summing to 100%). After this initial distribution was input, the SOCINT feature was presented by breaking the display down into four colored regions representing probabilistic category boundaries. Using these boundaries, participants applied the SOCINT rule and updated their probability distribution.

Once their revised probability distribution was entered, participants were required to generate a resource allocation. The resource allocation response was produced using the same interface as probability estimates. For instance, assuming that resources were allocated such that $\{1 = 40\%, 2 = 30\%, 3 = 20\%, 4 = 10\%\}$ and if the probe belonged to category 1 (i.e., that 1 was the "ground truth"), then the participant would receive a score of 40 out of 100, whereas if the probe instead belonged to category 2, they would score 30 points. After completing their resource allocation, the display was reset and a new trial started.

Participants completed 10 trials. Unlike Tasks 1–3, each trial was presented on a unique road network with all four category locations presented in a unique location.

2.5. Task 5. In Task 5, all five features were revealed to participants in each trial, with the HUMINT feature always revealed first (and the rest presented in a random order). Thus, participants began each trial with each category's centroid presented on the interface and the Bayesian optimal probability distribution already input on the right-side response panel (see Figure 7). The goal of Task 5 was to examine how participants fused multiple layers of information together. Unlike Task 4, the correct probability distribution for HUMINT was provided to participants. This was done both to reduce the variance in initial probabilities (due to the noisiness of spatial road distance judgments) and also to reduce participant fatigue. After perceiving HUMINT and being provided the correct probability distribution, each of the four remaining features (SOCINT, IMINT, MOVINT, and SIGINT on a single category) was revealed in a random order. After each feature was revealed, participants updated their probability distribution based on applying the corresponding decision rules. Similar to Task 4, after the final feature was revealed, participants allocated resources.

The same methodology was used as for Task 4, only with five layers of features presented instead of two. Participants reported likelihoods for each category {Aqua, Bromine, Citrine, or Diamond} based on the information revealed by the feature at each layer according to the rules of the PROBs handbook. Likelihoods were automatically normalized to a probability distribution (i.e., summing to 100%). After HUMINT was revealed, four other features (SOCINT, MOVINT, IMINT, and SIGINT) were revealed in random order. The SOCINT feature was presented by breaking the display down into four colored regions representing probabilistic category boundaries. The IMINT (IMagery INTelligence) feature was presented by showing either a government or military building at the probe location. The MOVINT



The **MOVINT layer** is now shown. Indicate the updated probability that each group is responsible for the attack at **Location 1**, and click Next to continue.

Exam: Pilot Exam Participant: TEST Phase: Mission 5 **(5/7)** Trial: **2/10**, Part: **2/13**

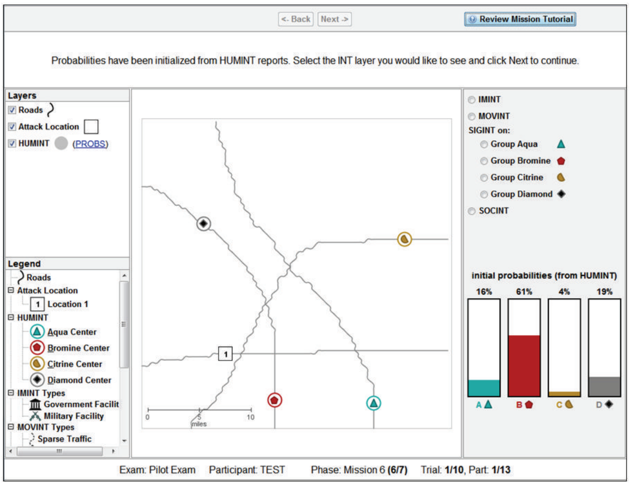
Figure 7: Sample output from Task 5. Participants must generate the likelihood that a probe event (denoted by the probe event "1") was produced by each category. The HUMINT layer is always displayed first, and the initial probability distribution based on road distance is provided to participants. Participants must update this initial distribution as new features are revealed. In the current example, the likelihoods of categories A and C are increased due to the MOVINT layer revealing sparse traffic at the probe event location (see PROBs rules in Table 1).

(MOVement INTelligence) feature was presented by showing either sparse or dense traffic at the probe location. Finally, the SIGINT (SIGnal INTelligence) feature was presented by showing either the presence or absence of chatter for a specific category at the probe location. After each feature was revealed, participants applied the relevant PROBs rule and updated their probability distribution.

After all feature layers were revealed and probability distributions were revised, participants were required to generate a resource allocation. The resource allocation response was produced using the same interface as in Task 4. After completing their resource allocation, the display was reset and a new trial started. Participants completed 10 trials. Note that participants needed to update likelihoods four times per trial (thus 40 times in total) in addition to a single resource allocation per trial (10 total). Similar to Task 4, each trial was presented on a unique road network with all four category locations presented in a unique location.

2.6. Task 6. In Task 6, participants were able to choose three of four possible features to be revealed, in addition to the order in which they are revealed (see Figure 8). The goal of Task 6 was to determine participants' choices and ordering in selecting features (which we refer to as layer selection). This methodology determined whether participants were biased to pick features whose corresponding decision rule confirmed their leading hypothesis or possibly maximized potential information gain. Participants were instructed to choose layers that maximized information at each step to increase the likelihood of a single category being responsible for the event.

As for Task 5, a trial began by perceiving the HUMINT layer and being provided the correct probability distribution. Participants must then choose a feature to be revealed (SOCINT, IMINT, MOVINT, or SIGINT on a single category). When participants chose the SIGINT layer, they needed to further specify which category they were inspecting (listening for chatter). After the chosen feature was revealed, participants updated their probability distribution based on applying the corresponding decision rules. This process was repeated twice more with different features, for a total of three layers being chosen. Participants must update category likelihoods {Aqua, Bromine, Citrine, or Diamond} after each layer was revealed based on the information provided by the corresponding feature at each layer according to the rules of the PROBs handbook. As in the other tasks, likelihoods were automatically normalized to sum to 100% across categories. Note that with only three layer selection choices, participants were not able to reveal one feature on each trial.



Probabilities have been initialized from HUMINT reports. Select the INT layer you would like to see and click Next to continue.

Exam: Pilot Exam Participant: TEST Phase: Mission 6 **(6/7)** Trial: **1/10**, Part: **1/13**

Figure 8: Sample output from Task 6. Participants must generate the likelihood that a probe event (denoted by the probe event "1") was produced by each category. The HUMINT layer is always displayed first, and the initial probability distribution based on road distance is provided to participants. Participants must update this initial distribution as new features are revealed. In the current example, the likelihoods of categories A and C are increased due to the MOVINT layer revealing sparse traffic at the probe event location.

After participants completed the process of choosing a feature and updating their likelihoods for each of three iterations, participants were required to generate a resource allocation. The resource allocation response was produced using the same interface as in Tasks 4-5. After completing their resource allocation, the display was reset and a new trial commenced. Participants completed 10 trials. Note that, with three layer selections, participants actually updated probabilities 30 times (3 times per trial), in addition to allocating resources once for each trial. Similar to Tasks 4-5, each trial was presented on a unique road network with all four category locations presented in a unique location.

3. An ACT-R Model of Sensemaking

3.1. Overview of ACT-R. Our aim has been to develop a functional model of several core information-foraging and hypothesis-updating processes involved in sensemaking. We do this by developing ACT-R models to specify how elementary cognitive modules and processes are marshaled to produce observed sensemaking behavior in a set of complex geospatial intelligence Tasks. These tasks involve an iterative process of obtaining new evidence from available sources and using that evidence to update hypotheses about potential outcomes. One purpose of the ACT-R functional model is to provide a roadmap of the interaction and sequencing of neural modules in producing task performance (see next section). A second purpose is to identify and understand a core set mechanisms for producing cognitive biases observed in the selection and weighting of evidence in information foraging (e.g., confirmation bias).

The ACT-R architecture (see Figure 9) is organized as a set of modules, each devoted to processing a particular kind of information, which are integrated and coordinated through a centralized production system module. Each module is assumed to access and deposit information into buffers associated with the module, and the central production system can only respond to the contents of the buffers not the internal encapsulated processing of the modules. Each module, including the production module, has been correlated with activation in particular brain locations [1]. For instance, the visual module (occipital cortex and others) and visual buffers (parietal cortex) keep track of objects and locations in the visual field. The manual module (motor cortex; cerebellum) and manual buffer (motor cortex) are associated with control of the hands. The declarative module (temporal lobe; hippocampus) and retrieval buffer (ventrolateral prefrontal cortex) are associated with the retrieval and awareness of information from long-term declarative memory. The goal buffer (dorsolateral prefrontal cortex) keeps track of the goals and internal state of the system in problem solving. Finally, the production system (basal ganglia) is associated with matching the contents of module buffers and coordinating their activity. The production includes components for pattern matching (striatum), conflict resolution (pallidum), and execution (thalamus). A production rule can be thought of as a formal specification of the flow of information from buffered information in the cortex to the basal ganglia and back again [16].

The declarative memory module and production system module, respectively, store and retrieve information that corresponds to declarative knowledge and procedural knowledge [17]. Declarative knowledge is the kind of knowledge that a person can attend to, reflect upon, and usually articulate in some way (e.g., by declaring it verbally or by gesture). Procedural knowledge consists of the skills we display in our behavior, generally without conscious awareness. Declarative knowledge in ACT-R is represented formally in terms of chunks [18, 19]. The information in the declarative memory module corresponds to personal episodic and semantic knowledge that promotes long-term coherence in behavior. In this sense a chunk is like a data frame, integrating information available in a common context at a particular point in time in a single representational structure. The goal module stores and retrieves information that represents the internal intention and problem solving state of the system and provides local coherence to behavior.

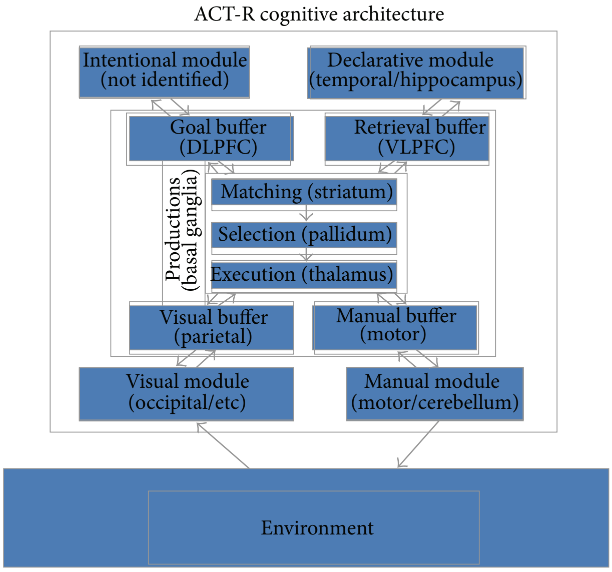

ACT-R cognitive architecture
Intentional module (not identified)
Declarative module (temporal/hippocampus)
Goal buffer (DLPFC)
Retrieval buffer (VLPFC)
Productions (basal ganglia)
Matching (striatum)
Selection (pallidum)
Execution (thalamus)
Visual buffer (parietal)
Manual buffer (motor)
Visual module (occipital/etc)
Manual module (motor/cerebellum)
Environment

Figure 9: ACT-R functions as a production system architecture with multiple modules corresponding to different kinds of perception, action, and cognitive information stores. Modules have been identified with specific brain regions. In addition to those shown above, the Imaginal module has been associated with posterior parietal activation.

Chunks are retrieved from long-term declarative memory by an activation process (see Table 2 for a list of retrieval mechanisms in ACT-R). Each chunk has a base-level activation that reflects its recency and frequency of occurrence. Activation spreads from the current focus of attention, including goals, through associations among chunks in declarative memory. These associations are built up from experience, and they reflect how chunks cooccur in cognitive processing. The spread of activation from one cognitive structure to another is determined by weighting values on the associations among chunks. These weights determine the rate of activation flow among chunks. Chunks are compared to the desired retrieval pattern using a partial matching mechanism that subtracts from the activation of a chunk its degree of mismatch to the desired pattern, additively for each component of the pattern and corresponding chunk value. Finally, noise is added to chunk activations to make retrieval a probabilistic process governed by a Boltzmann (softmax) distribution. While the most active chunk is usually retrieved, a blending process [20] can also be applied which returns a derived output reflecting the similarity between the values of the content of all chunks, weighted by their retrieval probabilities reflecting their activations and partial-matching scores. This blending process will be used intensively in the model since it provides both a tractable way to learn to perform decisions in continuous domains such as the probability spaces of the AHA framework and a direct abstraction to the storage and retrieval of information in neural models (see next section).

Table 2: The list of sub-symbolic mechanisms in the ACT-R architecture.

Mechanism	Equation	Description
Activation	$A_i = B_i + S_i + P_i + \varepsilon_i$	B_i: base-level activation reflects the recency and frequency of use of chunk i S_i: spreading activation reflects the effect that buffer contents have on the retrieval process P_i: partial matching reflects the degree to which the chunk matches the request ε_i: noise value includes both a transient and (optional) permanent components (permanent component not used by the integrated model)
Base level	$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + \beta_i$	n: the number of presentations for chunk i t_j: the time since the jth presentation d: a decay rate (not used by the integrated model) β_i: a constant offset (not used by the integrated model)
Spreading activation	$S_i = \sum_k \sum_j W_{kj} S_{ji}$, $S_{ji} = S - \ln\left(\text{fan}_{ji}\right)$	k: weight of buffers summed over are all of the buffers in the model j: weight of chunks which are in the slots of the chunk in buffer k W_{kj}: amount of activation from sources j in buffer k S_{ji}: strength of association from sources j to chunk i S: the maximum associative strength (set at 4 in the model) fan_{ji}: a measure of how many chunks are associated with chunk j
Partial matching	$P_i = \sum_k P M_{ki}$	P: match scale parameter (set at 2) which reflects the weight given to the similarity M_{ki}: similarity between the value k in the retrieval specification and the value in the corresponding slot of chunk i The default range is from 0 to −1 with 0 being the most similar and −1 being the largest difference
Declarative retrievals	$P_i = \dfrac{e^{A_i/s}}{\sum_j e^{A_j/s}}$	P_i: the probability that chunk i will be recalled A_i: activation strength of chunk i $\sum A_j$: activation strength of all of eligible chunks j s: chunk activation noise
Blended retrievals	$V = \min \sum_i P_i \left(1 - \text{Sim}_{ij}\right)^2$	P_i: probability from declarative retrieval Sim_{ij}: similarity between compromise value j and actual value i
Utility learning	$U_i(n) = U_i(n-1) + \alpha\left[R_i(n) - U_i(n-1)\right]$ $P_i = \dfrac{e^{U_i/s}}{\sum_j e^{U_j/s}}$	$U_i(n-1)$: utility of production i after its $n-1$st application $R_i(n)$: reward production received for its nth application $U_i(n)$: utility of production i after its nth application P_i: probability that production i will be selected U_i: expected utility of the production determined by the utility equation above U_j: the expected utility of the competing productions j

Production rules are used to represent procedural knowledge in ACT-R. That is, they specify procedures that represent and apply cognitive skill (know-how) in the current context and how to retrieve and modify information in the buffers and transfer it to other modules. In ACT-R, each production rule has conditions that specify structures that are matched in buffers corresponding to information from the external world or other internal modules. Each production rule has actions that specify changes to be made to the buffers.

ACT-R uses a mix of parallel and serial processing. Modules may process information in parallel with one another. So, for instance, the visual modules and the motor modules may both operate at the same time. However, there are two serial bottlenecks in process. First, only one production may be executed during a cycle. Second, each module is limited to placing a single chunk in a buffer. In general, multiple production rules can be applied at any point. Production utilities, learned using a reinforcement learning scheme, are used to select the single rule that fires. As for declarative memory retrieval, production selection is a probabilistic process.

Cognitive model development in ACT-R [21] is in part derived from the rational analysis of the task and information structures in the external environment (e.g., the design of the tasks being simulated or the structure of a graphical user interface), the constraints of the ACT-R architecture, and guidelines from previous models of similar tasks. A successful design pattern in specifying cognitive process sequencing in ACT-R [21] is to decompose a complex task to the level of unit tasks [22]. Card et al. [22] suggested that unit tasks control immediate behavior. Unit tasks empirically take about 10 seconds. To an approximation, unit tasks are where “the rubber of rationality meets the mechanistic road.” To an approximation, the structure of behavior above the unit task level largely reflects a rational structuring of the task within the constraints of the environment, whereas the structure within and below the unit task level reflects cognitive and biological mechanisms, in accordance with Newell's bands of cognition [23]. Accordingly, in ACT-R, unit tasks are implemented by specific goal types that control productions that represent the cognitive skills for solving those tasks.

ACT-R has been the basis for several decades of research on learning complex cognitive tasks such as algebra and programming [24, 25]. In general, the long-run outcome of learning such tasks is a large set of highly situation-specific productions whose application is sharply tuned by ACT-R utility mechanisms (a form of reinforcement learning). However, it is also generally assumed that achieving such expert levels of learning requires 1000s of hours of experience. We assume that the participants in the AHA tasks will not have the opportunity to achieve such levels of expertise. Instead, we hypothesize that participants will rely on direct recognition or recall of relevant experience from declarative memory to guide their thinking or, failing that, will heuristically interpret and deliberate through the rules and evidence provided in the challenge tasks. This compute-versus-retrieve process is another design pattern that typically structures ACT-R models [21]. The notion that learners have a general-purpose mechanism whereby situation-action-outcome observations are stored and retrieved as chunks in ACT-R declarative memory is derived from instance-based learning theory (IBLT) [26, 27]. Gonzalez et al. [26] present arguments that IBLT is particularly pertinent to modeling naturalistic decision making in complex dynamic situations, and many of those arguments would transfer to making the case that IBLT is appropriate for sensemaking.

Relevant to the Bayesian inspiration for the AHA tasks, ACT-R's subsymbolic activation formula approximates Bayesian inference by framing activation as log-likelihoods, base-level activation (B_i) as the prior, the sum of spreading activation and partial matching as the likelihood adjustment factor(s), and the final chunk activation (A_i) as the posterior. The retrieved chunk has an activation that satisfies the maximum likelihood equation. ACT-R provides constraint to the Bayesian framework through the activation equation and production system. The calculation of base levels (i.e., priors) occurs within both neurally and behaviorally consistent equations (see Table 2) providing for behaviorally relevant memory effects like recency and frequency while also providing a constrained mechanism for obtaining priors (i.e., driven by experience).

In addition, the limitations on matching in the production system provide constraints to the Bayesian hypothesis space and, as a result, the kinds of inferences that can be made. For instance, there are constraints on the kinds of matching that can be accomplished (e.g., no disjunction, matching only to specific chunk types within buffers), and, while user-specified productions can be task-constrained, the production system can generate novel productions (through proceduralization of declarative knowledge) using production compilation. In addition, the choice of which production to fire (conflict resolution) also constrains which chunks (i.e., hypotheses) will be recalled (limiting the hypothesis space) and are also subject to learning via production utilities.

It has been argued that ACT-R's numerous parameters do not provide sufficient constraint on modeling endeavors. However, the use of community and research-justified default values, the practice of removing parameters by developing more automatized mechanisms, and the development of common modeling paradigms—such as instance-based learning theory—mitigate these criticisms by limiting degrees of freedom in the architecture and thus constraining the kinds of models that can be developed and encouraging their integration.

3.2. ACT-R Prototyping for Neural Models. ACT-R can be used in a prototyping role for neural models such as Emergent, which uses the Leabra learning rule [28]. In ACT-R, models can be quickly developed and tested, and the results of these models then help inform modeling efforts and direct training strategies in Emergent models [29, 30]. ACT-R models can be created quickly because ACT-R models accept predominantly functional specifications, yet they produce neurally relevant results. The ACT-R architecture is also flexible enough that innovations made in neurocomputational models can be implemented (to a degree of abstraction) within new ACT-R modules [31].

There are several points of contact between ACT-R and Emergent, the most tangible of which is a commitment to neural localization of architectural constructs in both architectures (see Figure 9). In both architectures a central control module located in the basal ganglia collects inputs from a variety of cortical areas and outputs primarily to the frontal cortex, which maintains task relevant information [16, 32]. Additionally, both include a dedicated declarative/episodic memory system in the hippocampus and associated cortical structures. Lastly, both account for sensory and motor processing in the posterior cortex.

The architectures differ in that the brain regions are explicitly modeled in Emergent, whereas they are implicit in ACT-R. In ACT-R the basal ganglia are associated with the production system; the frontal cortex with the goal module; the parietal cortex with the imaginal module; the hippocampus with the declarative memory module; and finally the posterior cortices with the manual, vocal, aural, and vision modules. This compatibility of ACT-R and Emergent has been realized elsewhere by the development of SAL (Synthesis of ACT-R and Leabra/Emergent), a hybrid architecture that combines ACT-R and Emergent and exploits the relative strengths of each [33]. Thus, ACT-R connects to the underlying neural theory of Emergent and can provide meaningful guidance to the development of neural models of complex tasks, such as sensemaking.

In effect, ACT-R models provide a high-level specification of the information flows that will take place in the neural model between component regions implemented in Emergent. Since ACT-R models have been targeted at precisely this level of description, they can provide for just the right level of abstraction while ignoring many implementational details (e.g., number of connections) at the neural level.

Conceptually, the ACT-R architecture provides a bridge between the rational Bayesian level and the detailed neural level. In terms of Marr [34] levels of analysis, the Bayesian characterization of the task solutions is situated at the computational level, describing the computations that should be performed without specifying how. An ACT-R account of the tasks is at the algorithmic/representational level, specifying what representations are created, which mechanisms are used to manipulate them, and which structure constrains both representations and processes. Finally, a neural account is situated at the physical/implementational level, fully specifying all the details of how the computations are to be carried out in the brain. Thus, just as in Marr's analysis it would not make sense to try to bridge directly the highest and lowest levels; a functional cognitive architecture such as ACT-R provides a critical link between abstract computational specifications such as Bayesian rational norms and highly detailed neural mechanisms and representations.

Moreover, ACT-R does not just provide any intermediate level of abstraction between computational and implementational levels in a broad modular sense. Rather, just as the ACT-R mechanisms have formal Bayesian underpinnings, they also have a direct correspondence to neural mechanisms and representations. The fundamental characteristics of modern neural modeling frameworks are distributed representations, local learning rules, and training of networks from sets of input-output instances [35].

Distributed representations are captured in ACT-R through similarities between chunks (and other sets of values such as number magnitudes) that can be thought of as corresponding to the dotproduct between distributed representations of the corresponding chunks. The generalization process operates over distributed representations in neural networks, effectively matching the learned weights from the input units resulting in a unit containing the representation of the current input. This is implemented in ACT-R using a partial matching mechanism that combines a chunk's activation during the memory retrieval process with its degree of match to the requested pattern as determined by the similarities between chunk contents and pattern [36].

Local learning rules in neural networks are used to adjust weights between units based on information flowing through the network. The base-level and associative learning mechanisms in ACT-R perform a similar function in the same manner. Both have Bayesian underpinnings [37] but also direct correspondence to neural mechanisms. Base-level learning is used to adjust the activation of a chunk based on its history of use, especially its frequency and recency of access. This corresponds to learning the bias of a unit in a network, determining its initial activation which is added to inputs from other units. Associative learning adjusts the strengths of association between chunks to reflect their degree of coactivation. While the original formulation was Bayesian in nature, a new characterization makes the link to Hebbian-like learning explicit, in particular introducing the same positive-negative learning phases as found in many connectionist learning algorithms including Leabra [31].

Neural models are created by a combination of modeler-designed structure and training that adjusts the network's weights in response to external inputs. The instance-based learning approach in ACT-R similarly combines a representational structure provided by the modeler with content acquired from experience in the form of chunks that represent individual problem instances. The set of chunks stored in declarative memory as a result can be thought of as the equivalent to the set of training instances given to the neural network. While the network compiles those instances into weights during training, ACT-R can instead dynamically blend those chunks together during memory retrieval to produce an aggregate response that reflects the consensus of all chunks, weighted by their probability of retrieval reflecting the activation processes described above [20].

Thus, ACT-R models can be used to prototype neural models because they share both a common structure of information flow as well as a direct correspondence from the more abstract (hence tractable) representations and mechanisms at the symbolic/subsymbolic level and those at the neural level.

3.3. Cognitive Functions Engaged in the AHA Tasks. The integrated ACT-R model of the AHA tasks has been used to prototype many cognitive effects in neural models including generating category prototypes of centroids from SIGACT events (centroid generation) [29], spatial path planning along road networks [30], adjusting probabilities based on incoming information [29], choosing how many resources to allocate given a set of probabilities and prior experience [29], and selecting additional intelligence layers (see Table 3 for an overview). The model's output compared favorably with human behavioral data and provides a comprehensive explanation for the origins of cognitive biases in the AHA framework, most prominently the anchoring and adjustment bias. All the functions described below were integrated in a single ACT-R model that performed all 6 AHA tasks using the same parameters. That model was learned across trials and tasks. We will describe later in details how its performance in the later trials of a task can depend critically upon its experience in the earlier trials (even just the first trial), in particular leading to a distinct conservatism bias. Similarly, its performance in the later tasks depends upon its experience in earlier tasks, leading directly to probability matching bias in resource allocation.

The model performs the task in the same manner as human subjects. Instructions such as the probabilistic decision rules are represented in declarative memory for later retrieval when needed. The model perceives the events, represents them in the imaginal buffer, and then stores them in declarative memory where they influence future judgments. In Tasks 1–3, the model uses those past events in memory to generate the category centroid when given a probe. In Tasks 3-4, the model programmatically parses the map to represent the road network declaratively and then uses that declarative representation to generate paths and estimate road distances. In all tasks, probability adjustment is performed using the same instance-based mechanism, with experience from earlier tasks accumulated in memory for use in later tasks. Resource allocation is also performed in all tasks using the same instance-based approach, with results

Table 3: Overview of cognitive functions of ACT-R model.

Cognitive function	Overview of operation
Centroid generation Tasks: 1–3	Buffers implicated: blending, imaginal, and goal Biases instantiated: base-rate neglect, anchoring and adjustment The model generates a category centroid by aggregating overall of the perceived events (SIGACTs) in memory via the blended memory retrieval mechanism. Judgments are based on generating a centroid-of-centroids by performing a blended retrieval over all previously generated centroids, resulting to a tendency to anchor to early judgments. Because there is an equal number of centroids per category, this mechanism explicitly neglects base rate
Path planning Tasks: 3-4	Buffers implicated: retrieval, imaginal, and goal Biases instantiated: anchoring and adjustment The model parses the roads into a set of intersections and road segments. The model hill-climbs by starting at the category centroid and appends contiguous road segments until the probe event is reached. Road segment lengths are perceived veridically; however, when recalled the lengths are influenced by bottom-up perceptual mechanisms (e.g., curve complexity and length) simulated by a power law with an exponent less than unity. This leads to underestimation of longer and curvier segments, resulting in a tendency to anchor when perceiving long segments
Probability adjustment Tasks: 1–6	Buffers implicated: blending, imaginal, and goal Biases instantiated: anchoring in weighing evidence, confirmation bias The model represents the prior probability and multiplicative factor rule and then attempts to estimate the correct posterior by performing a blended retrieval over similar chunks in memory in a form of instance-based learning. The natural tendency towards regression to the mean in blended retrievals leads to anchoring bias in higher probabilities and confirmation bias in lower probabilities. The partial matching mechanism is used to allow for matches between the prior and similar values in DM
Resource allocation Tasks: 1–6	Buffers implicated: blending, imaginal, and goal Biases instantiated: probability matching The model takes the probability assigned to a category and then estimates an expected outcome by performing a blended retrieval using the probability as a cue. The outcome value of the retrieved chunk is the expected outcome for the trial. Next, an additional blended retrieval is performed based on both the probability and expected outcome, whose output is the resources allocation After feedback, the model stores the leading category probability, the resources allocated, and the actual outcome of the trial. Up to two counterfactuals are learned, representing what would have happened if a winner-take-all or pure probability matching resources allocation had occurred. Negative feedback on forced winner-take-all assignments in Tasks 1–3 leads to probability matching in Tasks 4–6
Layer selection Task: 4–6	Buffers implicated: blending, goal Biases instantiated: confirmation bias In Task 6, the model uses partial matching to find chunks representing past layer-selection experiences that are similar to the current situation (the distribution of probabilities over hypotheses). If that retrieval succeeds, the model attempts to estimate the utility of each potential layer choice by performing a blended retrieval over the utilities of past layer-choice outcomes in similar situations. The layer choice that has the highest utility is selected. If the model fails to retrieve past experiences similar to the current situations, it performs a "look-ahead" search by calculating the expected utility for some feature layers. The number of moves mentally searched will not often be exhaustive The blended retrieval mechanism will tend to average the utility of different feature layers based on prior experiences from Tasks 4 and 5 (where feature layers were provided to participants), in addition to prior trials on Task 6

from forced-choice selections in Tasks 1–3 fundamentally affecting choices in later Tasks 4–6. Finally, layer selection in Task 6 uses experiences in Tasks 4-5 to generate estimates of information gain and select the most promising layer. Thus the integrated model brings to bear constraints from all tasks and functions.

3.3.1. Centroid Generation. The ACT-R integrated model generates category centroids (i.e., the prototype or central tendency of the events) in Tasks 1–3 by aggregating overall of the representations of events (e.g., spatial-context frames) in memory via the blended memory retrieval mechanism. The goal buffer maintains task-relevant top-down information while the blending buffer creates/updates centroids from both the set of perceived SIGACTs to date and prior created centroids. Centroid generation approximates a stochastic least-MSE derived from distance and based on the 2D Cartesian coordinates of the individual SIGACTs. Specifically, the mismatch penalty (P_i) used in the blended retrieval is a linear difference:

$$P_i = \frac{2 \cdot |d_1 - d_2|}{* \max \_\text{range} *}, \tag{1}$$

where d is the perceived distance and $*\text{max\_range}*$ is the size of the display (100 units). The imaginal buffer (correlated with parietal activity) is used to hold blended chunks before being committed to declarative memory. When centroids are generated directly from SIGACTs, the blending process reflects a disproportionate influence of the most recent events given their higher base-level activation. A strategy to combat this recency bias consisted of generating a final response by performing a blended retrieval over the current and past centroids, thereby giving more weight to earlier SIGACTs. This is because the influence of earlier centroids has been compounded over the subsequent blended retrievals, essentially factoring earlier SIGACTs into more centroids. This second-order blended retrieval is done for each category across their prior existing centroids, which we refer to as the generation of a centroid-of-centroids. This blending over centroids effectively implements an anchoring-and-adjustment process where each new centroid estimate is a combination of the previous ones together with the new evidence. A fundamental difference with traditional implementation of anchoring-and-adjustment heuristic is that this process is entirely constrained by the architectural mechanisms (especially blending) and does not involve additional degrees of freedom. Moreover, because there are an equal number of centroid chunks (one per category created after each trial), there is no effect of category base rate on the model's later probability judgments, even though the base rate for each category is implicitly available in the model based on the number of recallable events.

3.3.2. Path Planning. The ACT-R model uses the declarative memory and visual modules to implement path planning, which simulate many of the parietal functionalities that were later implemented in a Leabra model [30]. Two examples include

(1) perceptually segmenting the road network so that the model only attends to task-relevant perceptual elements,

(2) implementing visual curve tracing to model the psychophysics of how humans estimate curved roads.

The model segments the road network in Tasks 3-4 into smaller elements and then focuses perceptual processes such as curve tracing, distance estimation, and path planning on these smaller road segments [38]. Specifically, the model identifies the intersections of different roads as highly salient HUMINT features and then splits the road network into road segments consisting of two intersections (as the ends of the segment), the general direction of the road and the length of road. Intersections are generally represented as a particular location on the display in Cartesian X-Y coordinates.

For each trial, the probe location is also represented as a local HUMINT feature, and, in Task 3, the individual events are represented as local HUMINT features for the purposes of calculating category centroids. At the end of each trial, the probe is functionally removed from the path planning model, although a memory trace of the previous location still remains in declarative memory.

We have implemented a multistrategy hill-climber to perform path planning. The model starts with a category centroid and appends contiguous road segments until the probe location is reached (i.e., the model is generating and updating a spatial-context frame). The path planning "decision-making" is a function of ACT-R's partial matching. In partial matching, a similarity is computed between a source object and all target objects that fit a set of matching criteria. This similarity score is weighted by the mismatch penalty scaling factor. The hill-climber matches across multiple values such as segment length and remaining distance to probe location. In general, for all road segments adjoining the currently retrieved intersection or category centroid, the model will tend to pick the segment where the next intersection is nearest to the probe. This is repeated until the segment with the probe location is retrieved. These strategies are not necessarily explicit but are instead meant to simulate the cognitive weights of different perceptual factors (e.g., distance, direction, and length) guiding attentional processes. The partial matching function generates probabilities from distances and calculates similarity between distances using the same mismatch penalty as in Tasks 1 and 2.

Human performance data on mental curve tracing [14] show that participants take longer to mentally trace along a sharper curve than a relatively narrower curve. This relation is roughly linear (with increased underestimation of total curve length at farther curves) and holds along the range of visual sizes of roads that were seen in the AHA tasks. This modeling assumption is drawn from the large body of the literature on visuospatial representation and mental scanning [39]. When road network is parsed, a perceived length is assigned to each road segment. This length is more or less represented veridically in declarative memory. The dissociation between a veridical perceived magnitude of distance and a postprocessed cognitive distance estimate is consistent with prior literature [40]. We represent a cognitive distance estimate using Stevens' Power Law [41]. Stevens' Power Law is a well-studied relationship between the magnitude of a stimulus and its perceived intensity and serves as a simple yet powerful abstraction of many low-level visual processes not currently modeled in ACT-R.

The function uses the ratio of "as the cow walks" distance to "as the crow flies" distance to create an estimate of curve complexity [41]. The higher the curve complexity, the curvier the road. To represent the relative underestimation of distance for curvier segments, this ratio is raised to an exponent of .82 [41–43]. The effect of this parameter is that, for each unit increase in veridical distance, the perceived distance is increased by a lesser extent. The closer the exponent to 1, the more veridical the perception, and the closer to zero, the more the distance will be underestimated. This value for curve complexity is then multiplied by a factor representing straight-line distance estimation performance (1.02) [43–45]:

$$D = \left\{1.02\left(\frac{\mathrm{Cow}_{\mathrm{Walk}}}{\mathrm{Crow}_{\mathrm{Flies}}}\right) + \mathrm{Crow}_{\mathrm{Flies}}\right\}^{.82}, \qquad (2)$$

where D is the cognitive judgment of distance for the road segment, $\mathrm{Cow}_{\mathrm{Walk}}$ is the veridical perception of the curvature of the road, and $\mathrm{Crow}_{\mathrm{Flies}}$ is the veridical perception of the Euclidean distance between the source and target locations.

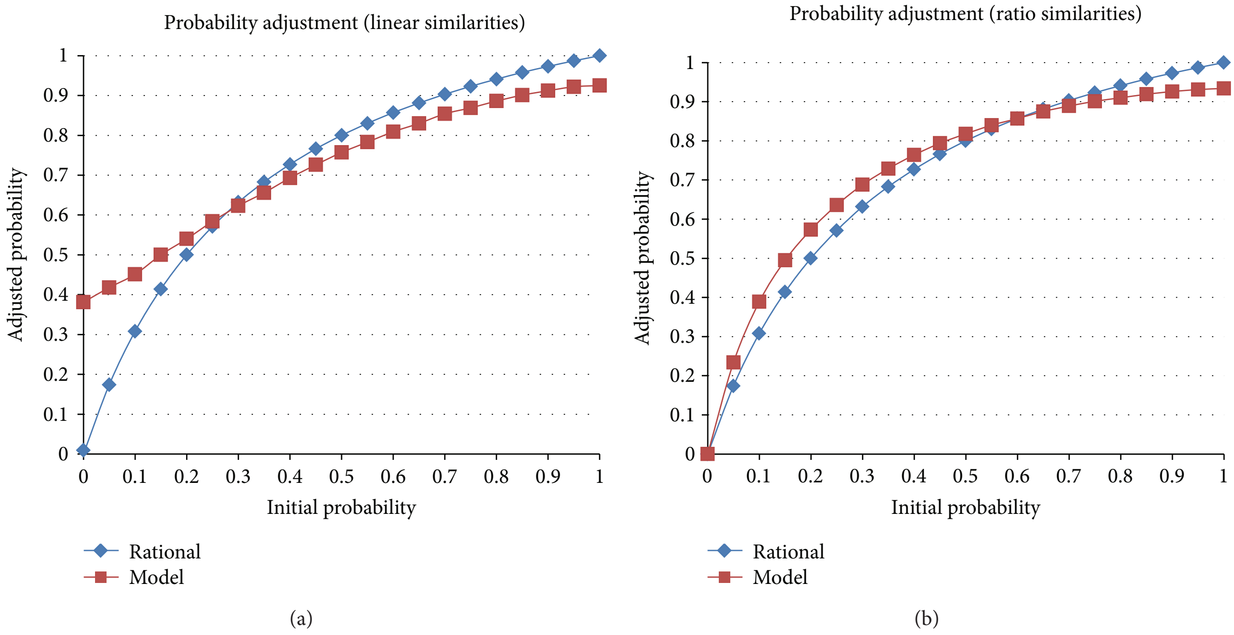


Figure 10: Results from an ACT-R model of probability adjustment with linear (a) and ratio (b) similarities.

The factor of 1.02 represents a slight overestimation of smaller straight-line distances. Similar to the exponent, any factor above unity represents an overestimation of distance and any factor below unity represents an underestimation of distance.

3.3.3. Probability Adjustment. Lebiere [20] proposed a model of cognitive arithmetic that used blended retrieval of arithmetic facts to generate estimates of answers without explicit computations. The model was driven by number of similarities that correspond to distributed representations for number magnitudes in the neural model and more generally to our sense of numbers [46]. It used partial matching to match facts related to the problem and blended retrievals to merge them together and derive an aggregate estimated answer. The model reproduced a number of characteristics of the distribution of errors in elementary school children, including both table and nontable errors, error gradients around the correct answer, higher correct percentage for tie problems, and, most relevant here, a skew toward underestimating answers, a bias consistent with anchoring and adjustment processes.

To leverage this approach for probability adjustment, the ACT-R model's memory was populated with a range of facts consisting of triplets: an initial probability, an adjustment factor, and the resulting probability. These triplets form the building blocks of the implementation of instance-based learning theory [47] and correspond roughly to the notion of a decision frame [3, 4]. In the AHA framework, the factor is set by the explicit rules of the task (e.g., an event in a category boundary is twice as likely to belong to that category). The model is then seeded with a set of chunks that correspond to a range of initial probabilities and an adjustment factor together with the posterior probability that would result from multiplying the initial probability by the adjustment factor, then normalizing it. When the model is asked to estimate the resulting probability for a given prior and multiplying factor, it simply performs a blended retrieval specifying prior and factor and outputs the posterior probability that represents the blended consensus of the seeded chunks. Figure 10 displays systematic results of this process, averaged over a thousand runs, given the variations in answers resulting from activation noise in the retrieval process. When provided with linear similarities between probabilities (and factors), the primary effect is an underestimation of the adjusted probability for much of the initial probability range, with an overestimate on the lower end of the range, especially for initial values close to 0. The latter effect is largely a result of the linear form of the number similarities function. While linear similarities are simple, they fail to scale both to values near zero and to large values.

A better estimate of similarities in neural representations of numbers is a ratio function, as reflected in single cell recordings [1]. This increases dissimilarities of the numbers near zero and scales up to arbitrarily large numbers. When using a ratio similarity function, the effects from the linear similarity function are preserved, but the substantial overestimate for the lower end of the probability range is considerably reduced. While the magnitude of the biases can be modulated somewhat by architectural parameters such as the mismatch penalty (scaling the similarities) or the activation noise (controlling the stochasticity of memory retrieval), the effects themselves are a priori predictions of the architecture, in particular its theoretical constraints on memory retrieval.

Particular to the integrated model of the AHA tasks, the mismatch penalty (P_i) was computed as a linear difference:

$$P_i = 2 * \left| M_k - M_j \right|, \quad (3)$$

where M_k is the possible target probability and M_j is the probability in the blended retrieval specification. As will be described below, this linear difference matches extremely well to human behavioral data.

The results of the ACT-R mechanism for probability adjustment provided a benchmark against which neural models were evaluated and were used to generate training instances for the neural model which had already embodied the proper bias. In our opinion, this constitutes a novel way to use functional models to quickly prototype and interpret neural models.

3.3.4. Resource Allocation. The resource allocation mechanism in the model makes use of the same instance-based learning paradigm as the probability adjustment mechanism. This unified mechanism has no explicit strategies but instead learns to allocate resources according to the outcomes of past decisions. The model generates a resource allocation distribution by focusing on the leading category and determining how many resources to allocate to that category. The remaining resources are divided amongst the remaining three categories in proportion to their assigned probabilities. This instance-based learning not only occurs during Tasks 4–6 but also in Tasks 1–3 for forced-choice allocations. Because of this, the model has some prior knowledge to draw upon in Task 4 when it first has the opportunity to select how many resources to assign to the leading category.

As mentioned above, this instance-based model has the same structure as the model of probability adjustment. Representation of a trial instance consists of three parts: a decision context (in this case, the probability of the leading category), the decision itself (i.e., the resource allocation to the leading category), and the outcome of the decision (i.e., the payoff resulting from the match of that allocation to the ground truth of the identity of the responsible category). This representation is natural because all these pieces of information are available during a resource allocation instance and can plausibly be bound together in episodic memory. However, the problem is how to leverage it to make decisions.

Decision-making (choice) models based on this instance-based learning approach iterate through a small number of possible decisions, generating outcome expectancies from the match of context and decision, and then choose the decision with the highest expected outcome [47, 48]. Control models apply the reverse logic: given the current context and a goal (outcome) state, they match context and outcome to generate the expected action (usually a control value from a continuous domain) that will get the state closest to the goal [49, 50]. However, our problem does not fit either paradigm: unlike choice problems, it does not involve a small number of discrete actions but rather a range of possible allocation values, and, unlike control problems, there is no known goal state (expected outcome) to be reached.

Our model's control logic takes a more complex hybrid approach, involving two steps of access to experiences in declarative memory rather than a single one. The first step consists of generating an expected outcome weighted over the available decisions given the current context. The second step will then generate the decision that most likely leads to that outcome given to the context. Note that this process is not guaranteed to generate optimal decisions, and indeed people do not. Rather, it represents a parsimonious way to leverage our memory of past decisions in this paradigm that still provides functional behavior. A significant theoretical achievement of our approach is that it unifies control models and choice models in a single decision-making paradigm.

When determining how many resources to apply to the lead category, the model initially has only the probability assigned to that category. The first step is to estimate an expected outcome. This is done by performing a blended retrieval on chunks representing past resource allocation decisions using the probability as a cue. The outcome value of the retrieved chunk is the expected outcome for the trial. Next, based on the probability assigned to the leading category and the expected outcome, an additional blended retrieval is performed. The partial matching mechanism is leveraged to allow for nonperfect matches to contribute to the estimation of expected outcome and resource quantity. The resource allocation value of this second blended allocate chunk is the quantity of resources that the model assigns to the leading category. After feedback is received, the model learns a resource allocation decision chunk that associates the leading category probability, the quantity of resources assigned to the leading category, and the actual outcome of the trial (i.e., the resource allocation score for that trial). Additionally, up to two counterfactual chunks are committed to declarative memory. The counterfactuals represent what would have happened if a winner-take-all resource assignment had been applied and what would have happened if a pure probability-matched resource assignment (i.e., using the same values as the final probabilities) had been applied. The actual nature of the counterfactual assignments is not important; what is essential is to give the model a broad enough set of experience representing not only the choices made but also those that could have been made.

The advantage of this approach is that the model is not forced to choose between a discrete set of strategies such as winner-take-all or probability matching; rather, various strategies could emerge from instance-based learning. By priming the model with the winner-take-all and probability matching strategies (essentially the boundary conditions), it is possible for the model to learn any strategy in between them, such as a tendency to more heavily weigh the leading candidate (referred to as PM+), or even suboptimal strategies such as choosing 25% for each of the four categories (assuring a score of 25 on the trial) if the model is unlucky enough to receive enough negative feedback so as to encourage risk aversion [47].

3.3.5. Layer Selection. Layer selection in Task 6 depends on learning the utilities of layer choices in Tasks 4-5 and relies on four processes: instance-based learning (similar to probability adjustment and resource allocation mechanisms), difference reduction heuristic, reinforcement learning, and cost-satisfaction. During Tasks 4–6 participants were asked to update probability distributions based on the outcome of each layer (i.e., feature and relevant decision rule). In Tasks 4-5, participants experienced 20 instances of the SOCINT rule and 10 instances each of the IMINT, MOVINT, and SIGINT rules. They also had a variable number of instances from Task 6 based on their layer selections. Some of the layers and outcomes might support their preferred hypothesis, but some of them might not. Based on the results of each layer's outcome, the gain towards the goal of identifying a single category might vary, and those experiences affect future layer selection behavior through reinforcement learning.

A rational Bayesian approach to Tasks 4–6 might involve the computation of expected information gains (EIGs) computed overall possible outcomes that might result from the selection of a feature layer. Under such a rational strategy, SIGINT and SOCINT layers would require more calculation cost than IMINT and MOVINT. In particular, the calculation of EIG for SIGINT requires considering four categories with two outcomes each, and the calculation of EIG for SOCINT requires four outcomes; however, the calculation of EIG for an IMINT or MOVINT layer requires consideration of only two outcomes. We assume participants might consider the cognitive costs of exploring layer selection outcomes in preferring certain layer selection.

We chose to use a difference reduction heuristic (i.e., hill-climbing) because we assume that an average person is not able to compute and maintain the expected information gain for all layers. A hill-climbing heuristic enables participants to focus on achieving states that are closer to an ideal goal state with the same requirement for explicit representation, because all that needs to be represented is the difference between the current state and a preferred (i.e., goal) state.

In Task 6, all prior instances were used to perform evaluations of layer selection. First, the model attends to the current problem state, including the distribution of likelihood of attacks, as represented in the goal (Prefrontal Cortex; PFC) and imaginal (Parietal Cortex; PC) buffers. Then, the model attempts to retrieve a declarative memory chunk (Hippocampus/Medial Temporal Lobe; HC/MTL) which encodes situation-action-outcome-utility experiences of past layer selections. This mechanism relies on partial matching to retrieve the chunks that best match the current goal situation and then on blending to estimate the utility of layer-selection moves based on past utilities. If the retrieval request fails, then the model computes possible layer-selection moves (i.e., it performs a look-ahead search) using a difference-reduction problem-solving heuristic. In difference reduction, for each mentally simulated layer-selection action, the model simulates and evaluates the utility of the outcome (with some likelihood of being inaccurate). Then the model stores a situation-action-outcome-utility chunk for each mentally simulated move. It is assumed that the number of moves mentally searched will not often be exhaustive. This approach is similar to the use of counterfactuals in the resource allocation model.

The ACT-R model of Task 6 relies on the use of declarative chunks that represent past Tasks 4, 5, and 6 experiences. This is intended to capture a learning process whereby participants have attended to a current probability distribution, chosen a layer, revised their estimates of the hypotheses, and finally assessed the utility of the layer selection they just made. The model assumes that chunks are formed from these experiences each representing the specific situation (probability distribution over groups), selected intelligent layer, and observed intelligence outcome and information utility, where the utilities are computed by the weighted distance metric (d) below:

$$d = \sum_{i \in \text{Hypotheses}} p_i \left(1 - p_i\right), \tag{4}$$

where, each p is a posterior probability of a group attack based on rational calculation, and zero is the optimum. The ACT-R model uses this weighted distance function and assumes that the participant's goal is to achieve certainty on one of the hypotheses (i.e., $p_i = 1$).

At a future layer selection point, a production rule will request a blended/partial matching retrieval from declarative memory based on the current situation (probability distribution over possible attacking groups). ACT-R will use a blended retrieval mechanism to partially match against previous experience chunks and then blend across the stored information utilities for each of the available intelligence layer choices. For each layer, this blending over past experience of the information utilities will produce a kind of expected information utility for each type of intelligence for specific situations. Finally, the model compares the expected utility of different intelligence layers and selects the one with the highest utility.

The ACT-R model performs reinforcement learning throughout Tasks 4 to 6. After updating probability distribution based on a layer and its outcomes, the model evaluates whether it has gained or lost information by comparing the entropy of the prior distribution with the posterior distribution. If it has gained information, the production for the current layer receives a reward if it has lost, it receives a punishment. This reinforcement learning enables the model to acquire a preference order for the selection of intelligence layers, and this preference order list was used to determine which layer should be explored first in the beginning of the layer selection process.

3.4. *Cognitive Biases Addressed.*

Anchoring and confirmation biases have been long studied in cognitive psychology and the intelligence communities [9, 51–55]. As we have already mentioned, these biases emerge in several ways in the ACT-R model of AHA tasks (see Table 4 for an overview). In general, our approach accounts for three general sources of biases.

The first source of bias is the architecture itself, both in terms of mechanisms and limitations. In our model, a primary mechanistic source of bias is the ACT-R blending mechanism that is used to make decisions by drawing on instances of similar past problems. While it provides a powerful way of aggregating knowledge in an efficient manner, its consensus-driven logic tends to yield a regression to the mean that often (but not always) results in an anchoring bias. Limitations of the architecture such as working memory capacity and attentional bottlenecks can lead to ignoring some information that can result in biases such as base-rate

Table 4: Source of cognitive biases in the ACT-R integrated model of the AHA tasks.

Cognitive bias	Mechanism	Source of bias in functional model (ACT-R)
Confirmation bias	Attentional effect (seeking)	Feature selection behavior such as selecting SIGINT too early. Blended retrieval during layer choice using stored utilities.
	Overscaling in rule application (weighing)	Bias in blended retrieval of mappings from likelihood factor to revised probability (low value). Weighted-distance utilities used for layer selections shows confirmation bias in weighing.
Anchoring in learning	Underscaling in rule application	Bias in blended retrieval of mappings from likelihood factor to revised probability (high values)
	Centroid computation	Inertia from centroid estimates to consolidated values to DM. Productions encoding thresholds in distance for centroid updating
Representativeness	Base-rate neglect	Base rate not a cue for matching to a category. Compute distance to category centroid rather than cloud of events. Blended retrievals ignore number of events
Probability matching	Resource allocation	Use of instance-based learning leads to tendency of risk aversion against winner-take-all instances, leading to the tendency for the blended retrieval of instances between pure probability matching and winner-take-all

neglect (ignoring background frequencies when making a judgment of conditional probabilities).

The second source of bias is the content residing in the architecture, most prominent strategies in the form of procedural knowledge. Strategies can often lead to biases when they take the form of heuristic that attempt to conserve a limited resource, such as only updating a subset of the probabilities in order to save time and effort, or overcome a built-in architectural bias, such as the centroid-of-centroid strategy intended to combat the recency effect in chunk activations that in turn leads to an anchoring bias.

The third and final source of biases is the environment itself, more specifically its interaction with the decision-maker. For instance, the normalization functionality in the experimental interface can lead to anchoring bias if it results in a double normalization. Also, the feedback provided by the environment, or lack thereof, can lead to the emergence or persistence of biases. For instance, the conservatism bias that is often seen in the generation of probabilities could persist because subjects do not receive direct feedback as to the accuracy of their estimates.

In the next subsections we discuss in detail the sources of the various biases observed.

3.4.1. Anchoring and Adjustment. Anchoring is a cognitive bias that occurs when individuals establish some beliefs based on some initial evidence and then overly rely on this initial decision in their weighting of new evidence [54]. Human beings tend to anchor on some estimates or hypotheses, and subsequent estimates tend to be adjustments that are influenced by the initial anchor point—they tend to behave as if they have an anchoring + adjustment heuristic. Adjustments tend to be insufficient in the sense that they overweight the initial estimates and underweight new evidence.

Anchoring and adjustment in learning (AL) can occur in the first three tasks due to the nature of the task, specifically the iterative generation of the centroids of each category across each trial. For each trial, participants estimate a centroid for the events, perceived to date by that category, then observe a new set of events and issue a revised estimate. This process of issuing an initial judgment and then revising might lead to anchoring and adjustment processes. Thus, in Tasks 1–3, anchoring can occur due to the centroid-of-centroid strategy to prevent being overly sensitive to the most recent events.

Tasks 4–6 can also elicit anchoring biases. Anchoring bias in weighing evidence might be found when participants revise their belief probabilities after selecting and interpreting a particular feature. The estimates of belief probabilities that were set prior to the new feature evidence could act as an anchor, and the revised (posterior) belief probabilities could be insufficiently adjusted to reflect the new feature (i.e., when compared to some normative standards). Insufficient adjustment may happen because the blended retrieval mechanism tends to have a bias towards the mean.

The model also updates only the probabilities corresponding to the positive outcomes of the decision rules. For example, if it is discovered that the probe occurs on a major road, the model would update the probabilities for categories A and B and neglect to adjust downward the probabilities for categories C and D. This neglect is assumed to result from a desire to save labor by relying on the interface normalization function and by the difficulty of carrying out the normalization computations mentally. In turn, this is a cause of an underestimation of probabilities (anchoring) that results from the normalization of the probabilities in the interface.

3.4.2. Confirmation Bias. Confirmation bias is typically defined as the interpretation of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand [53], the tendency for people to seek information and cues that confirm the tentatively held hypothesis or belief and not seek (or discount) those that support an opposite conclusion or belief [56], or the seeking of information considered supportive of favored beliefs [53]. Studies [57–59] have found evidence of confirmation bias in tasks involving intelligence analysis, and there is a common assumption that many intelligence failures are the result of confirmation bias in particular [9, 60].

Confirmation bias in weighing evidence might occur in the probability adjustment process in Tasks 4–6. For example, we have seen that the probability adjustment process applied to small probability values sometimes resulted in overadjustment. Certain strategies for probability adjustment might also result in confirmation bias. When applying a particular feature, such as IMINT (which supports two hypotheses), participants may only apply the adjustment to the preferred hypothesis while neglecting the other category that is also supported by evidence or weight the evidence too strongly in favor of the preferred category.

Confirmation bias can also occur in the evidence seeking process in Task 6 as the participants might select intelligence layers that maximize information gains about the current preferred category. For instance, when applying the IMINT and MOVINT rules, one could only apply the adjustment to the preferred hypothesis (assuming it is one of the two receiving favorable evidence from that layer) while neglecting the other categories also supported by the evidence. This strategic decision could reflect the desire both to minimize effort and to maximize information gain.

A significant difference in layer selection features is the SIGINT feature, which requires the selection of one particular category to investigate. If that feature is applied to the leading category and chatter is detected, then the category likelihood gains considerable weight (by a factor of 7). However, if no chatter is detected, then the category likelihood is strongly downgraded, which throws considerable uncertainty over the decision process. Thus the decision to select the SIGINT layer too early (before a strong candidate has been isolated) or to apply it to strictly confirm the leading category rather than either confirm or invalidate a close second might be construed as a form of confirmation bias in layer selection.

3.4.3. Base-Rate Neglect. Base-rate neglect is an error that occurs when the conditional probability of a hypothesis is assessed without taking into account the prior background frequency of the hypothesis' evidence. Base-rate neglect can come about from three sources.

(1) Higher task difficulties and more complex environments can lead to base-rate neglect due to the sheer volume of stimuli to remember. To reduce memory load, some features may be abstracted.

(2) Related to the above, there can be base-rate neglect due to architectural constraints. For instance, short-term memory is generally seen to have a capacity of 7± 2 chunks of information available. Once more chunks of information need to be recalled, some information may either be abstracted or discarded.

(3) Finally, there can be explicit knowledge-level strategic choices made from an analysis of (1) and (2) above.

The strategic choice (3) of the ACT-R model leads to base-rate neglect in calculating probabilities for Tasks 1–3. In particular, the fact that the ACT-R model generates probabilities based on category centroids leads to base-rate neglect. This is because base-rate information is not directly encoded within the category centroid chunk. The information is still available within the individual SIGACTs stored in the model, but it is not used directly in generating probabilities.

3.4.4. Probability Matching. We endeavored to develop a model that leveraged subsymbolic mechanisms that often give rise naturally to probability matching phenomena [61]. Subsymbolic mechanisms in ACT-R combine statistical measures of quality (chunk activation for memory retrieval, production utility for procedural selection) with a stochastic selection process, resulting in behavior that tends to select a given option proportionately to its quality rather than in a winner-take-all fashion. This approach is similar to stochastic neural models such as the Boltzmann Machine [35].

In our model, resource allocation decisions are based not on discrete strategies but rather on the accumulation of individual decision instances. Strategies then are an emergent property of access to those knowledge bases. Moreover, to unify our explanation across biases, we looked to leverage the same model that was used to account for anchoring (and sometimes confirmation) bias in probability adjustment.

Results also approximate those seen by human participants: a wide variation between full probability matching and winner-take-all, several individual runs tending towards uniform or random distributions, and the mean falling somewhere between probability matching and winner-take-all (closer to matching).

Probability matching in resource allocation occurs due to the trade-off inherent in maximizing reward versus minimizing risk. A winner-take-all is the optimal strategy overall; however there are individual trials with large penalties (a zero score) when a category other than the one with the highest probability is the ground truth. When such an outcome occurs prominently (e.g., in the first trial), it can have a determinant effect on subsequent choices [47].

4. Data and Results

The results of the ACT-R model on the AHA tasks were compared against 45 participants who were employees of the MITRE Corporation. All participants completed informed consent and debriefing questionnaires that satisfied IRB requirements. To compare the ACT-R model to both human participants and a fully Bayesian rational model, several metrics were devised by MITRE [62–64], which are summarized below.

As an overall measure of uncertainty across a set of hypotheses, we employed a *Negentropy* (normalized negative entropy) metric, N, computed as

$$N = \frac{(E_{\max} - E)}{E_{\max}}, \tag{5}$$

where E is the Shannon entropy computed as

$$E = -\sum_{h} P_h * \log_2 P_h, \tag{6}$$

where the summation is over the probabilities, P_h, assigned to hypotheses. Negentropy can be expressed on a scale of 0% to 100%, where $N = 0\%$ implies maximum uncertainty (i.e., maximum entropy or a uniform probability distribution over hypotheses) and $N = 100\%$ implies complete certainty (i.e., zero entropy or a distribution in which one hypothesis is assigned a maximum probability of 1). The normalization is provided by $E_{\max} = 2$ in the case of four hypotheses (Tasks 2–6) and $E_{\max} = 1$ in the case of two hypotheses (Task 1).

Comparisons of human and normative (e.g., Bayesian) assessments of certainty as measured by Negentropy permit the evaluation of some cognitive biases. For instance, one can compare human Negentropy N_H to Negentropy for a rational norm N_Q following the revision of probabilities assigned to hypotheses after seeing new intelligence evidence. If $N_H > N_Q$, then the human is exhibiting a confirmation bias because of overweighing evidence that confirms the most likely hypothesis. On the other hand, if $N_H < N_Q$, then the human is exhibiting conservatism which might arise from an anchoring bias.

In addition to measuring biases, we also compared the probability estimation and resource allocation functions of the model against both human participants and a Bayesian rational model (i.e., an optimal model). The Kullback-Leibler Divergence (K) is a standard information-theoretic measure for comparing two probability distributions like those of a human (P) and model (M). K_{PM} measures the amount of information (in bits) by which the two distributions differ, which is computed as follows:

$$\begin{aligned} K_{PM} &= E_{PM} - E_P \\ &= -\sum_h P_h * \log_2 M_h + \sum_h P_h * \log_2 P_h, \end{aligned} \tag{7}$$

where, similar to the Negentropy measure, E_{PM} is the cross-entropy of human participants (P) and the ACT-R model (M) and E_P is the entropy of human participants. It is important to note that $K_{PM} = 0$ when both distributions are the same, and K_{PM} increases as the two distributions diverge. K ranges from zero to infinity, but K is typically less than 1 unless the two distributions have large peaks in different hypotheses.

A normalized measure of similarity (S) on a 0–100% scale similar to that of Negentropy can be computed from K:

$$S = 100\% * 2^{-K}. \tag{8}$$

As the divergence K ranges from zero to infinity, the similarity S ranges from 100% to 0%. Thus S_{QP} and S_{QM} can be useful for comparing the success of humans or models in completing the task (compared by their success relative against a fully rational Bayesian model). This measure will be referred to as an S1 score.

To address the overall fitness of the model output compared with human data, the most direct measure would be a similarity comparing the human and model distributions (S_{PM}) directly. However, this would not be a good measure as it would be typically higher than 50% (K is typically less than 1); thus we scaled our scores on a relative basis by comparing against a null model. A null model (e.g., a uniform distribution, $R = \{0.25, 0.25, 0.25, 0.25\}$) exhibits maximum entropy, which implies "random" performance in sensemaking. Thus S_{PR} was used to scale as a lower bound in computing a relative success rate (RSR) measure as follows:

$$\text{RSR} = \frac{(S_{PM} - S_{PR})}{(100\% - S_{PR})}. \tag{9}$$

The model's RSR was zero if S_{PM} is equal to or less than S_{PR}, because in that case a null model R would provide the same or better prediction of the human data as the model. The RSR for a model M will increase as S_{PM} increases, up to a maximum RSR of 100% when $S_{PM} = 100\%$. For example, if a candidate model M matches the data P with a similarity score of $S_{PM} = 80\%$ and the null model R matches P with a similarity score of $S_{PR} = 60\%$, then the RSR for model M would be $(80 - 60)/(100 - 60) = (20/40) = 50\%$.

In Task 6, because each participant potentially receives different stimuli at each stage of the trial as they choose different INTs to receive, RSR was not appropriate. Instead, the model was assessed against a relative match rate (RMR), which is defined below.

After receiving the common HUMINT feature layer at the start of each trial in Task 6, human participants have a choice amongst four features (IMINT, MOVINT, SIGINT, or SOCINT). The next choice is among three remaining features, and the last choice is among two remaining features. Thus there are $4 * 3 * 2 = 24$ possible sequences of choices that might be made by a subject on a given trial of Task 6. For each trial, the percentage of subjects that chooses each of the 24 sequences was computed. The modal sequence (maximum percentage) was used to define a benchmark $(t, s_{\max})$ for each trial (t), where F is the percentage of a sequence and $s_{\max}$ refers to the sequence with maximum F for trial t. For each trial, the model predicted a sequence of choices s_{mod}, and the percentage value of $F(t, s_{\text{mod}})$ for this sequence was computed from the human data. In other words, $F(t, s_{\text{mod}})$ is the percentage of humans that chose the same sequence as the model chose, on a given trial t:

$$\text{RMR}(t) = \frac{F(t, s_{\text{mod}})}{F(t, s_{\max})}. \tag{10}$$

For example, assume a model predicts a sequence of choices s_{mod} on a trial of Task 6. Assume also that 20% of human subjects chose the same sequence, but a different sequence was the most commonly chosen by human subjects, for example, by 40% of subjects. In that case $F(t, s_{\text{mod}}) = 20\%$ and $F(t, s_{\max}) = 40\%$, so $\text{RMR}(t) = 20\%/40\% = 50\%$.

Finally, a measure of resource allocation was derived by assigning a value (S2) based on the resources allocated to the category that was the ground truth. Thus if a participant (or model) was assigned a resource allocation of $\{A\% = 40, B\% = 30, C\% = 20, D\% = 10\}$ and the ground truth was category B, then the S2 score for that trial would be 30%. Thus, to maximize the score, an optimal model or participant would need to assign 100% of their resources to the ground truth (i.e., adopt a winner-take-all strategy to resource allocation).

4.1. Data. The integrated ACT-R AHA model performs (and learns incrementally across) all 6 tasks using the same

Table 5: S1, S2, RSR (RMR for Task 6), and linear regression (r^2) scores broken down by task and layer.

Task		S1 score			S2 score		RSR/RMR
	Model	Human	R^2	Model	Human	r^2	
1	78.1	68.7	.929*	55.0	69.1	.219	.650
2	68.2	53.7	.313*	78.7	79.1	.990*	.799
3	82.6	74.5	.001	45.2	45.3	.253*	.595
4-1	92.2	75.6	.730*				.761
4-2	92.7	76.2	.461*	47.7	44.0	.510*	.906
5-1	96.6	68.1	.037				.856
5-2	91.5	77.4	.078				.776
5-3	85.3	69.8	.115				.780
5-4	82.3	66.3	.262*	40.4	45.2	.637*	.618
6	91.2	91.0	.867*	34.8	31.2	.902*	.788

*$P < .01$.

knowledge constructs (production rules and chunks; other than those it learns as part of executing the task) and parameters. The results comparing human and model performance presented below are broken down by task and expanded on trial-by-trial and layer-by-layer analyses. For instance, while a similar instance-based representation is used across all tasks for probability adjustment, resource allocation, and layer selection, the output from the path planning mechanism is only used in Tasks 3 and 4. Thus it is best to examine Task 3 and Task 4-1 (the first layer of Task 4) alone in determining the efficacy of the path planning mechanism.

The model was run the same number of times as participants in the dataset (45) with the average model response were compared to the average human performance. The natural variability in the model (stochastic elements influencing instance-based learning) approximates some of the individual differences of the human participants. While the average distribution of the ACT-R model is slightly peakier than the average human (the ACT-R model is closer to Bayesian rational than humans are), the overall fits (based on RSR/RMR) are quite high, with an overall score over .7 (a score of 1 indicates a perfect fit [63, 64]; see Table 5). In addition, a linear regression comparing model and human performance at each block of each layer indicates that the model strongly and significantly predicts human behavior on AHA tasks.

Supporting these results, the trial-by-trial performance of the model (see Figure 11) predicted many of the variations seen in users' data. While the ACT-R model tended to behave more rationally than human participants (i.e., the model exhibited a higher S1 score), the model tended to capture much of the individual variation of human participants across trials (the S1 scores on Task 2 and S2 scores on Task 3 being the exceptions).

In addition to the fit to human data based on probabilities (S1/RSR) and resource allocation (S2), the model was also compared to human participants in terms of the anchoring and adjustment and confirmation biases (see Figure 12). Whenever both the human behavioral data and model exhibit a lower Negentropy than the Bayesian rational model, they are both exhibiting anchoring bias (and conversely they exhibit confirmation bias when they have a higher Negentropy). As shown below, the ACT-R model significantly predicts not only the presence or absence of a bias but also the quantity of the bias metric, reflected in an overall $R^2 = .645$ for Negentropy scores across all tasks.

4.1.1. *Tasks 1 and 2.*

In Task 1, the ACT-R model produces a probability distribution and forced-choice resource allocation for each trial. The probability distribution is based on the blended probability adjustments using instance-based learning as described above and results in an increased prevalence of anchoring (i.e., less peaky distributions) over the normative solution in a manner similar to (yet stronger than) human data.

Similar to Task 1, in Task 2 the model follows the general trends of human participants for both S1 and especially S2 scores. With 4 groups to maintain in Task 2, we assume that there is more base-rate neglect in humans (which results in ACT-R from the centroid representation of groups that loses base-rate information), which increases the RSR score to .799. However, the R^2 for S1 drops from .929 in Task 1 to .313 in Task 2 because the ACT-R model does not capture the same trial-by-trial variability despite being closer to mean human performance.

In Task 1, the ACT-R model exhibited a mean Negentropy score ($N_M = .076$), well below that of the Bayesian solution ($N_Q = .511$); thus, there was an overall strong trend towards anchoring and adjustment in learning (AL) for the model. Humans exhibited a similar AL bias ($N_H = .113$). Additionally, on a trial-by-trial comparison of the model to the Bayesian solution, both humans and the ACT-R model showed AL for each individual trial.

In Task 2 the results were similar ($N_Q = .791$, $N_M = .206$, $N_H = .113$) with both the model and humans exhibiting anchoring and adjustment in learning in every trial.

4.1.2. *Task 3.*

In Task 3 the ACT-R model was first required to generate category centroids based on a series of events and then was required to use the path planning mechanism

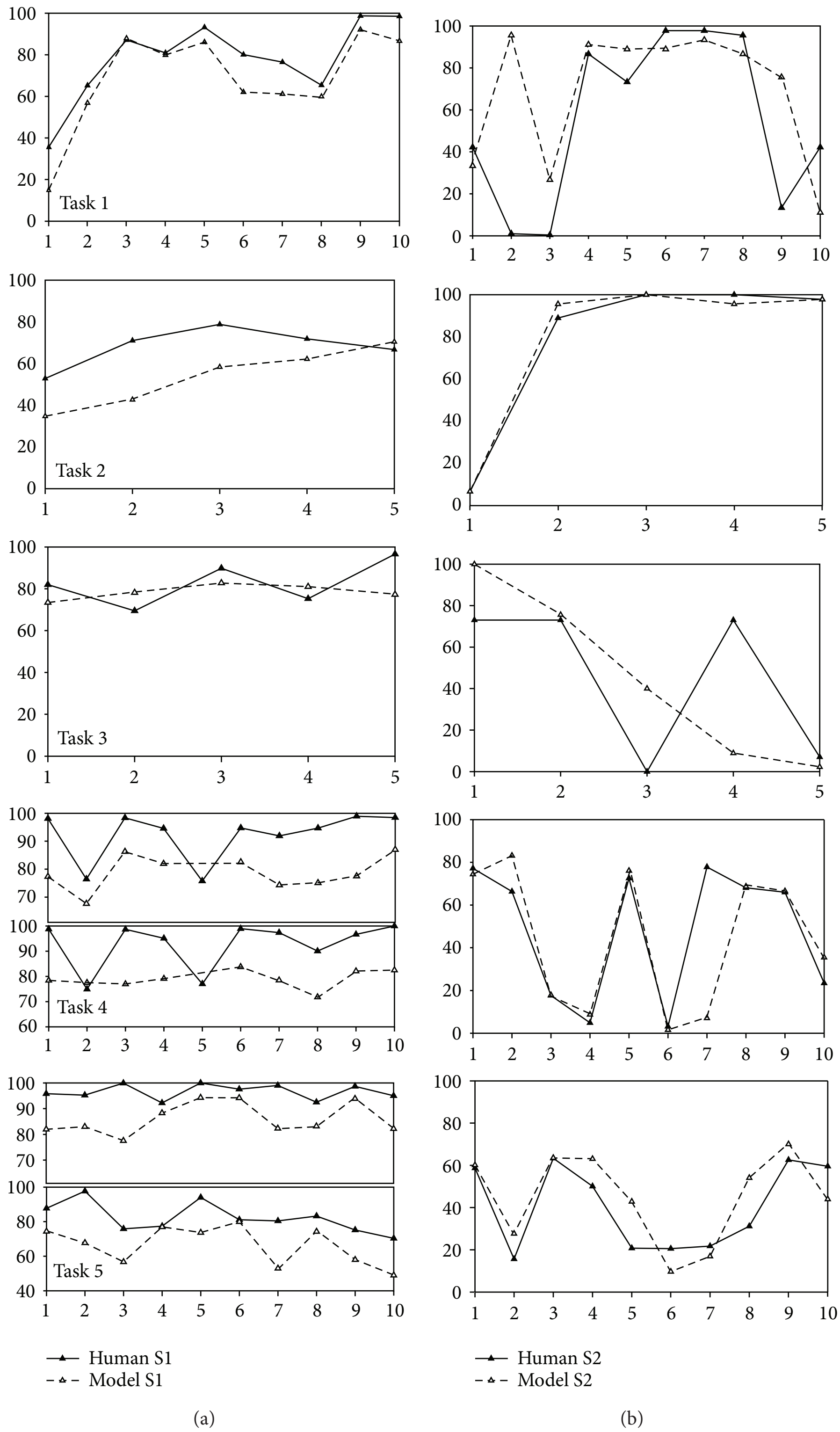
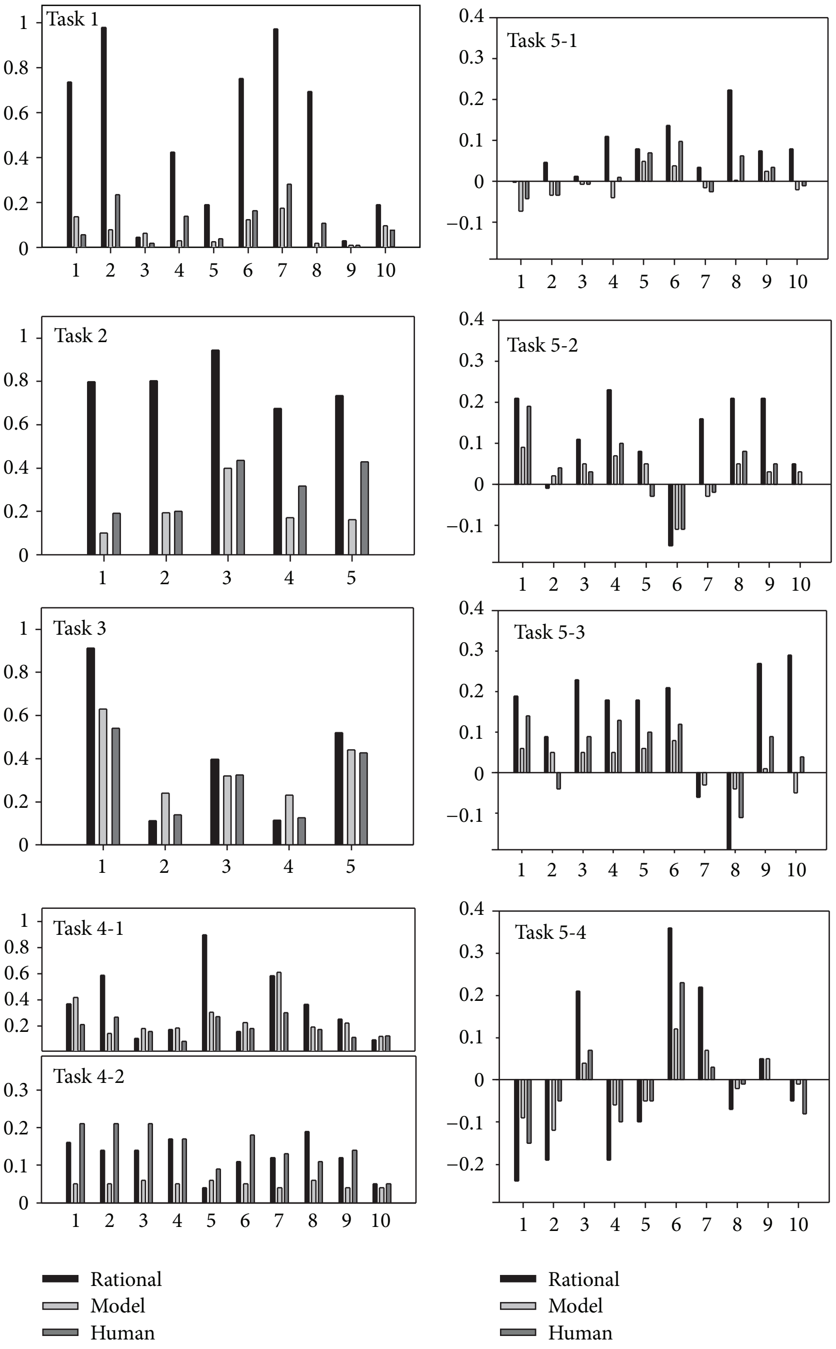


Figure 11: (a) is the trial-by-trial (horizontal axis) fit between the ACT-R model and human data for Tasks 1–5 using the S1 metric (vertical axis), which compares humans and model to Bayesian rational. (b) is the fit for the S2 metric determining resource allocation score. For Tasks 4-5, the top tile is the fit for the first feature layer, and the bottom tile is the fit for the final feature layer.



(a) (b)

Figure 12: Trial-by-trial Negentropy scores for Tasks 1–5 (Δ Negentropy between layers for Tasks 4-2 and 5) for the fully rational Bayes outcome, the ACT-R model, and human participants. Values less than normative (i.e., Bayesian rational) are considered an anchoring bias, and values greater than normative are considered confirmation bias.

to estimate the distance between each category centroid and a probe location. While the model captured average human performance on the task, it was not able to capture individual human behavior. This was in part due to wide variability and negative skew in the raw human data and a difficulty in the ACT-R model correctly placing category centroids when events fell across multiple roads.

However, when examining bias metrics, the ACT-R model exhibited both AL and confirmation biases as did human participants. Both ACT-R and human participants exhibited an AL bias on Trials 1, 3, and 5 and confirmation bias on Trials 2 and 4. Overall, both the model and humans exhibit a similar AL (N_Q = .412, N_M = .372, and N_H = .311). Thus, while the model was not capturing the exact distance estimates of human participants, it was able to capture the variability in the bias metrics.

4.1.3. Task 4. In Task 4, the issue with centroid generation over multiple roads is avoided since centroids are provided by the task environment, resulting in a HUMINT layer RSR = .761 and R^2 = .730. Thus, the path-planning mechanism itself is functioning correctly and providing excellent fits to human data. In addition, the model provided excellent fits to the second layer (the SOCINT layer) in Task 4, with an RSR fit of .905.

Beginning with Task 4, layer 2, the measure of anchoring and adjustment (Delta Negentropy) is based on whether category probabilities were revised sufficiently by following the probabilistic decision rules. There was an overall trend towards anchoring and adjustment in both learning and inference, with a slight trend towards confirmation bias for the humans. The main difference is when using SOCINT; the ACT-R model tends to exhibit an anchoring bias while human participants tended to exhibit a confirmation bias when applying the SOCINT layer. We speculate that the reason why humans would exhibit a confirmation bias on SOCINT, which is the easiest of the rules to apply, might be that it has a compelling visual interpretation that participants are more likely to trust.

Also, beginning with Task 4, resource allocation judgments are a distribution instead of a forced-choice. The model learns the outcomes of probability matching (PM) versus winner-take-all (WTA; forced-choice) through experience on Tasks 1–3 in the form of IBL chunks. From this experience, the model adopts a strategy (not a procedural rule but emergent from blended retrieval of chunks) that is somewhere between PM and WTA, with a bias towards PM. Based on the S2 fits for Tasks 4–6 (see Table 5), the resource allocation mechanism, which also relies on the same instance-based learning approach as the probability adjustment mechanism, provides an excellent match to human data.

4.1.4. Task 5. In Task 5 the RSR fits for Layers 1–3 are quite high (.856, .776, and .780, resp.) with some drop-off in Layer 4 (.618) due to human participants' distributions being closer to uniform and an RSR singularity (a near-uniform Bayesian, human, and model distribution leading to all nonperfect fits receiving a near-zero score since the random model near-perfect predicts human behavior). It may also be the case that humans, after getting several pieces of confirmatory and disconfirmatory evidence, express their uncertainty by flattening out their distribution in the final layer rather than applying each layer mechanically.

As seen in the Delta Negentropy graphs for each layer (see Figure 12), ACT-R correctly predicts the overall trend of anchoring ($N_H < N_Q$ and $N_M < N_Q$) for each layer:

Layer 1: $N_q = .080$, $N_h = .016$, $N_m = -.007$

Layer 2: $N_q = .110$, $N_h = .033$, $N_m = .025$

Layer 3: $N_q = .138$, $N_h = .056$, $N_m = .024$

Layer 4: $N_q = .000$, $N_h = -.007$, $N_m = -.011$

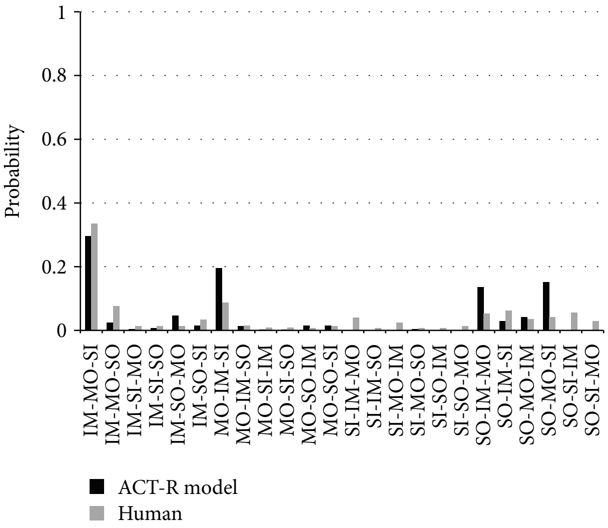


Figure 13: Layer selection sequences both the ACT-R model and human data (IM for IMINT, MO for MOVINT, SI for SIGINT, and SO for SOCINT).

Across each layer, the model correctly predicts anchoring on all 10 trials of Layer 2, correctly predicts anchoring on 8 trials of Layer 3 and correctly predicts the confirmation on the other 2 trials, correctly predicts anchoring on 8 trials of Layer 4 and correctly predicts confirmation on the other 2, and correctly predicts anchoring on 4 trials of Layer 5 and correctly predicts confirmation on 5 other trials. Over 40 possible trials, ACT-R predicts human confirmation and anchoring biases on 39 of the trials (trial 10 of Layer 5 being the only exception).

4.1.5. Task 6. In Task 6, both the model and participants are able to choose 3 feature layers before specifying a final probability distribution. Figure 13 shows the probability distribution of layer selection sequences for our ACT-R model and human data. To measure the similarity of the probability distribution of layer selection sequences between the ACT-R model and human data, we performed Jensen-Shannon divergence analysis, which is a method of measuring the similarity between two distributions. The divergence between the two distributions is .35, indicating that the ACT-R model strongly predicts the human data patterns.

5. Generalization

To determine the generalizability of the integrated ACT-R model of the AHA tasks, the same model that produced the above results was run on novel stimuli in the same AHA framework. The results of the model were then compared to the results of a novel sample gathered from 103 students at Penn State University. This new data set was not available before the model was run, and no parameters or knowledge structures were changed to fit this data set. Unlike the original 45-participant dataset, the Penn State sample used only people who had taken course credit towards a graduate Geospatial Intelligence Certificate. Overall, the RSR and R^2

Table 6: Set of S1, S2, RSR (RMR in Task 6), and linear regression (r^2) scores broken down by task for novel dataset and participants.

Task	S1 score			S2 score			RSR/RMR
	Model	Human	R^2	Model	Human	r^2	
1	81.7	80.7	.011	59.1	63.4	.141	.625
2	68.5	78.9	.347*	54.2	54.6	.765*	.534
3	72.1	79.7	.121	34.7	73.8	.701*	.692
4	94.4	87.6	.006	47.5	46.7	.992*	.893
5	84.5	84.5	.000	42.0	45.5	.943*	.864
6	85.3	88.3	.447*	48.4	44.6	.990*	.854

*$P < .01$.

fits on S2 scores improved while the R^2 fits on S1 scores dropped (see Table 6). The increase in RSR was mainly due to the Penn State population behaving more rationally (i.e., higher S1 scores; see Figure 14) than the population from the initial dataset. This is consistent with the increased education and experience of the Penn State sample. That said, the Penn State sample most likely utilized some different strategies in solving the tasks, as the trial-by-trial S1 fits were not as close, implying some difference in reasoning that the ACT-R model was not capturing.

Overall, the improved model fits indicate that the ACT-R model of the AHA tasks is able to capture average human performance at the task level for S1 scores and at the trial-by-trial level for S2 scores. Conversely, this justifies the reliability of the AHA tasks as a measure of human performance in a sensemaking task environment.

Finally, the ACT-R model fits for anchoring and confirmation biases (see Figure 15) were also similar in the Penn State dataset. The model correctly predicted both the presence and degree of anchoring on every block in Tasks 1–3 and followed similar trial-by-trial trends for both anchoring and confirmation in Tasks 4-5. R^2 of Negentropy scores was a similar .591 to the original dataset.

6. Conclusion

The decision-making literature has established a lengthy list of cognitive biases under which human decision making empirically deviates from the theoretical optimum. Those biases have been studied in a number of theoretical (e.g., binary choice paradigms) and applied (e.g., medical diagnosis and intelligence analysis) settings. However, as the list of biases and experimental results grows, our understanding of the mechanisms producing these biases has not followed pace. Biases have been formulated in an ad hoc, task- and domain-specific manner. Explanations have been proposed ranging from the use of heuristic to innate individual preferences. What is lacking is an explicit, unified, mechanistic, and theoretical framework for cognitive biases that provides a computational understanding of the conditions under which they arise and of the methods by which they can be overcome.

In this paper, we present such a framework by developing unified models of cognitive biases in a computational cognitive architecture. Our approach unifies results along a pair of orthogonal dimensions. First, the cognitive architecture provides a functional computational bridge from qualitative theories of sensemaking to detailed neural models of brain functions. Second, the framework enables the integration of results across paradigms from basic decision making to applied fields. Our basic hypothesis is that biases arise from the interaction of three components: the task environment, including the information and feedback available as well as constraints on task execution, the cognitive architecture, including cognitive mechanisms and their limitations, and the use of strategies including heuristic as well as formal remediation techniques. This approach unifies explanations grounded in neurocognitive mechanisms with those assuming a primary role for heuristic. The goal is to derive a unified understanding of the conditions under which cognitive biases appear as well as those under which they can be overcome or unlearned. To achieve this unification, our model uses a small set of architectural mechanisms, leverages them using a coherent task modeling approach (i.e., instance-based learning), performs a series of tasks using the same knowledge structures and parameters, generalizes across different sets of scenarios and human participants, and quantitatively predicts a number of cognitive biases on a trial-to-trial basis.

In particular, we show biases to be prevalent under system 1 (automatic) processes [65] and that a unified theory of cognition can provide a principled way to understand how these biases arise from basic cognitive and neural substrates. As system 2 (deliberative) processes make use of the same cognitive architecture mechanisms in implementing access to knowledge and use of strategies, we expect biases to also occur, in particular as they relate to the interaction between the information flow and the structure of the architecture. However, at the same time, we show that system 2 processes can provide explicit means to remediate most effects of the biases, such as in the centroid-of-centroid generation strategy, where a strong recency bias is replaced with an (slight) anchoring bias.

Moreover, it is to be expected that a rational agent learns and adapts its strategies and knowledge, its metacognitive control (e.g., more deliberate processing of information), and its use of the task environment (e.g., using tools to perform computations or serve as memory aids) so as to at least reduce the deteriorating effects of these biases. However, biases are always subjective, in that they refer to an implicit assumption about the true nature of the world. For instance, the emergence of probability matching in the later tasks can

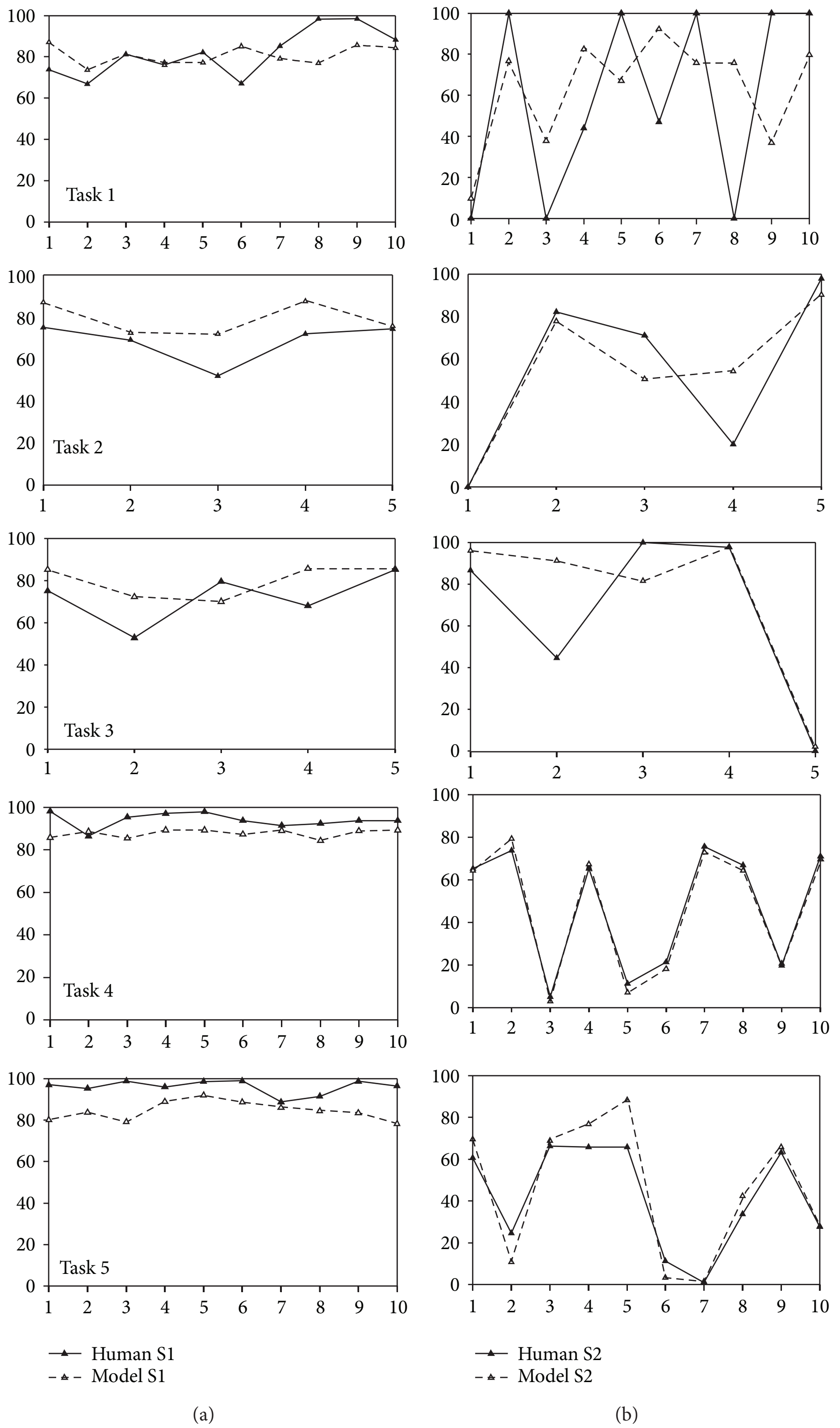
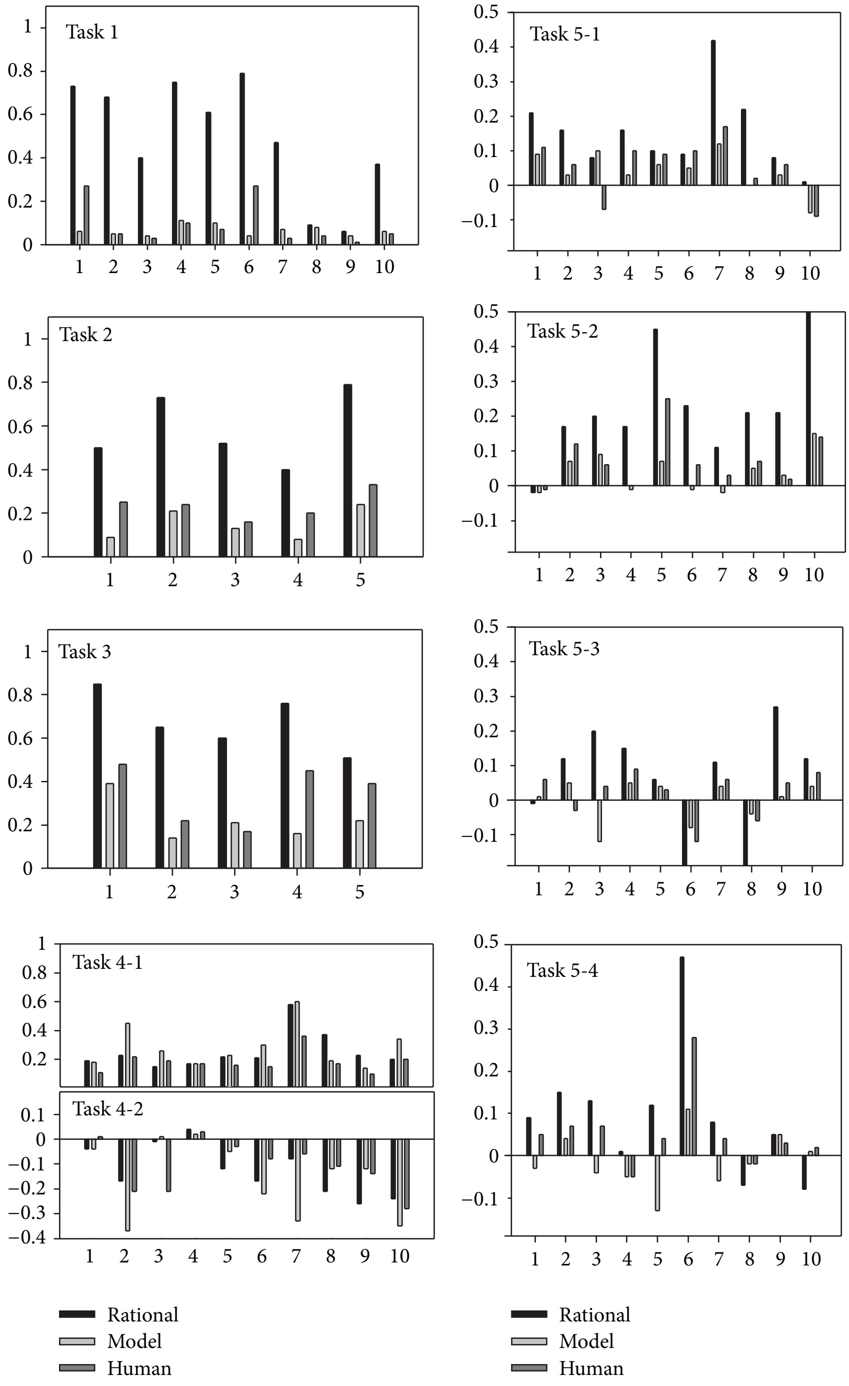


Figure 14: (a) is the trial-by-trial fit between the ACT-R model and human data for Tasks 1–5 using the S1 metric, which compares humans and model to Bayesian rational. (b) is the fit for the S2 metric determining resource allocation score. For Tasks 4-5, the graph represents the final layer fit. These results are for the final Penn State dataset.



(a) (b)

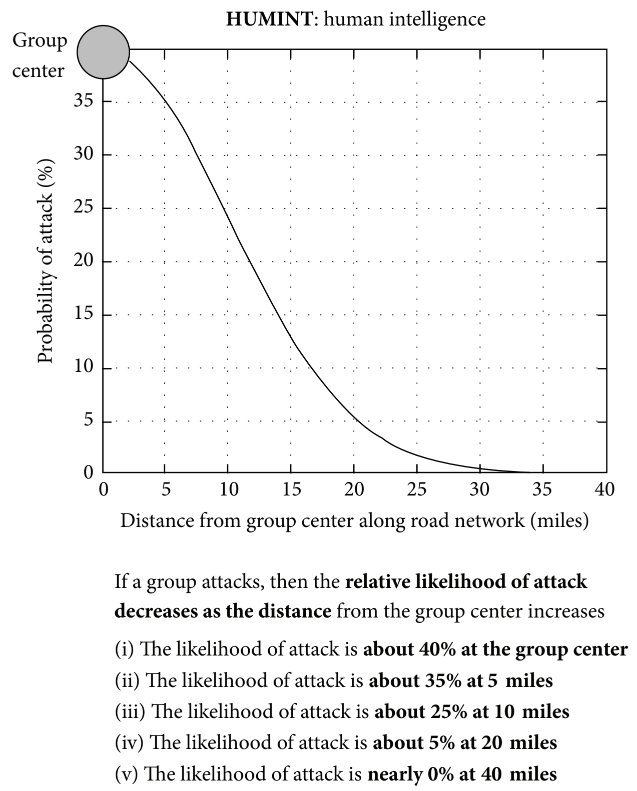
Figure 15: Trial-by-trial Negentropy scores for Tasks 1–5 (Δ Negentropy between layers for Tasks 4-2 and 5) for the fully rational Bayes outcome, the ACT-R model, and human participants. These results are for the Penn State dataset.

be seen as a response to the uncertainty of the earlier tasks and the seemingly arbitrary nature of the outcomes. Thus, people respond by hedging their bets, a strategy that can be seen as biased in a world of exact calculations but one that has shown its robust adaptivity in a range of real-world domains such as investing for retirement. As is often the case, bias is in the eye of the beholder.

Appendix

PROBs Handbook

Figures 16, 17, 18, 19, and 20 were the individual pages of the PROBs (Probabilistic Decision Rules) handbook explaining how to update category likelihoods based on the information revealed by each feature.


HUMINT: human intelligence

If a group attacks, then the **relative likelihood of attack decreases as the distance** from the group center increases

(i) The likelihood of attack is **about 40% at the group center**

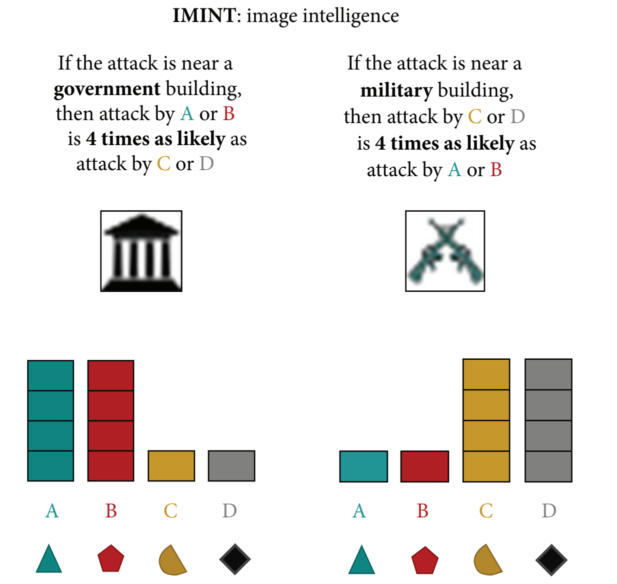
(ii) The likelihood of attack is **about 35% at 5 miles**

(iii) The likelihood of attack is **about 25% at 10 miles**

(iv) The likelihood of attack is **about 5% at 20 miles**

(v) The likelihood of attack is **nearly 0% at 40 miles**

Figure 16: The HUMINT feature, representing distance along a road network between a category and a probe event.


IMINT: image intelligence

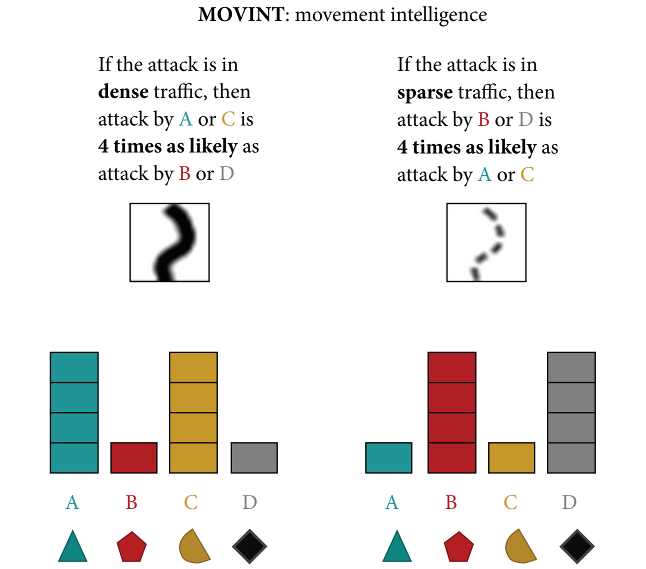
If the attack is near a **government** building, then attack by A or B is **4 times as likely** as attack by C or D

If the attack is near a **military** building, then attack by C or D is **4 times as likely** as attack by A or B

Figure 17: The IMINT feature, representing imagery of government or military buildings located at a probe event location.

Acknowledgments

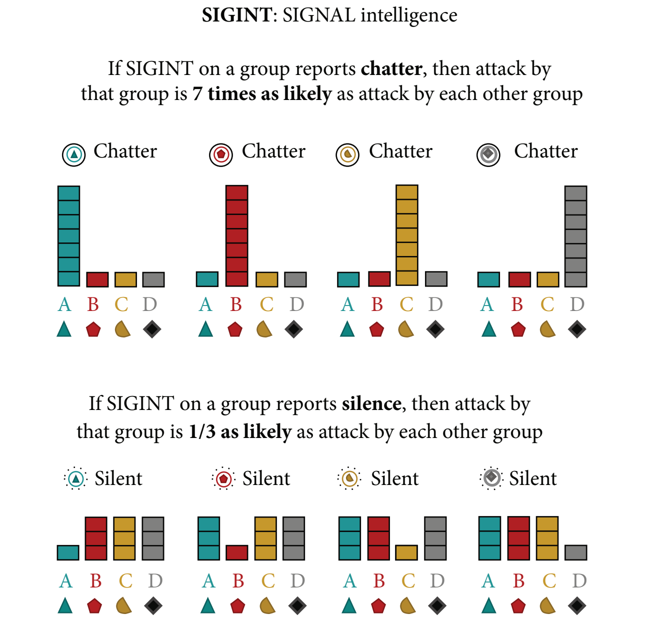
This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract no. D10PC20021. The U.S. Government is


MOVINT: movement intelligence

If the attack is in **dense** traffic, then attack by A or C is **4 times as likely** as attack by B or D

If the attack is in **sparse** traffic, then attack by B or D is **4 times as likely** as attack by A or C

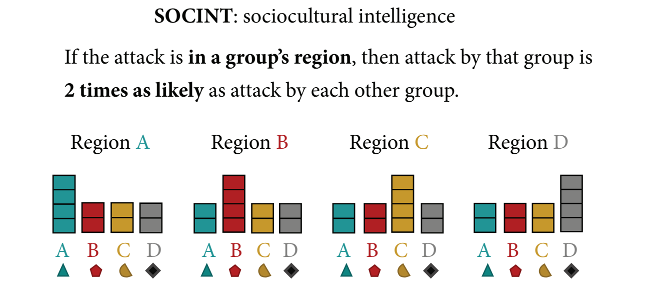
Figure 18: The MOVINT feature, representing vehicular movement information located at a probe event location.


SIGINT: SIGNAL intelligence

If SIGINT on a group reports **chatter**, then attack by that group is **7 times as likely** as attack by each other group

If SIGINT on a group reports **silence**, then attack by that group is **1/3 as likely** as attack by each other group

Figure 19: The SIGINT feature, representing chatter located at a probe event location. Note that SIGINT must be picked for a specific category.


SOCINT: sociocultural intelligence

If the attack is **in a group's region**, then attack by that group is **2 times as likely** as attack by each other group.

Figure 20: The SOCINT feature, representing sociocultural information about the region boundary for each category.

authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

[1] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, Oxford, UK, 2007.

[2] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–1060, 2004.

[3] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: alternative perspectives," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 70–73, 2006.

[4] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 2: a macrocognitive model," *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 88–92, 2006.

[5] P. Pirolli and S. K. Card, "The sensemaking process and leverage points for analyst technology," in *Proceedings of the International Conference on Intelligence Analysis*, McLean, Va, USA, May 2005.

[6] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in *Proceedings of the INTERACT and CHI Conference on Human Factors in Computing Systems*, pp. 269–276, April 1993.

[7] B. Dervin, "From the mind's eye of the user: the sense-making of qualitative-quantitative methodology," in *Sense-Making Methodology Reader: Selected Writings of Brenda Dervin*, B. Dervin, L. Foreman-Wenet, and E. Lauterbach, Eds., pp. 269–292, Hampton Press, Cresskill, NJ, USA, 2003.

[8] K. Weick, *Sensemaking in Oragnizations*, Sage, Thousand Oaks, Calif, USA, 1995.

[9] R. J. Heuer, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, Washington, DC, USA, 1999.

[10] S. G. Hutchins, P. Pirolli, and S. K. Card, "What makes intelligence analysis difficult? A cognitive task analysis of intelligence analysts," in *Expertise Out of Context*, R. R. Hoffman, Ed., pp. 281–316, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2007.

[11] L. G. Militello and R. J. B. Hutton, "Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands," *Ergonomics*, vol. 41, no. 11, pp. 1618–1641, 1998.

[12] J. M. Schraagen, S. F. Chipman, and V. L. Shalin, Eds., *Cognitive Task Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2000.

[13] M. I. Posner, R. Goldsmith, and K. E. Welton Jr., "Perceived distance and the classification of distorted patterns," *Journal of Experimental Psychology*, vol. 73, no. 1, pp. 28–38, 1967.

[14] C. Lefebvre, R. Dell'acqua, P. R. Roelfsema, and P. Jolicæur, "Surfing the attentional waves during visual curve tracing: evidence from the sustained posterior contralateral negativity," *Psychophysiology*, vol. 48, no. 11, pp. 1510–1516, 2011.

[15] F. G. Ashby and W. T. Maddox, "Complex decision rules in categorization: contrasting novice and experienced performance," *Journal of Experimental Psychology*, vol. 18, no. 1, pp. 50–71, 1992.

[16] A. Stocco, C. Lebiere, R. C. O'Reilly, and J. R. Anderson, "The role of the anterior prefrontal-basal ganglia circuit as a biological instruction interpreter," in *Biologically Inspired Cognitive Architectures 2010*, A. V. Samsonovich, K. R. Jóhannsdóttir, A. Chella, and B. Goertzel, Eds., vol. 221 of *Frontiers in Artificial Intelligence and Applications*, pp. 153–162, 2010.

[17] G. Ryle, *The Concept of Mind*, Hutchinson, London, UK, 1949.

[18] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[19] H. A. Simon, "How big is a chunk?" *Science*, vol. 183, no. 4124, pp. 482–488, 1974.

[20] C. Lebiere, "The dynamics of cognition: an ACT-R model of cognitive arithmetic," *Kognitionswissenschaft*, vol. 8, no. 1, pp. 5–19, 1999.

[21] N. Taatgen, C. Lebiere, and J. R. Anderson, "Modeling paradigms in ACT-R," in *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, R. Sun, Ed., Cambridge University Press., New York, NY, USA, 2006.

[22] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1983.

[23] A. Newell, *Unified Theories of Cognition*, Harvard University Press, Cambridge, Mass, USA, 1990.

[24] J. R. Anderson, "Acquisition of cognitive skill," *Psychological Review*, vol. 89, no. 4, pp. 369–406, 1982.

[25] J. R. Anderson, *Rules of the Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1993.

[26] C. Gonzalez, J. F. Lerch, and C. Lebiere, "Instance-based learning in dynamic decision making," *Cognitive Science*, vol. 27, no. 4, pp. 591–635, 2003.

[27] C. Lebiere, C. Gonzalez, and W. Warwick, "Metacognition and multiple strategies in a cognitive model of online control," *Journal of Artificial General Intelligence*, vol. 2, no. 2, pp. 20–37, 2010.

[28] R. C. O'Reilly, T. E. Hazy, and S. A. Herd, "The leabra cognitive architecture: how to play 20 principles with nature and win!," in *Oxford Handbook of Cognitive Science*, S. Chipman, Ed., Oxford University Press, Oxford, UK.

[29] M. Ziegler, M. Howard, A. Zaldivar et al., "Simulation of anchoring bias in a spatial estimation task due to cholinergic-neuromodulation," submitted.

[30] Y. Sun and H. Wang, "The parietal cortex in sensemaking: spatio-attentional aspects," in press.

[31] R. Thomson and C. Lebiere, "Constraining Bayesian inference with cognitive architectures: an updated associative learning mechanism in ACT-R," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci '13)*, Berlin, Germany, July-August 2013.

[32] C. Lebiere and J. R. Anderson, "A connectionist implementation of the ACT-R production system," in *Proceedings of the 15th Annual Meeting of the Cognitive Science Society (CogSci '93)*, pp. 635–640, Lawrence Erlbaum Associates, June 1993.

[33] D. J. Jilk, C. Lebiere, R. C. O'Reilly, and J. R. Anderson, "SAL: an explicitly pluralistic cognitive architecture," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 20, no. 3, pp. 197–218, 2008.

[34] D. Marr, "Simple memory: a theory for archicortex," *Philosophical Transactions of the Royal Society of London B*, vol. 262, no. 841, pp. 23–81, 1971.

[35] D. E. Rumelhart, J. L. McClelland, and PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, MIT Press, Cambridge, Mass, USA, 1986.

[36] C. Lebiere, J. R. Anderson, and L. M. Reder, "Error modeling in the ACT-R production system," in *Proceedings of the 16th Annual Meeting of the Cognitive Science Society*, pp. 555–559, Lawrence Erlbaum Associates, 1994.

[37] J. R. Anderson, *The Adaptive Character of Thought*, Lawrence Erlbaum Associates, 1990.

[38] A. Klippel, H. Tappe, and C. Habel, "Pictorial representations of routes: chunking route segments during comprehension," in *Spatial Cognition III*, C. Freksa, W. Brauer, C. Habel, and K. F. Wender, Eds., vol. 2685 of *Lecture Notes in Computer Science*, pp. 11–33, 2003.

[39] S. M. Kosslyn, *Image and Brain: The Resolution of the Imagery Debate*, MIT Press, Cambridge, Mass, USA, 1994.

[40] W. C. Gogel and J. A. da Silva, "A two-process theory of the response to size and distance," *Perception & Psychophysics*, vol. 41, no. 3, pp. 220–238, 1987.

[41] W. M. Wiest and B. Bell, "Stevens's exponent for psychophysical scaling of perceived, remembered, and inferred distance," *Psychological Bulletin*, vol. 98, no. 3, pp. 457–470, 1985.

[42] J. A. da Silva, "Scales for perceived egocentric distance in a large open field: comparison of three psychophysical methods," *The American Journal of Psychology*, vol. 98, no. 1, pp. 119–144, 1985.

[43] J. A. Aznar-Casanova, E. H. Matsushima, N. P. Ribeiro-Filho, and J. A. da Silva, "One-dimensional and multi-dimensional studies of the exocentric distance estimates in frontoparallel plane, virtual space, and outdoor open field," *The Spanish Journal of Psychology*, vol. 9, no. 2, pp. 273–284, 2006.

[44] C. A. Levin and R. N. Haber, "Visual angle as a determinant of perceived interobject distance," *Perception & Psychophysics*, vol. 54, no. 2, pp. 250–259, 1993.

[45] R. Thomson, *The role of object size on judgments of lateral separation [Ph.D. dissertation]*.

[46] S. Dehaene and L. Cohen, "Language and elementary arithmetic: dissociations between operations," *Brain and Language*, vol. 69, no. 3, pp. 492–494, 1999.

[47] C. . Lebiere, C. Gonzalez, and M. Martin, "Instance-based decision making model of repeated binary choice," in *Proceedings of the 8th International Conference on Cognitive Modeling (ICCM '07)*, Ann Arbor, Mich, USA, July 2007.

[48] I. Erev, E. Ert, A. E. Roth et al., "A choice prediction competition: choices from experience and from description," *Journal of Behavioral Decision Making*, vol. 23, no. 1, pp. 15–47, 2010.

[49] D. Wallach and C. Lebiere, "Conscious and unconscious knowledge: mapping to the symbolic and subsymbolic levels of a hybrid architecture," in *Attention and Implicit Learning*, L. Jimenez, Ed., John Benjamins Publishing, Amsterdam, Netherlands, 2003.

[50] C. Lebiere, C. Gonzalez, and W. Warwick, "A comparative approach to understanding general intelligence: predicting cognitive performance in an open-ended dynamic task," in *Proceedings of the 2nd Conference on Artificial General Intelligence (AGI '09)*, pp. 103–107, Arlington, Va, USA, March 2009.

[51] J. Klayman, "Varieties of confirmation bias," in *Decision Making from a Cognitive Perspective*, J. Busemeyer, R. Hastie, and D. L. Medin, Eds., vol. 32 of *Psychology of Learning and Motivation*, pp. 365–418, Academic Press, New York, NY, USA, 1995.

[52] J. Klayman and Y.-W. Ha, "Confirmation, disconfirmation, and information in hypothesis testing," *Psychological Review*, vol. 94, no. 2, pp. 211–228, 1987.

[53] R. S. Nickerson, "Confirmation bias: a ubiquitous phenomenon in many guises," *Review of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[54] A. Tversky and D. Kahneman, "Judgment under uncertainty: heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[55] P. Wason, "On the failure to eliminate hypotheses in a conceptual task," *The Quarterly Journal of Experimental Psychology*, vol. 12, no. 3, pp. 129–140, 1960.

[56] C. D. Wickens and J. G. Hollands, *Engineering Psychology and Human Performance*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2000.

[57] B. A. Cheikes, M. J. Brown, P. E. Lehner, and L. Adelman, "Confirmation bias in complex analyses," Tech. Rep., MITRE Center for Integrated Intelligence Systems, Bedford, Mass, USA, 2004.

[58] G. Convertino, D. Billman, P. Pirolli, J. P. Massar, and J. Shrager, "Collaborative intelligence analysis with CACHE: bias reduction and information coverage," Tech. Rep., Palo Alto Research Center, Palo Alto, Calif, USA, 2006.

[59] M. A. Tolcott, F. F. Marvin, and P. E. Lehner, "Expert decision-making in evolving situations," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 3, pp. 606–615, 1989.

[60] C. Grabo and J. Goldman, *Anticipating Surprise*, Rowman & Littlefield, 2004.

[61] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1998.

[62] K. Burns, "Mental models and normal errors," in *How Professionals Make Decision*, H. Montgomery, R. Lipshitz, and B. Brehmer, Eds., pp. 15–28, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2004.

[63] MITRE Technical Report, "IARPA's ICArUS program: phase 1 challenge problem design and test specification," in progress.

[64] MITRE Technical Report, "A computational basis for ICArUS challenge problem design," in progress.

[65] D. Kahneman and S. Frederick, "A model of heuristic judgment," in *The Cambridge Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrison, Eds., pp. 267–293, Cambridge University Press, New York, NY, USA, 2005.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 149329, 18 pages
http://dx.doi.org/10.1155/2013/149329

Research Article

Strategic Cognitive Sequencing: A Computational Cognitive Neuroscience Approach

Seth A. Herd, Kai A. Krueger, Trenton E. Kriete, Tsung-Ren Huang, Thomas E. Hazy, and Randall C. O'Reilly

Department of Psychology, University of Colorado Boulder, Boulder, CO 80309, USA

Correspondence should be addressed to Seth A. Herd; seth.herd@gmail.com

Received 22 December 2012; Revised 7 May 2013; Accepted 28 May 2013

Academic Editor: Giorgio Ascoli

We address strategic cognitive sequencing, the "outer loop" of human cognition: how the brain decides what cognitive process to apply at a given moment to solve complex, multistep cognitive tasks. We argue that this topic has been neglected relative to its importance for systematic reasons but that recent work on how individual brain systems accomplish their computations has set the stage for productively addressing how brain regions coordinate over time to accomplish our most impressive thinking. We present four preliminary neural network models. The first addresses how the prefrontal cortex (PFC) and basal ganglia (BG) cooperate to perform trial-and-error learning of short sequences; the next, how several areas of PFC learn to make predictions of likely reward, and how this contributes to the BG making decisions at the level of strategies. The third models address how PFC, BG, parietal cortex, and hippocampus can work together to memorize sequences of cognitive actions from instruction (or "self-instruction"). The last shows how a constraint satisfaction process can find useful plans. The PFC maintains current and goal states and associates from both of these to find a "bridging" state, an abstract plan. We discuss how these processes could work together to produce strategic cognitive sequencing and discuss future directions in this area.

1. Introduction

Weighing the merits of one scientific theory against another, deciding which plan of action to pursue, or considering whether a bill should become law all require many cognitive acts, in particular sequences [1, 2]. Humans use complex cognitive strategies to solve difficult problems, and understanding exactly how we do this is necessary to understand human intelligence. In these cases, different strategies composed of different sequences of cognitive acts are possible, and the choice of strategy is crucial in determining how we succeed and fail at particular cognitive challenges [3, 4]. Understanding strategic cognitive sequencing has important implications for reducing biases and thereby improving human decision making (e.g., [5, 6]). However, this aspect of cognition has been studied surprisingly little [7, 8] because it is complex. Tasks in which participants tend to use different strategies (and therefore sequences) necessarily produce data that is less clear and interpretable than that from a single process in a simple task [9]. Therefore, cognitive neuroscience tends to avoid such tasks, leaving the neural mechanisms of strategy selection and cognitive sequencing underexplored relative to the large potential practical impacts.

Here, we discuss our group's efforts to form integrative theories of the neural mechanisms involved in selecting and carrying out a series of cognitive operations that successfully solve a complex problem. We dub this process strategic cognitive sequencing (SCS). While every area of the brain is obviously involved in some of the individual steps in some particular cognitive sequences, there is ample evidence that the prefrontal cortex (PFC), basal ganglia (BG), and hippocampus and medial temporal lobe (HC and MTL) are particularly important for tasks involving SCS (e.g., [10–14]). However, exactly how these brain regions allow us to use multistep approaches to problem solving is unknown. The details of this process are clearly crucial to understanding that process well enough to help correct dysfunctions, to better train it, and to perhaps eventually reproduce it in artificial general intelligence (AGI).

We present four different neural network models, each of a computational function that we consider crucial for strategic cognitive sequencing. The first two models address how sequences are learned and selected: how the brain selects which of a small set of known strategic elements to use in a given situation. The first, "model-free learning," is a model of how dopamine-driven reinforcement learning in the PFC and BG can learn short cognitive sequences entirely through trial and error, with reward available only at the end of a successful sequence. The second, "PFC/BG decision making" (PBDM), shows how cortical predictions of reward and effort can drive decision making in the basal ganglia for different task strategies, allowing a system to quickly generalize learning from selecting strategies on old tasks to new tasks with related but different strategies. The last two models apply to selecting *what* plans or actions (from the large set of possibilities in long-term semantic memory) will be considered by the two "which" systems. The third model, "instructed learning," shows how episodic recall can work with the PFC and BG to memorize sequences from instructions, while the last "subgoal selection" model shows how semantic associative processes in posterior cortex can select representations of "bridging states" which also constitute broad plans connecting current and goal states, each of which can theoretically be further elaborated using the same process to produce elaborate plan sequences.

Because these models were developed somewhat separately, they and their descriptions address "actions," "strategies," "subgoals," and "plans." We see all of these as sharing the same types of representations and underlying brain mechanism, so each model actually addresses all of these levels. All of these theories can be applied either to individual actions or whole sequences of actions that have been previously learned as a "chunk" or plan. This hierarchical relationship between sequence is well understood at the lower levels of motor processing (roughly, supplementary motor areas tend to encode sequences of primary motor area representations, while presupplementary motor areas encode sequences of those sequences); we assume that this relationship holds to higher levels, so that sequences of cognitive actions can be triggered by a distributed representation that loosely encodes that whole sequence and those higher-level representations can then unfold as sequences themselves using identically structured brain machinery, possibly in slightly different, but parallel brain areas.

Before elaborating on each model, we clarify the theoretical framework and background that have shaped our thinking. After describing each model, we further tie each model to our overall theory of human strategic cognitive sequencing and describe our planned future directions for modeling work that will tie these individual cognitive functions into a full process that learns and selects sequences of cognitive actions constituting plans and strategies appropriate for novel, complex mental tasks, one of humans' most impressive cognitive abilities.

2. Theoretical Framework

These models synthesize available relevant data and constitute our attempt at curren best-guess theories. We take a computational cognitive neuroscience approach, in which artificial neural network models serve to concretize and specify our theories. The models serve as cognitive aids in a similar way to diagramming and writing about theories but also serve to focus our inquiries on the computational aspects of the problem. These theories are constrained not only by the data we specifically consider here but also by our use of the Leabra modeling framework [15, 16]. That framework serves as a cumulative modeling effort that has been applied to many topic areas and serves to summarize a great deal of data on neural function. This framework serves as a best-guess theory on cortical function, and individual models represent more specific, but still empirically well-supported and constrained theories of PFC, basal ganglia, reward system, and hippocampal function. Here, we extend these well-developed theories to begin to address SCS.

We also take our constraints from purely cognitive theories of cognitive sequencing. Work on production system architectures serves as elaborate theories of how human beings sequence cognitive steps to solve complex problems [17–19]. The numerous steps by which a production system model carries out a complex task such as air traffic control [20] are an excellent example of cognitive sequencing. Our goal here is to elaborate on the specific neural mechanisms involved, and in so doing, we alter those theories somewhat while still accounting for the behavioral data that has guided their creation.

Neural networks constitute the other class of highly specified and cumulative theories of cognition. However, these are rarely applied to the type of tasks we address here, in which information must be aggregated from step to step, but in arbitrary ways (e.g., first figure out center of a set of points, then calculate the distance from that center of points to an another point, and then based on that distance, estimate the likelihood that the point shares properties with the set). This is essentially because neural networks perform information processing in parallel and so offer better explanations of single-step problem solving. Indeed, we view humans' ability to use strategic cognitive sequences as an exaptation of our ancestral brain machinery, one that makes us much smarter by allowing us to access a range of strategies that lower animals largely cannot use [21, 22].

Because of the weaknesses in each approach and the paucity of other mechanistically detailed, cumulative models of cognition, we take inspiration from the well-developed theories from production systems about how cognitive steps are sequenced [17–19, 23] while focusing on artificial neural network-centered theories on the specifics of how individual cognitive actions are performed. This perspective is influenced by prior work on hybrid theories and cognitive architectures based on ACT-R and Leabra networks for a different purpose [24]. ACT-R [18] is the most extensively developed production system architecture and the one which most explicitly addresses physiology, while Leabra is arguably the most extensively developed and cumulative theory of neural function that spans from the neural to cognitive levels.

In ACT-R, the sequence of cognitive actions is determined by which production fires. This in turn is based upon the "fit" between the conditions of each production and the current state of the cognitive system (which also reflects the state of the environment through its sensory systems). This function has been proposed to happen in the basal ganglia (BG) [25, 26], and this has been borne out through matches with human neuroimaging data [25]. While it is possible that the BG is solely responsible for action selection in well-practiced cases [27], we focus on the learning process and so on less well-practiced cases. In our neural network framework, we divide this functionality between cortical and BG areas, with the cortex (usually PFC) generating a set of possible cognitive actions that might be performed next (through associative pattern matching or "constraint satisfaction"), while the basal ganglia decides whether to perform each candidate action, based on its prior relationship to reward signals in similar circumstances.

In modeling this process, we draw upon previous work from our group in modeling the mechanisms and computations by which the PFC and BG learn to maintain useful information in working memory [28–32]. The prefrontal cortex basal ganglia working memory (PBWM) models developed by O'Reilly and colleagues integrate a wealth of electrophysiological, anatomical, and behavioral data, largely from animal work. Working memory also appears to be a large component of executive function, because in many cases a specific task is performed by virtue of maintaining an appropriate task set [33], in effect remembering what to do. Those maintained representations bias other brain processing through constraint satisfaction. Because it explains the deep question of how we learn our executive function (EF), this theory makes progress in dispelling the "homunculus" [30], by explaining how complex cognitive acts are performed by a collection of systems, each of which supplies a small part of the overall intelligence, decision making, and learning.

In essence, the PBWM framework extends the wealth of knowledge on the role of the basal ganglia in motor control to address working memory and executive function. This is possible because there are striking regularities across areas of frontal cortex, so that the anatomy of cortex and basal ganglia that subserves motor function is highly similar to prefrontal and anterior BG areas known to subserve WM and EF [34]. This anatomy is thought to help select potential motor actions by "gating" that information through thalamus back to cortex, amplifying it and so cleanly selecting one of the several possible candidate actions represented in the cortex (e.g., [35]). The core hypothesis of PBWM is that these same circuits help select which representations will be actively maintained in PFC by fostering local reverberant loops in the cortex, and between cortex and thalamus, and by triggering intrinsic maintenance currents that enable self-sustained persistent firing in cortical pyramidal neurons. The reinforcement learning mechanisms by which BG learns which actions are rewarding also apply to learning what to remember and so what to do.

The primary value and learned value (PVLV) model of dopamine release as change in reward prediction [36, 37] is also a key component of PBWM and is in turn based on electrophysiological and behavioral data from a collection of subcortical areas known to be involved (e.g., [38–41]). The known properties of dopamine release indicate that it serves as a reward prediction error signal [42] which has informational properties that make it useful for driving learning [43, 44]. This system learns to signal when a new set of representations will likely lead to reward in a biologically realistic variant of the function of the better-known temporal difference (TD) algorithm when it is supplemented with "eligibility trace" information (e.g., [45]). This reward prediction function is crucial, because the difficulty in assessing the benefit of an action (whether it be cognitive or behavioral) is that the actual reward achieved by that action very often occurs later in time and so cannot be used directly as a learning signal [46, 47]. Instead, the system learns to perform actions that are predicted to gain reward. This reinforcement learning trains the striatum and works alongside the more powerful associative and error-driven learning within the PFC portion of PBWM that learns the representations (and therefore the associative semantics) of candidate actions to take.

In the remainder of the paper, we present an overview of four models that elaborate on this process in several ways. The first addresses how the learning mechanisms described previously and elaborated upon in works by various workers in our group [36, 37, 48, 49] can learn short sequences of cognitive actions, when they are sequentially dependent and so must be performed in the right order to achieve reward. The second describes how the hippocampus can achieve instructed learning, participating in the constraint satisfaction process of deciding which action to consider performing, as when we perform a novel task based on memorized instructions. The third model considers how slow, cortical associative learning can contribute to that same "which" process by using constraints of the current state and the goal to arrive at a subgoal that can serve as a viable next step in the sequence. Finally, we close with some discussion of the state of this research and the many remaining questions.

3. Model-Free Reinforcement Learning

Model-free reinforcement learning (RL) can be defined at a high level as learning which actions (which we take to include cognitive "actions") produce reward, without any other knowledge about the world [43]. While the learning mechanisms we describe here are purely trial and error, the same learning mechanisms apply to model-driven or "hypothesis-driven" learning as well. For instance, the same learning principles apply when using actions, explicit plans from memorized instructions, or semantic associations as outlined in the final two models we describe later.

In hopes of better understanding how this process could occur in neural tissue, we have leveraged the prefrontal cortex basal ganglia working memory framework, or PBWM [28–32]. Under this account, a basic actor-critic architecture [43, 50] naturally arises between the prefrontal cortex (PFC), the basal ganglia (BG), and the midbrain dopamine system as modeled by our PVLV system described previously. PVLV serves as the critic, evaluating the state of the network and providing dopamine bursts or dips for better than and worse than expected outcomes, respectively. The BG system is naturally situated to perform the functions of the actor based on its known role in selecting motor actions (and by

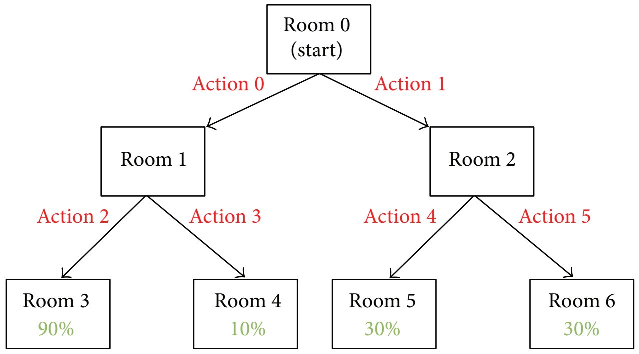


Figure 1: Simple state-based room navigation task. The percentages of the last level of rooms at the bottom of the figure represent the probability that the agent will get rewarded if it chooses the path that leads to the respective rooms.

hypothesis, selecting cognitive actions with analogous neural mechanisms in more anterior regions of PFC). Using the critic's input, the BG learns from experience a policy of updating segregated portions of the PFC as task contingencies change. The PFC is able to maintain past context and provides a temporally extended biasing influence on the other parts of the system. It is helpful to view this entire process as a "gating" procedure: the BG gating controls that are being actively maintained within the PFC, and therefore subsequently biasing (controlling processing in) other cortical areas. When the gate is closed, however, the contents of the PFC are robustly maintained and relatively protected from competing inputs. Importantly, as task contingencies change and the actor determines that a change is needed, the gate can be opened allowing new, potentially more task appropriate, content into the PFC.

The simple RL-based learning of the PBWM framework allows us to easily and naturally investigate one manner in which the brain may be capable of utilizing model-free RL in order to solve a simple task. In short, the network must learn to maintain the specific actions taken and evaluate this sequence based on either the success or failure of a simulated agent to attain reward. The simple example task we use is a basic state-based navigation task (abstracted at the level of "rooms" as states) in which a simulated agent must navigate a state space with probabilistic rewards as inspired by the work of Fu and Anderson [51] (see Figure 1). The goal of the task is simply to learn an action policy that leads to the highest amount of reward. To achieve this, the agent must make a choice in each room/state it visits to move either to the next room to the left or the next room to right but always moving forward. The only rooms that contain reward are at the final level (as in most tasks). The structure of the reward is probabilistic, so a single room is the most consistent provider of reward (Room 3 in Figure 1), but the others have a lower chance to be rewarding as well. In order for the PBWM framework to ultimately succeed, it must be able to maintain a short history of the actions it took and reward or punish these action choices in the final presence or absence of reward. This is a simple task, but a learning in this way is a valuable tool when the system must learn basic actions first in order to succeed at more extensive cognitive sequencing tasks.

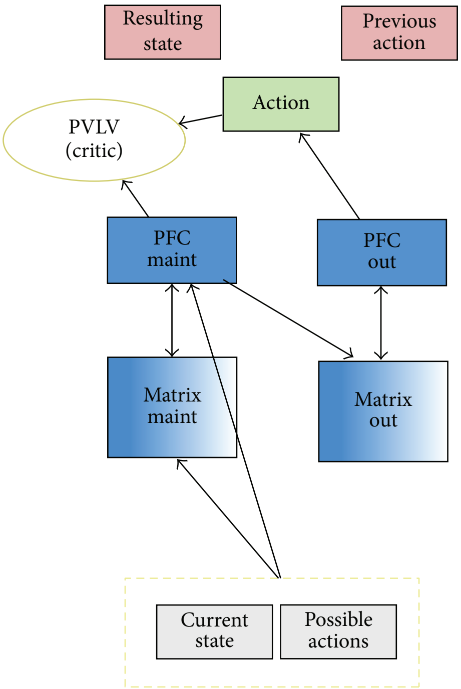


Figure 2: Model-free network architecture. Based on both current state and possible actions, the "matrix maint" determines what to maintain in PFC. Based on the stored information in PFC, "matrix out" determines the next chosen action via PFC out. PVLV (consisting of multiple biological systems) evaluates the actions (critic) and helps train the model. See text for in-depth description and functions of the various components of the network. Detailed network architecture is highly similar to the PBDM model discussed later.

3.1. Description of the Model. The model-free RL network is depicted in Figure 2. The ultimate goal of the network is to receive reward by determining the best action to take given the reward structure of the simulated environment. There are many models of reinforcement learning in similar domains, and the PBWM and PBDM models have been applied to learning in superficially similar domains. However, some very important differences make the setup of this model unique. Most importantly, the final outcome (what Room the network ends up in based on the action chosen) of the network is not determined in the standard neural network manner of having activation propagate through units and having a competition that determines the winner. Instead, the network chooses an action via the *action* layer, which is the only traditional output layer in the network. The possible actions can be thought of as any atomic action that may result in a change of state, such as "go left" or "go right." After the network chooses an action, a state transition table is used to determine the outcome of the action. More specifically, the network makes a decision about what action to take, and program code determines what the effect is on the environment of the simulated agent. The outcome is reported back to the network via the *resulting state* layer, but for display purposes only (not used in any computation). The example trial stepped through later in this section will help to clarify this process.

3.1.1. Network Layer Descriptions

(i) Action layer: this is the output of the network. The chosen action is used via a state transition table to choose a new room. In the current simulation, the room choice is completely deterministic based on the action.

(ii) CurrentState layer: this is a standard input layer. The CurrentState is the current state (room) that the model is occupying.

(iii) PossibleActions layer: this is the second input layer. The layer is used to specify what "legal" actions are based on the current state that the network is occupying. Importantly, PossibleActions provides the main signal to the simulated basal ganglia to determine the gating policy, as well as the main input to the PFC. This ensures that only legal actions should be chosen (gated) at any given time.

(iv) PreviousAction layer (display only): this is a display only layer. It maintains the last action choice that the network made. This can be useful to understand how the network arrived to its current state.

(v) ResultingState layer (display only): this is a display only layer. The ResultingState is the "room" that the simulated agent will arrive in based on the action that the network produced. The final room is used to determine if the agent should receive reward.

(vi) PVLV layers: the PVLV layer(s) represents various brain systems believed to be involved in the evaluative computations of the critic [36].

(vii) PFC maint and PFC out: simulated prefrontal cortex layers, the maint PFC is used to actively maintain information overextended delay period. The PFC out layer models the process of releasing this information, allowing it to affect downstream cortical areas and drive actual responses.

(viii) Matrix maint and matrix out: these layers are used to model the basal ganglia system and represent the actor portion of the network. They learn to gate portions of the PFC, through experience, using the information provided from the PVLV system.

3.1.2. Task Example

(1) The current state (room) is presented to the network via the CurrentState layer. The inputs correspond to different rooms as shown in Figure 1, where Room 0 corresponds to the first unit in CurrentState layer, Room 1 to the second, Room 2 to the third, and so forth.

(2) Using the CurrentState and the actions maintained within the PFC, the network must decide to go to the room to the left or the room to the right. This decision is reflected by activation in the action layer.

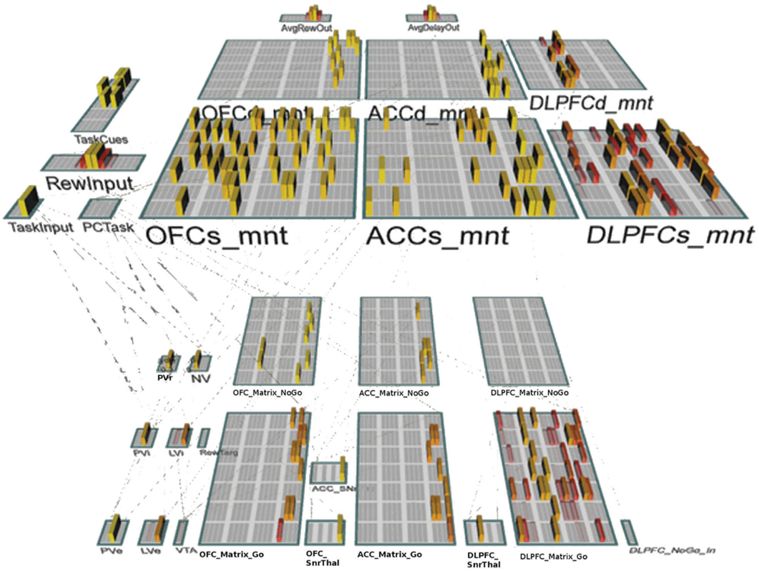
(3) The action that is chosen by the network is used to determine where the simulated agent is in the current state space, and this is accomplished using a standard transition table to look up the next room. The actions are deterministic and move the agent directly to the room based only on the action.

(4) The resulting state of the agent is returned to the network via activation in the CurrentState layer indicating the result of the action. Return to Step 2 unless the agent reaches a terminal room.

(5) If the room reached by the agent is a final room, the reward probabilities for that room are used to determine the likelihood of reward to the agent.

(6) Repeat from Step 1 until task is reliably learned.

3.2. Results. The network is capable of quickly learning the optimal policy of action sequences that optimize its reward on this task. To assess the ability of the network to solve this task, we set up a testing structure which allowed the network 75 "chances" to solve the task per epoch (block). At the end of the epoch, the average rate of reward was recorded for the simulated agent. This was repeated until either the agent received an average reward greater than 85% of the time or for 25 epochs (blocks), whichever came first. Ten simulated agents were ran, and 8 out of the 10 reached criteria of 85% average reward within 25 epochs. On average, it took 4 epochs to achieve this feat. While this may not appear to be a surprising result, the complex nature of the biologically realistic network made this far from a forgone conclusion. Indeed, many insights were gained about the nature of how the actor must balance its exploring of the state space with gaining reward. If the network randomly gets reward in one of the low-reward states, it must still be willing to explore its environment in order to confirm this finding. Conversely, if the network is in the high-reward state and does not receive reward, the (relative) punishment for this nonreward needs to allow a possible return to this same state at some point in the future in order to discover the optimal action policy. The limits of the framework are apparent in the 2 networks that did not reach criteria. In both of these cases, the agent randomly reached the low probability area of state space. In most cases, the agent is able to successfully explore other options and thus find the more rewarding rooms. However, the current PBWM framework will occasionally fail if reward is not present early enough in exploration process. We are investigating biologically inspired mechanisms to bootstrap the learning in more efficient ways. Encouraged by our initial framework, we are actively investigating how a simple model-free approach to learning basic sequences could be utilized by the human brain in order to scaffold up to more complex and interesting sequences. We are hopeful that concentrating on the relevant biological data and learning will provide us with useful insights to help us better understand how people



PBDM net: showing activations

Figure 3: PBDM decision-making model. This figure shows the PBDM network and the components it models. The bars show the activation strengths of each of the units in the model for a particular point in time.

are capable of such effortless sequencing of extended, diverse, and complex action plans.

We hypothesize that this type of learning aids in cognitive sequencing by allowing humans to discover useful simple sequences of cognitive actions purely by trial and error. While this learning does not likely account for the more impressive feats of human cognition, since these seem to require substantial semantic models of the relevant domain and/or explicit instruction in useful sequences, we feel that understanding what the brain can accomplish without these aids is necessary to understanding how the many relevant mechanisms work together to accomplish useful strategic cognitive sequencing.

4. Prefrontal Cortex Basal Ganglia Decision-Making (PBDM) Model

In the PBDM model, we primarily address decision making at the level of task strategies (task set representations in PFC, primarily dorsolateral PFC (DLPFC)). Decision making is important in many areas, but the selection of strategies for complex tasks is our focus. We believe that the same mechanisms apply to making decisions in many different domains.

The main idea behind PBDM is to computationally model the interactions between basal ganglia and medial prefrontal areas that represent particularly relevant information for making action plan or strategy decisions. Anterior cingulate cortex (ACC) and orbitofrontal cortex (OFC) serve as activation-based monitors of task affective value parameters [52, 53], including action effort in the ACC [54], and probability of reward in the OFC. These then project to the basal ganglia that controls updating in the DLPFC, giving it the necessary information to select choices in favor of lower effort and higher reward strategies. Because the ACC and OFC are themselves PFC areas with inputs from the same type of basal ganglia/thalamic circuits as motor and working memory areas, they are hypothesized to be able to rapidly update and maintain their value representations and, with a single gating action, change the evaluation to reflect new important information. This confers great flexibility and rapid adaptability to rapidly changing circumstances. Within this framework, several questions remain: what, more precisely, do the ACC and OFC represent? How can these representations drive appropriate gating behavior in the DLPFC BG? How are appropriate representations engaged in novel task contexts?

In the initial version of the PBDM model, described in more detail later and shown in Figure 3, we adopt simple provisional answers to these questions while recognizing that these likely underestimate the complexity of what happens in the real system. In particular, while ACC is often (and in our model) assumed to represent effort, its true role is more complex. The current state of knowledge on these issues is reviewed thoroughly by Kennerley and Walton [55]. The ACC and OFC in our model compute running time-averaged estimates of effort and reward probability, respectively, based on phasic inputs on each trial. If a task is ongoing, the ACC just increases its running average of time-effort by one. When a reward value is received or not (when otherwise expected), the OFC increments its running average estimate of reward probability. We have four different stripes within the ACC and OFC, each of which receives input from and so has a representation determined by one of the task strategies represented in the parietal cortex. These are thought of as very general strategies for dealing with spatial information, and over a lifetime of experience, we build up reasonable estimates of how effortful and rewarding they are on average in similar tasks. In order to in part capture the importance of context, there is also a randomly updated set of task features, which represent specific details about each different task that the model learns to perform. Over time, the model learns to pay attention to the ACC/OFC value representations in selecting a task strategy and pay less attention to these idiosyncratic task cues. Having done so, the model can then generalize to novel task contexts, by paying attention to the underlying spatial task values and ignoring the novel task features. Then, as the novel task progresses, actual experienced reward and effort drive the ACC and OFC representations, providing a more accurate picture for decision making going forward. This is the overall model we think applies to subjects as they engage in novel tasks with multiple possible strategies.

We conceptualize this PBDM process as engaging when people are actively and explicitly considering a new strategy or similar decision. We model an abstract spatial task, in which the strategies consist of individual spatial properties of groups of similar items. Strategies consist of considering one or a combination of these properties. There are 4 different strategies considered (listed by increasing order of both effort and reward probability; the precise values vary by task): Distance Only, Distance + BaseRate, Distance + Radius, and Distance + BaseRate + Radius. These are merely example strategies associated with a hypothetical spatial estimation task and are therefore sometimes also simply referred to as strategies 0 to 3, respectively; the task is not implemented for this model outside of entirely hypothetical probabilities of success (reward) and level of effort (time to implement). The weights for the PBDM component are trained to model a long history of experience with these hypothetical reward and effort values. After this learning (and purely through it), the OFC reward representations primarily bias the Go pathway, while the ACC effort representations bias the NoGo pathway. It is this balance between Go and NoGo that then ultimately determines the strategy selected. In our models, we observe that different random initial weights produce different individual preferences along this tradeoff.

The network performs various tasks (which switch every 10 trials during pretraining, simulating the intermixed variety of spatial tasks a person encounters during their daily life). The probability of reward and the number of trials required are determined by the selected strategy, the task representation that the DLPFC maintains. In reality, the possible strategies and therefore the representational space would be much larger, but we have narrowed it down to just 4 different states in a localist representation, (called Distance, Dist + Base Rate, Dist + Radius, and Dist + BaseRate + Radius; the original relation of these strategies to a particular task is irrelevant since the base task was abstracted to only the strategy component for this model). The inner loop per trial consists of "performing" the task in question, which happens through task-specific areas responding to the DLPFC task representation. We model that process here only at the most abstract level: each strategy takes an amount of time and has a probability of success that varies for each possible task. Thus, the PBDM network only experiences the overall feedback parameters: number of trials and probability of reward at the end of those trials. We do not model the process of carrying out these strategies; each of the models here could also be applied to understanding how a particular strategy unfolds into an appropriate sequence of cognitive actions.

The overall behavior is thus as follows: select a DLPFC task representation, run a number of blank trials (blank since we assume that the lower-level processes that carry out the strategy have little influence on this level of cortical machinery) according to the "effort" parameter (representing task performance), then receive reward with given probability determined by the PCTask representation that the DLPFC task representation drives, and then repeat. Over time, the BG gating units for the DLPFC are shaped by the effort/delay and reward parameters, to select DLPFC stripes, and associated reps that are associated with greater success and shorter delays.

The BG "Matrix" layer units control gating in DLPFC and so, ultimately, make final decisions on strategy choice. They receive inputs from the ACC and OFC, which learn over time to encode, using dynamic activation-based updating, running time averages of reward and effort, associated with the different strategies on the different tasks. Because we assume that mental effort is equal per unit time across strategies, the effort integration is identical to time integration in this case. Critically, because this is done in activation space, these can update immediately to reflect the current PCTask context. Over time, the BG learns weights that associate each OFC and ACC unit with its corresponding probability of success or effort. Thus, an immediate activation-based update of the ACC and OFC layers will immediately control gating selection of the DLPFC layers, so that the system can quickly change its decision making in response to changing task contexts [52, 56, 57].

Thus, the early part of the network training represents a developmental time period when the ACC and OFC are learning to perform their time-averaging functions, and the DLPFC BG is learning what their units/representations correspond to in terms of actual probability of reward and effort experienced. Then, in the later part, as the DLPFC task representations continue to be challenged with new task cue inputs (different specific versions of this task space), the learned ACC/OFC projections into DLPFC BG enable it to select a good task strategy representation on the first try.

4.1. Details of Network Layer Functions

(i) TaskInput: generalized task control information about the inner loop task being performed projects to DLPFC. We assume that this information comes from abstract semantic representations of the task at hand; this is likely represented in a variety of posterior and prefrontal regions, depending on the type of task. Use the following units/localist representations:

(a) PERF—performing current task-signals that DLPFC should not update the task representation (see DLPFC NoGo In later); this repeats for the number of trials a given PCTask strategy requires and metes out the delay/effort associated with a given strategy.

(b) DONE—done performing current task-reward feedback will be received in RewInput to OFC and PVe (PVLV); note that there is a "cortical" distributed scalar value representation of reward (RewInput), in addition to the subcortical one that goes directly into the reward learning system (PVe); conceptually these are the same representation, but their implementation differs.

(c) CHOICE—DLPFC should choose a new task representation, based on influences from TaskCues, ACC, and OFC states; the newly gated DLPFC representation will then drive a new PCTask representation, which will then determine how many PERF trials are required and the probability of reward for the next DONE state.

(ii) TaskCues: these are random bit patterns determined by the cur_task_no state, which drives DLPFC (both cortex and BG); they represent all the sensory, contextual, and instructional cues associated with a given specific task.

(iii) PCTask reflects the actual task parameters. In this example, these are Distance, Dist + BaseRate, Dist + Radius, and Dist + BaseRate + Radius, but more generally this would represent a much larger space of task representations that have associated reward and effort parameters for different tasks. This may also reflect a combination of posterior cortical and also more posterior DLPFC representations that provide topdown biasing to these PC task representations and maintain them over shorter durations. The DLPFC in the model is the more anterior "outer loop" DLPFC that maintains higher-level, longer-duration task representations that are "unfolded" into useful sequences by other processes, including but likely not limited to those we address in the models here.

(iv) RewInput: scalar val of reward input level activated during the DONE trial; this also has a −1 state that is activated whenever the network is in PERF task mode, and this is what triggers the incrementing of delay/effort in the ACC layer (i.e., both OFC and ACC feed off of this same basic RewInput layer, pulling out different information). This is overall redundant with signals in PVLV but packages them in a simple way for OFC/ACC to access and for us to manipulate for various experiments.

(v) OFC computes running time average of reward probability/magnitude; only updated when reward occurs (DONE trials), otherwise maintains the current estimate for PERF and CHOICE trials. The network learns coarse-coded distributed representation of this value, not in a scalar value format, through a "decoding" layer (AvgRewOut) that is in scalar value format. But it is the distributed representation that projects to DLPFC to bias its processing. It is not exactly clear what AvgRewOut corresponds to biologically, but the general idea is that there are autonomic level states in the brainstem, and so forth, that compute low-level time averages based on physiological variables (e.g., longer time average sucrose concentration in the blood), and that is what drives the OFC to learn to compute activation-based running time averages. See (vii) for the way this representation learns to affect DLPFC gating.

(vi) ACC computes running time-average interval between reward trials which constitutes total effort on each task, since we assume roughly equal effort per time. It is updated on each performance trial and maintained during the DONE and CHOICE trials; each time step increases activation. As with OFC, this layer learns coarse-coded distributed representation of this value, not in a scalar value format, through a "decoding" layer (AvgDelayOut), which again reflects longer time-average metabolic cost variables.

(vii) DLPFC encodes current task strategy and learns representations entirely through reinforcement learning stabilization. It receives information about each task from TaskCues; the Matrix layer also receives from ACC and OFC and learns over time to select task representations associated with good values of ACC and OFC (i.e., values of those that have been associated with rewards in the past). DLPFC also projects to PCTask, which in turn projects to ACC and OFC and "conditionalizes" (makes appropriate to the particular task) their representations.

(viii) DLPFC_NoGo_In is our one "hack." It turns on NoGo (strongly) whenever a task is being performed to ensure that the matrix does not update DLPFC midtask. This hard-coded behavior is simply the assumption that the DLPFC task set representation remains active during task performance; that is, people maintain one task set without switching strategies midway through more general learning: when you decide on a strategy, stick with it until you are done (or until it gets "frustrating" by consuming too much time).

4.2. Results

4.2.1. Reward-Only Optimization: OFC Proof of Concept Test. The first proof of concept test sets the probability of reward to .2, .4, .6, and .8 for PCTask units 0–3, respectively (labeled "Distance only," "+BaseRate," "+radius," and "Combined," resp.), with delay set to a constant 1 trial (.2 parameter × 5 trials max delay) for all options. Thus, the best strategy is to select strategy 3, based on OFC inputs. As shown in Figure 4, the network does this through a period of exploration followed by "exploitation" of strategy 3, which is selected automatically and optimally immediately, despite changing TaskCues inputs. All of the batches (10/10) exhibited this same qualitative behavior, with a few stabilizing on strategy 2 instead of 3. This was the second-best strategy, and the fact that the model stabilized on this in some cases shows the stochastic process of sampling success that likely contributes to the selection of nonoptimal strategies in

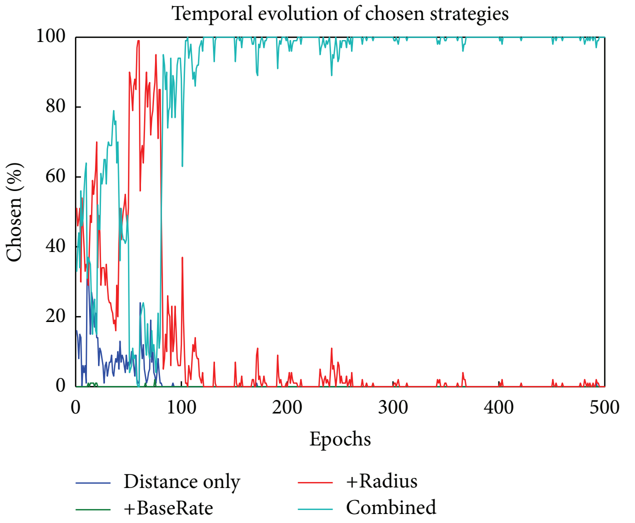


Figure 4: Developmental learning trajectory of PCTask selection. Early in learning it explores the different strategies, and later it learns to predominantly select the one (green line, strategy 3 ("Combined")) that produces the best results.

some real-life cases (since after the model stabilizes, it will not learn about potentially better strategies without some sort of external perturbation to force resampling). None stabilized on 0 or 1, since they have substantially lower reward probabilities. As shown in Figure 5, the weights into the Matrix Go stripe that gates DLPFC learned to encode the high-value OFC representations associated with the strategy 3 OFC representation.

4.2.2. Delay-Only Optimization: ACC Proof of Concept Test. Next, we set probability to .6 for all strategies and set the delay factors to 1, 2, 3, and 4 trials of delay, respectively, for strategies 0–3. Without any PVLV feedback at all during the PERF trials, the network does appear to be sensitive to this delay factor, with strategy 0 (1 trial delay) being chosen preferentially. However, this preference is somewhat weak, and to produce stronger, more reliable preferences, we added a direct dopaminergic cost signal associated with delay, as has been shown empirically [58]. This modulation decreased the size of a DA reward burst in proportion to effort/delay (with a small weighting term). In our proof of concept test, this small modulation produced 50% of networks preferring the first (least delay) strategy.

4.2.3. Balanced Reward and Delay (Actual Use Case). To simulate a plausible situation where there is a tradeoff between effort and reward, we set the reward factors to .4, .6, .6, and .8 and the delay factors to .2, .4, .6, and .8. This resulted in a mix of different strategies emerging over training across different random initial weights ("batches") (proportions shown in Figure 6), with some preferring the low-effort, low-reward distance only option, while others going for the full Distance + BaseRate + Radius high-effort, high-reward case, and others falling in between. The particular results are highly stochastic and a product of our particular choices of reward and effort values; it is easy to push these preferences around by using different weightings of effort versus time.

4.3. Discussion. The PBDM model shows how rapid updating in prefrontal cortex (as captured in the PBWM models and related work on persistent firing in PFC) can aid in decision making by allowing the system to use contextually appropriate representations of predicted reward and effort to drive decisions on task strategy. If the context (e.g., physical environment and task instructions) remains the same, then new learning in the ACC and OFC slowly updates the values of predicted reward and effort through weight-based learning. If, however, the context changes, representations in ACC and OFC will be "gated out," so that a new set of neurons learns about the new context. Detailed predictions about the old context are thus preserved in the synaptic weights to that now silent units (because the learning rule we use, and most others, does not adjust weights to inactive neurons/units).

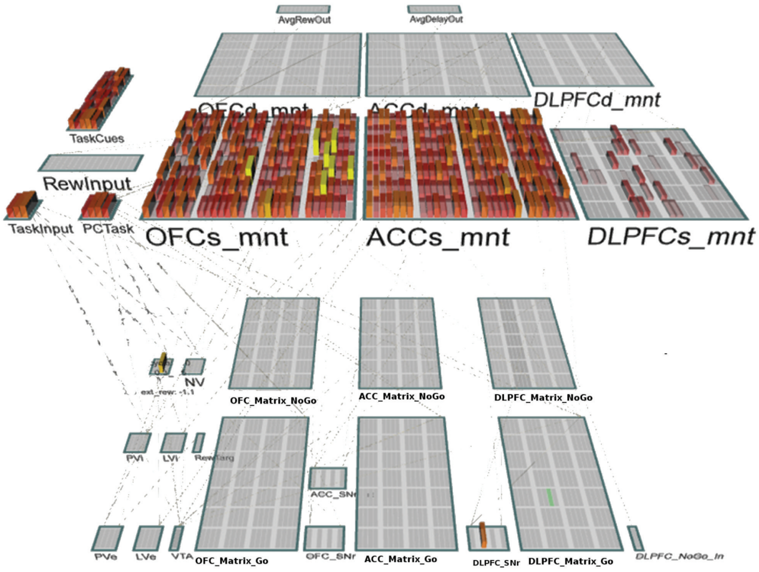
One way in which this preservation of contextually dependent ACC and OFC representations could be extremely useful is in interaction with episodic memory in the HC. We believe that predictive representations could also be retrieved to ACC and OFC from episodic memory in the hippocampus, a form of PFC-HC interaction similar to but importantly different from that we model in the "instructed learning" model.

This model primarily addresses the "strategic" component of strategic cognitive sequencing, but this type of effortful decision making, bringing the whole predictive power of cortex online to estimate payoff and cost of one possible sequence component, could help bootstrap learning through the mechanisms in either or both of the instructed learning and "model-free" models.

5. Instructed Learning

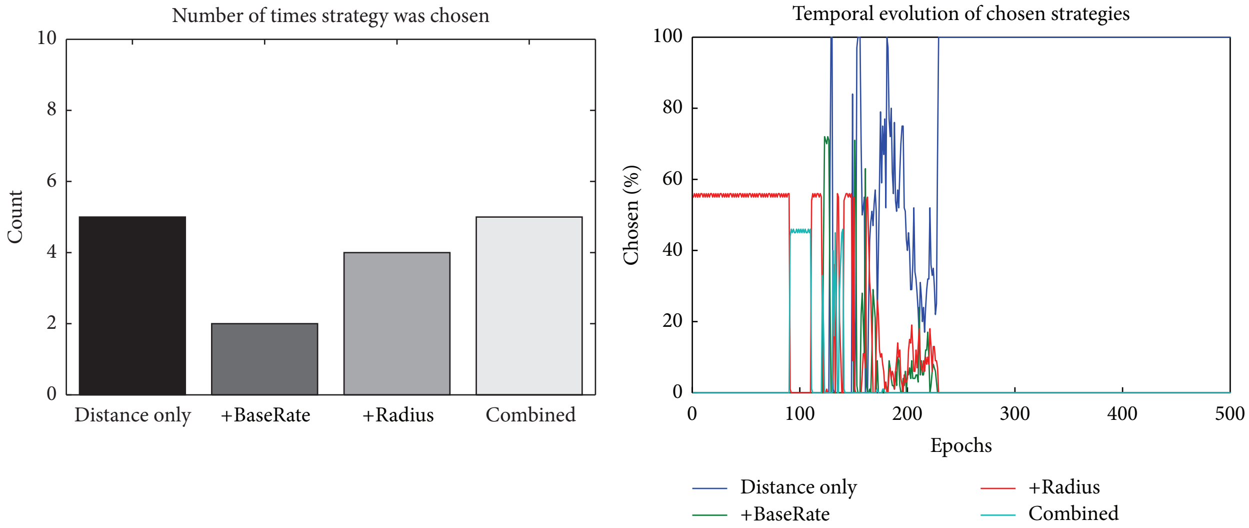
One source of complex, strategic cognitive sequences is learning them directly from instruction [59–61]. Humans have the remarkable ability to learn from the wisdom of others. We can take advice or follow instruction to perform a particular cognitive sequence. One such example may be observed daily by cognitive scientists who conduct human laboratory experiments. Most normal participants can well implement instructions of an arbitrary novel task with little or no practice. However, in the cognitive neuroscience of learning, reinforcement learning has been the central research topic and instructed learning appears to have been relatively understudied to date. In this section, we contrast reinforcement and instructed learning and outline the dynamics of instruction following in a biologically realistic neural model.

Reinforcement learning adapts behavior based on the consequences of actions, whereas instructed learning adapts behavior in accordance with instructed action rules. As a result, unlike the slow, retrospective process of trial and error in reinforcement learning, instructed learning tends to be fast, proactive, and error-free. In the brain, the neurotransmitter dopamine signals reward prediction errors for the basal ganglia to carry out reinforcement learning of



PBDM net: showing weight strengths

Figure 5: PBDM decision-making model. This figure shows the weights from the respective units to a unit in the DLPFC_Matrix_Go layer (green, lower right). It depicts the strength of weights towards the end of learning, at which point there are particularly strong connections from the core OFC distributed representations, which represent strategy's predicted reward value, established through learning.



(a)

(b)

Figure 6: Balanced reward and delay. The left graph shows the number of times a strategy was chosen over 16 repeats with random initial weights, while the graph on the right shows the temporal evolution of selection for one randomly chosen network. The variability in the equilibrium strategy choice stems from the balance between reward and delay (the higher the reward, the higher the delay) making each strategy approximately equally rational to choose. As discussed in the reward-only case previously, the particular, random history of reward plays a large role in determining the ultimate strategy choice.

reward-linked actions (for a discussion, see [62]). As for instructed learning, the human posterior hippocampus underlies verbal encoding into episodic memory [63] and use of conceptual knowledge in a perceptually novel setting [64].

Compared to reinforcement learning, instructed learning appears effortless. Why is learning so arduous in one mode but effortless in another? How exactly do we perform complex novel tasks on the first attempt? We propose that instruction offers nothing but a new plan of recombining old tricks that have been acquired through other forms of learning. In other words, instructions quickly assemble rather than slowly modify preexisting elements of perceptual and motor knowledge. For example, we can immediately follow the instruction: "press the left button when seeing a triangle; press the right button when seeing a square," in which the action of button press is a preexisting motor

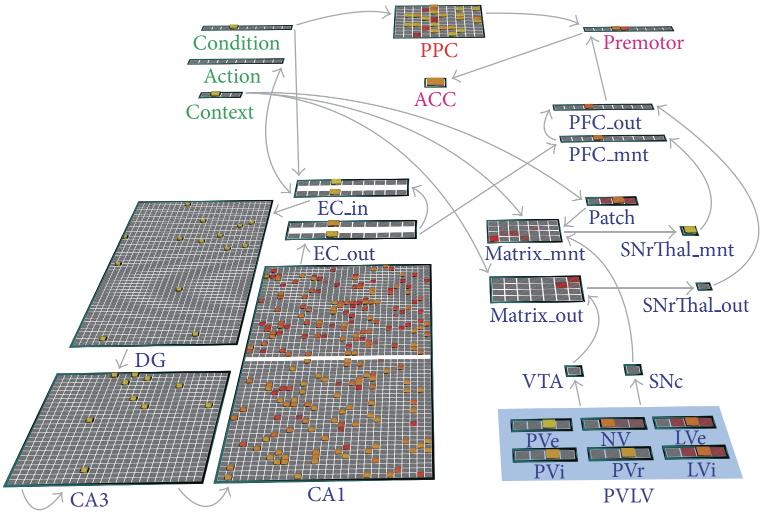


Figure 7: The instructed learning model. The model consists of two interactive learning pathways. The hippocampal-prefrontal pathway (i.e., lower part in the diagram) processes newly instructed conditional-action rules, whereas the parietal pathway (i.e., upper part in the diagram) processes habitual actions. The actions suggested by each of these pathways are then gated by the PFC portion.

skill, and visual recognition and categorization of shapes are also an already learned perceptual ability. Note also that understanding the instruction requires a previously learned mapping from language (e.g., the verbal command of "press") to actual behavior (e.g., the motor execution of "press").

To further study how instruction following is carried out from neural to behavioral levels, we constructed a model of instructed learning based upon known neuroanatomical and neurophysiological properties of the hippocampus and the prefrontal-basal ganglia circuits (Figure 7). Specifically, the model basal ganglia (BG) carries out reinforcement learning of motor execution (abstracted in the model to premotor); the model hippocampus rapidly encodes instructions as action episodes that can be contextually retrieved into the prefrontal cortex (PFC) as a goal for guiding subsequent behavior. Unlike a single-purpose neural network that slowly rewires the whole system to learn a new sensorimotor transformation, this general purpose instructable model separates motor from plan representations and restricts plan updating to lie within the fast-learning hippocampus, which is known to rapidly bind information into episodic memories.

As a concrete example, the proposed model is instructed with 10 novel pairs of if-then rules (e.g., if you see A, then do B) and evaluated for its success in performing conditional actions (e.g., do B) when encountering a specific condition (e.g., seeing A). In the model, each of the "Condition," "Action," and "Premotor" layers consists of 10 localist representations of conditions, verbal actions, and (pre-)motor outputs, respectively. The model is pretrained with action-to-motor mappings (i.e., from verbal commands to premotor responses) during the Pretraining stage and then trained with condition-to-action mappings (i.e., if-then rules) during the Instruction stage. Finally, during the Performance stage, it is tested with Condition-to-Motor mappings without any inputs from the "Action" layer. The simulation results are shown in Figure 8. The model quickly learns an if-then rule in just few trials during the Instruction stage, and without further practice, it makes no error in carrying out these instructions for response during the Performance stage, just as human subjects often do after being presented with clear instructions and a short practice period.

Inside the model, learning occurs in multiple parts of the architecture. During the Pretraining stage, the hippocampus learns to perform identity mapping for relaying information from the "Action" layer to the corresponding motor representations in the PFC layers. Meanwhile, BG learns to open the execution gate for PFC to output a motor decision to the "Premotor" layer. During the Instruction stage, the hippocampus associates inputs from the "Condition" and "Action" layers and learns each condition-action pair as a pattern. During the Performance stage, all the model components work together using mechanisms of pattern completion, and the hippocampus recalls instructions about what action to do based on retrieval cues from the "Condition" layer, and its downstream PFC either maintains a retrieved premotor command in working memory when BG closes the execution gate or further triggers a motor response in the "Premotor" layer when BG opens the execution gate.

Compared to earlier work on instructable networks [65], our model further explicates how different parts of the brain system coordinate to rapidly learn and implement instructions. Albeit simple, our instructed learning mechanisms can support strategic cognitive sequencing in that a cognitive sequence can be constructed from an ordered set of instructed or self-instructed operations. Beside sequential behavior, the model is being extended to also explain the interactions between instructions and experience (e.g., [66–69]) in the context of confirmation bias and hypothesis testing. The modeled ability of the hippocampus to memorize specific contingencies in one shot undoubtedly contributes an important piece of our ability to learn complex goal-oriented sequences of cognitive actions. Beyond simply memorizing

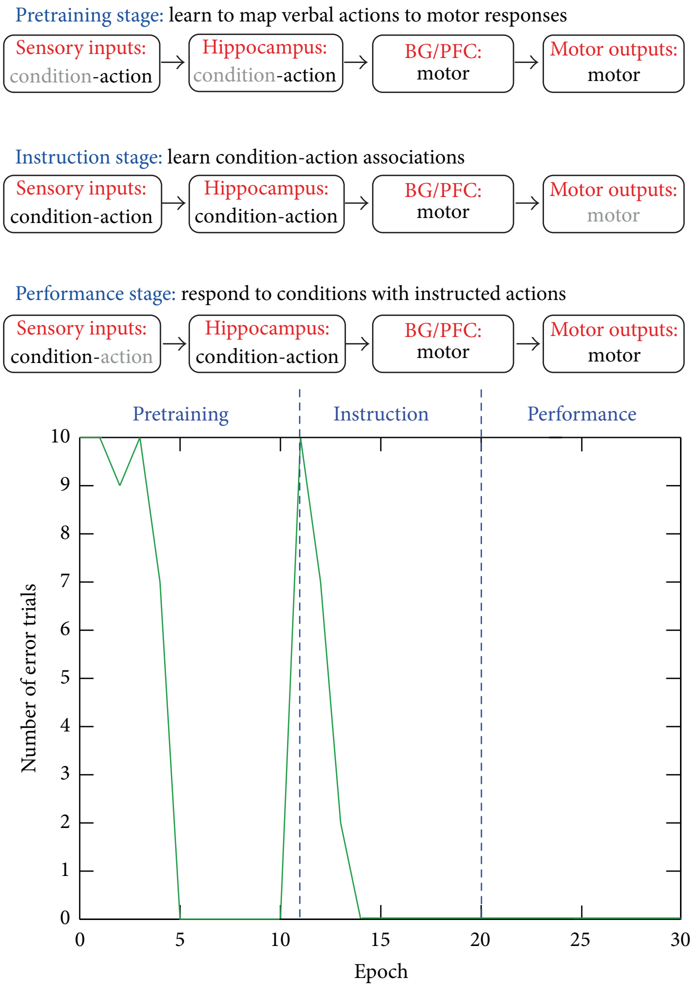


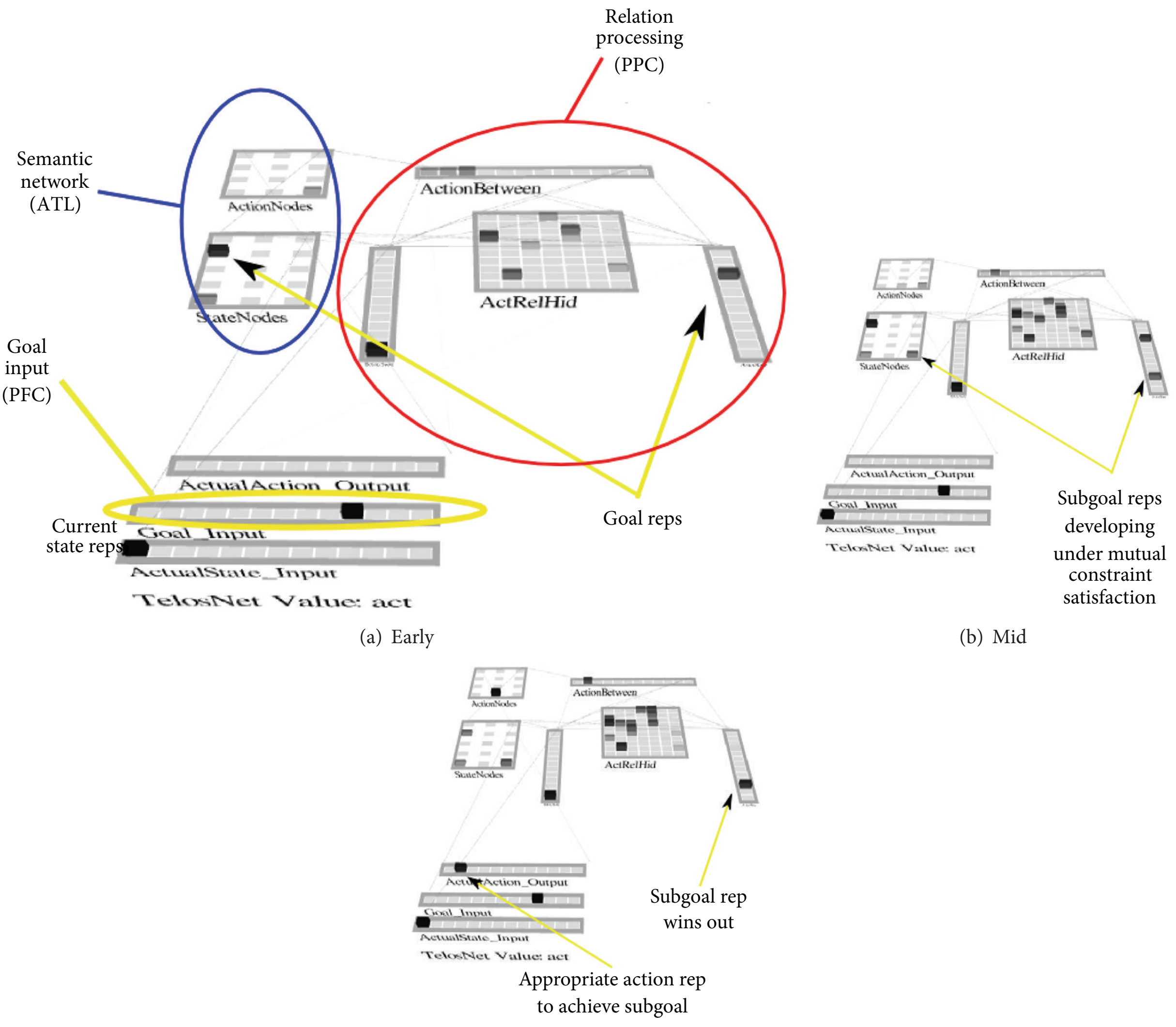
FIGURE 8: Instructed learning stages and simulation results. In the upper panel, black and grey texts denote present and absent representations, respectively. In the bottom panel, each epoch consists of 10 trials. Note that no error is produced during the Performance stage, since the prememorized mappings can be recalled perfectly after four repetitions during the Instruction period.

instructions given by others, it can also aid in "self-instructed" learning by remembering successful steps learned by trial and error or other means for assembly into new sequences.

6. Planning through Associative Discovery of Bridging States

We explore the idea that the active maintenance of long-term goals in the PFC can work in conjunction with a network's semantic knowledge to identify relevant subgoals and then use those individual subgoals in a similar manner to bias action selection in the present. One fundamental question motivates this research. Given some ultimate goal, possibly associated with explicit reward, how does the system identify subgoals that lead to the final goal? Our hypothesis revolves around semantics, that is, knowledge about how the world works. Our model uses this knowledge to perform constraint satisfaction by using active representations of the current state (where I am) and the desired goal (where I want to be) to associatively arrive at a representation of a subgoal that "bridges" between the two states. This subgoal can serve as the focus for a strategy or plan to achieve the larger goal.

6.1. Description of the Model. There is a tension that exists between the temporal sequencing over one or more subgoals versus a multiple constraint-satisfaction approach that does things all in one step. It seems clear that both can be involved and can be important. So, when does the brain do one versus the other? We have adopted the following heuristic as a kind of corollary of Occam's razor. In general, the brain will by default try to do things in a single time step if it can; as an initial hypothesis, we suspect that bridging over a single subgoal is probably about as much as can be done in this way. When no such plan exists, a more complex process of navigating the modeled task-space through stepwise simulations of intervening states can be undertaken; because this process is among the most complex that humans undertake, a model that does this in a biologically realistic way is a goal for future research. Thus, our initial objective here is to try to demonstrate a one-step constraint satisfaction solution to



(a) Early

(b) Mid

(c) Late

Figure 9: Subgoaling through constraint satisfaction. This figure shows settling of activations of the current state and goal state in both the *Semantic Network* (see text) and *Relation Area* (see text). (a) shows activations early in the settling process of a trial. (b) Activations midway into settling for a trial. The activation of two units in the rightmost Goal layer shows the constraint satisfaction process selecting two plausible subgoals. (c) Activations late in settling when they have converged. The network has settled onto the single most relevant subgoal through constraint satisfaction (simultaneous association from the current state and maintained goal state).

a simple three-state problem: current state and end state to subgoal ("bridging") state.

Another major issue is the tension that exists between state representations sometimes having to compete with one another (e.g., "What is the current state?,") versus sometimes needing to coexist as in spreading activation so as to represent a full motor plan or model of state space (e.g., representing all three of the states in the previous three-state problem). The solution we have settled on is a division of labor between a relation processing area, possibly in the posterior parietal cortex (PPC, circled in red in Figure 9), and a semantic association area, possibly in the anterior temporal lobe (ATL, circled in blue). Because many brain areas are involved in semantics, the precise areas can be expected to vary with the semantic domain, but the mechanisms we describe are expected to be general across those variances. Figure 9 later illustrates these two distinct areas. The PFC (not explicitly modeled) is envisioned to represent the goal state and thus to bias processing in these two areas. The relation processing area is based on the ideas described in "Semantic Cognition" by Rogers and McClelland [70].

Training: the network is trained on the semantics of the State-Action-State triad relation (parietal/anterior temporal cortex) but includes connections to the semantic part of the network. The idea is that the relation area will learn the specific role relations between the states (before, after) and the actions (between states), while the semantic area will learn simple associations between the nodes. The former is dominated by a tight inhibitory competition, while the latter is more free to experience spreading activation. In this way, pre-training on all the individual S-A-S relations enables the bridging over an intermediate subgoal state and biases the correct action in the current state, under the biasing influence of the goal state.

As illustrated in Figure 9(a), which shows a network trained only on pairwise transitions between adjacent states, when a current state and a remote goal state are input, both are activated in both the semantic network and relation engine early in settling. At this very early stage of settling, there are three action units active in the ActionBetween layer (Relation Engine), which are all of the possible actions that can be taken in the current state (S0). Later in settling (Figure 9(b)), a third state unit comes on, which is the intermediate state between the current state and the goal. It becomes the only active unit due to a constraint satisfaction process that includes both bottom-up input from the current state and top-down input from the goal state. This in turn drives the intermediate state unit ON in AfterState layer in the RelationEngine module.

Finally, late in settling (Figure 9(c)), the intermediate state outcompetes the goal unit in the AfterState layer due to the attractor associated with the prior training of contiguous state transitions. This is associated with the third action unit in the ActionBetween and ActionNodes (Semantic Network) layers. This is the correct answer. This model illustrates how constraint satisfaction to find bridging states can work as one component of more complex planning.

6.2. Discussion. Subgoals in this context are conceived as a version of "cold" goals, defined as teleological representations of a desired state of the world that, in and of itself, does not include primary reward. Thus, in a sense, cold goals (here subgoals) are "just a means to an end."

In thinking about the role of subgoals, a number of important issues can be identified. First, as already noted, a fundamental issue concerns how brain mechanisms create useful subgoals, if they are not provided externally. In addition, a second important issue is whether there are one or more biologically plausible mechanisms for rewarding the achievement of subgoals. This in turn has two subcomponents: (1) learning how to achieve subgoals in the first place (e.g., how to grind coffee in support of making coffee in the morning) and (2) learning how/when to exploit already familiar subgoal in the service of achieving a master goal (e.g., learning that having ground coffee is a precursor to enjoying a nice fresh cup of hot coffee for yourself and/or receiving kudos from your significant other). It is interesting to note that these two learning categories exhibit a mutual interdependence. Usually, learning how to achieve subgoals must precede learning to exploit them, although an interesting alternative can sometimes occur: if a learner is allowed to use its what-if imagination. For example, if a learner can do thought experiments like: "IF I had ground coffee, and cold water, and a working coffee maker, THEN I could have hot coffee." Thinking about it over and over could transfer (imagined) value from the hot coffee to the ground coffee, and so forth, *which then* could be used as secondary reinforcement to motivate the learning of instrumental subgoals. This scenario-spinning behavior is not modeled in any realistic cognitive model of which we are aware; achieving this will be difficult but an important step toward understanding human intelligence.

A third critical issue is how subgoals are actually used by the system (in a mechanistic sense) in the service of pursuing the master goal. Here, the simple idea that serves as a kind of working hypothesis in our work is that the active maintenance of subgoals can serve to bias the behavior that produces their realization in a kind teleological "pull of the future" way. Finally, there then still needs to be some sort of cognitive sequencing control mechanism organizing the overall process, that is, the achievement of each subgoal in turn. Ultimately, in our way of thinking, this whole process can be biased by keeping the master goal in active maintenance throughout the process.

In sum, this model demonstrates a rough draft of one aspect of human high-level planning: abstract state representations allow constraint satisfaction processes based on associative learning to find a bridging state between current and goal states. We hypothesize that this process is iterated at different levels of abstraction to create more detailed plans as they are needed. However, we do not as yet have a model that includes the movement between different levels of plan abstraction. The other models presented here represent some of the mechanisms needed for this process but have yet to be integrated into a coherent, let alone complete, model of human planning.

Explaining how brains perform planning requires understanding the computational demands involved. The more abstract literature on the algorithmic and computational properties of planning in artificial intelligence research has thoroughly explored the problem space of many types of planning (e.g., [71–73]). Consistent with this proposed biological model, algorithmic constraint satisfaction solvers have been an important part of AI planning algorithms (e.g., [74, 75]). Other extensions and combinations of these models are also suggested by AI planning work; search-based algorithms (e.g., [76, 77]) show that sequencing, storing, and retrieval of state (as in the model-free and instructed sequencing model) are essential for flexible planning. We address some such possible combinations and extensions later.

7. General Discussion

The four models here represent an incomplete start at fully modeling human strategic cognitive sequencing. A full model would explain how basic mammalian brain mechanisms can account for the remarkable complexity and flexibility of human cognition. It would address the use of elaborate cognitive sequences which constitute learned "programs" for solving complex problems and how people generalize this ability to new problems by selecting parts of these sequences to construct appropriate strategies for novel tasks in related domains. A complete model is thus a long-term and ambitious project, but one with important implications for understanding human cognition.

The following primarily addresses the limitations in the work described and our plans to extend these models toward a more complete explanation of complex human sequential cognition. Although learning was integral to all presented models, demonstrating the feasibility of bootstrapping such flexible cognitive systems, the learning in these initial models was mostly still domain specific: models were trained within the class of tasks to be performed from a naive state. While the instructed model could generalise to a variety of unseen if-then rules and the constraint satisfaction model generalizes to unseen state-goal pairings, they were both only trained on their respective tasks.

In future work, we plan to extend this to a more sophisticated pre-training or scaffolding of networks that are more general and ecologically valid. Instead of beginning training of specific task structures from a naive network, the idea is to train the networks on a large variety of distinct tasks, progressing from simple to complex. The PBDM model, for instance, was trained in a relatively ecologically valid way but did not learn increasing complexity of tasks as it mastered simple ones as humans do. With increasing number of tasks trained, the network should learn to extract commonality between tasks, abstracting the essence of tasks into distinct representations. While it remains unclear what these task representations might look like on the finer biological scale, either from experimentation or computational modeling, it seems likely that representations for some basic computational building blocks of cognitive sequencing exist.

Such representations must, at an abstract level, include some of those found in any standard computer programming language, such as sequencing, loops, storing, and recalling of state. While the models presented here cannot accomplish any of these functions as they stand, we already have a rough basis for these basic task building blocks. All of the previous "program flow" functions can be seen as subsets of conditional branching (e.g., if you have not yet found the goal object, use a sequence that looks for it). The other models presented here (planning, model-free sequence learning, and decision making) address important aspects of how sequences are learned and used, but the instructed learning model alone is enough to understand one way in which the brain can exhibit such program flow control once a relevant sequence is learned. This behavior requires extending the model to store and use state information. This minor extension would include working memory updates in the potential actions and make action pairs conditional on those working memory representations as well as sensory inputs.

Dayan [78] has already explored this behavior in a more abstract version of PBWM. This model includes storage actions and dependency upon stored information consistent with the role for which PBWM was primarily developed, understanding how mechanisms evolved for gating motor actions control storage in working memory. Making memorized pairings dependent upon state information in working memory is also straightforward, and known basal ganglia connectivity suggests such a convergence of information between prefrontal working memory and posterior sensory cortices for the purpose of gating decisions. Dayan [78] also includes a match detection function to allow nonmatch criteria that do not arise naturally from the associative nature of neural networks, an important consideration for our future development of these models.

The models presented here are also generally consistent with the most well-developed models in this domain, procedural models such as ACT-R [60], from which our approach draws inspiration. While our work is generally compatible, we hope to provide more constraints on these theories by considering the wealth of data on detailed aspects of neural function.

In particular, our learning neural network approach will also allow us to constrain theories of exactly what representations are used to produce cognitive sequences by how they are learned. By studying learning over a large number of tasks, we aim to address the question of how these representations emerge on a developmental time scale from a young infant to the fully developed capability of an adult. This focus addresses *learning to learn*, a phenomenon that has both been extensively studied in psychology as well as in machine learning and robotics [79–81]. In both cases, learning to learn transfers beneficial information from a group of tasks to new ones, speeding up learning of new tasks. While in machine learning, many different algorithms have been proposed to achieve transfer learning or learning to learn, a good proportion is based upon representational transfer [79, 82]; that is, due to the efficient and general representations learned in prior tasks, new tasks can be learned more rapidly or more effectively instructed.

To address these questions, we will draw on our and others' work on learning of abstract categories from sensory data (e.g., [83]). Generalizing from prior learning usefully categorizes novel sensory inputs through neural processing that is now relatively well understood. Such category generalization, when combined with the models presented here, offers one explanation of learning to learn. When strategic cognitive sequencing is performed based upon categorical representations (e.g., substitute "input A" in the instructed learning model for "signal to stop and wait for instructions"), learning will generalize to new sensory inputs that can be correctly categorized. This type of generalized matching bears a resemblance to the variable matching rule in recent versions of ACT-R (e.g., "if the word was (word X, previously stored), press the red button"). Modeling this process in greater neural detail will provide more constraints on what types of generalization and matching can be learned and performed by realistic neural networks.

Perhaps because such high-level cognition inevitably involves interactions between many brain regions, computational modeling and other forms of detailed theory construction have, as yet, made little progress. However, the enormous accumulation of work aimed at understanding the contributions from individual brain areas have rendered this complex but important domain a potentially productive target for detailed modeling and computational-level theory.

Acknowledgments

The authors thank members of the CCN Lab at CU Boulder for helpful comments and discussion. The paper is supported by ONR N00014-10-1-0177, NIH MH079485, and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract no. D10PC20021.

The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US Government.

References

[1] A. M. Owen, "Tuning in to the temporal dynamics of brain activation using functional magnetic resonance imaging (fMRI)," *Trends in Cognitive Sciences*, vol. 1, no. 4, pp. 123–125, 1997.

[2] T. Shallice, "Specific impairments of planning," *Philosophical transactions of the Royal Society of London B*, vol. 298, no. 1089, pp. 199–209, 1982.

[3] L. Roy Beach and T. R. Mitchell, "A contingency model for the selection of decision strategies," *The Academy of Management Review*, vol. 3, no. 3, pp. 439–449, 1978.

[4] P. Slovic, B. Fischhoff, and S. Lichtenstein, "Behavioral decision theory," *Annual Review of Psychology*, vol. 28, pp. 1–39, 1977.

[5] M. Chi and K. VanLehn, "Meta-cognitive strategy instruction in intelligent tutoring systems: how, when, and why," *Educational Technology & Society*, vol. 13, no. 1, pp. 25–39, 2010.

[6] J. M. Unterrainer, B. Rahm, R. Leonhart, C. C. Ruff, and U. Halsband, "The tower of London: the impact of instructions, cueing, and learning on planning abilities," *Cognitive Brain Research*, vol. 17, no. 3, pp. 675–683, 2003.

[7] A. Newell, "You can't play 20 questions with nature and win: projective comments on the papers of this symposium," in *Visual Information Processing*, W. G. Chase, Ed., pp. 283–308, Academic Press, New York, NY, USA, 1973.

[8] L. B. Smith, "A model of perceptual classification in children and adults," *Psychological Review*, vol. 96, no. 1, pp. 125–144, 1989.

[9] M. J. Roberts and E. J. Newton, "Understanding strategy selection," *International Journal of Human Computer Studies*, vol. 54, no. 1, pp. 137–154, 2001.

[10] J. Tanji and E. Hoshi, "Role of the lateral prefrontal cortex in executive behavioral control," *Physiological Reviews*, vol. 88, no. 1, pp. 37–57, 2008.

[11] A. Dagher, A. M. Owen, H. Boecker, and D. J. Brooks, "Mapping the network for planning: a correlational PET activation study with the tower of London task," *Brain*, vol. 122, no. 10, pp. 1973–1987, 1999.

[12] O. A. van den Heuvel, H. J. Groenewegen, F. Barkhof, R. H. C. Lazeron, R. van Dyck, and D. J. Veltman, "Frontostriatal system in planning complexity: a parametric functional magnetic resonance version of Tower of London task," *NeuroImage*, vol. 18, no. 2, pp. 367–374, 2003.

[13] A. Dagher, A. M. Owen, H. Boecker, and D. J. Brooks, "The role of the striatum and hippocampus in planning: a PET activation study in Parkinson's disease," *Brain*, vol. 124, no. 5, pp. 1020–1032, 2001.

[14] K. Shima, M. Isoda, H. Mushiake, and J. Tanji, "Categorization of behavioural sequences in the prefrontal cortex," *Nature*, vol. 445, no. 7125, pp. 315–318, 2007.

[15] R. C. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind By Simulating the Brain*, The MIT Press, Cambridge, Mass, USA, 2000.

[16] R. C. O'Reilly, T. E. Hazy, and S. A. Herd, "The leabra cognitive architecture: how to play 20 principles with nature and win!," in *The Oxford Handbook of Cognitive Science*, S. Chipman, Ed., Oxford University Press, In press.

[17] A. Newell and H. A. Simon, *Human Problem Solving*, Prentice Hall, Englewood Cliffs, NJ, USA, 1972.

[18] J. R. Anderson, *Rules of the Mind*, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1993.

[19] R. Morris and G. Ward, *The Cognitive Psychology of Planning*, Psychology Press, 2005.

[20] C. Lebiere, J. R. Anderson, and D. Bothell, "Multi-tasking and cognitive workload in an act-r model of a simplified air traffic control task," in *Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation*, 2001.

[21] T. Suddendorf and M. C. Corballis, "Behavioural evidence for mental time travel in nonhuman animals," *Behavioural Brain Research*, vol. 215, no. 2, pp. 292–298, 2010.

[22] S. J. Shettleworth, "Clever animals and killjoy explanations in comparative psychology," *Trends in Cognitive Sciences*, vol. 14, no. 11, pp. 477–481, 2010.

[23] D. Klahr, P. Langley, and R. Neches, Eds., *Production System Models of Learning and Development*, The MIT Press, Cambridge, Mass, USA, 1987.

[24] D. J. Jilk, C. Lebiere, R. C. O'Reilly, and J. R. Anderson, "SAL: an explicitly pluralistic cognitive architecture," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 20, no. 3, pp. 197–218, 2008.

[25] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–1060, 2004.

[26] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, New York, NY, USA, 2007.

[27] H. H. Yin and B. J. Knowlton, "The role of the basal ganglia in habit formation," *Nature Reviews Neuroscience*, vol. 7, no. 6, pp. 464–476, 2006.

[28] M. J. Frank, B. Loughry, and R. C. O'Reilly, "Interactions between frontal cortex and basal ganglia in working memory: a computational model," *Cognitive, Affective and Behavioral Neuroscience*, vol. 1, no. 2, pp. 137–160, 2001.

[29] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly, "By carrot or by stick: cognitive reinforcement learning in Parkinsonism," *Science*, vol. 306, no. 5703, pp. 1940–1943, 2004.

[30] T. E. Hazy, M. J. Frank, and R. C. O'Reilly, "Banishing the homunculus: making working memory work," *Neuroscience*, vol. 139, no. 1, pp. 105–118, 2006.

[31] T. E. Hazy, M. J. Frank, and R. C. O'Reilly, "Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system," *Philosophical Transactions of the Royal Society B*, vol. 362, no. 1485, pp. 1601–1613, 2007.

[32] R. C. O'Reilly and M. J. Frank, "Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia," *Neural Computation*, vol. 18, no. 2, pp. 283–328, 2006.

[33] K. Sakai, "Task set and prefrontal cortex," *Annual Review of Neuroscience*, vol. 31, pp. 219–245, 2008.

[34] G. E. Alexander, M. R. DeLong, and P. L. Strick, "Parallel organization of functionally segregated circuits linking basal ganglia and cortex," *Annual Review of Neuroscience*, vol. 9, pp. 357–381, 1986.

[35] M. J. Frank, "Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism," *Journal of Cognitive Neuroscience*, vol. 17, no. 1, pp. 51–72, 2005.

[36] R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz, "PVLV: the primary value and learned value Pavlovian learning algorithm," *Behavioral Neuroscience*, vol. 121, no. 1, pp. 31–49, 2007.

[37] T. E. Hazy, M. J. Frank, and R. C. O'Reilly, "Neural mechanisms of acquired phasic dopamine responses in learning," *Neuroscience and Biobehavioral Reviews*, vol. 34, no. 5, pp. 701–720, 2010.

[38] J. M. Fuster and A. A. Uyeda, "Reactivity of limbic neurons of the monkey to appetitive and aversive signals," *Electroencephalography and Clinical Neurophysiology*, vol. 30, no. 4, pp. 281–293, 1971.

[39] E. K. Miller, "The prefrontal cortex and cognitive control," *Nature Reviews Neuroscience*, vol. 1, no. 1, pp. 59–65, 2000.

[40] T. Ono, K. Nakamura, H. Nishijo, and M. Fukuda, "Hypothalamic neuron involvement in integration of reward, aversion, and cue signals," *Journal of Neurophysiology*, vol. 56, no. 1, pp. 63–79, 1986.

[41] S. A. Deadwyler, S. Hayashizaki, J. Cheer, and R. E. Hampson, "Reward, memory and substance abuse: functional neuronal circuits in the nucleus accumbens," *Neuroscience and Biobehavioral Reviews*, vol. 27, no. 8, pp. 703–711, 2004.

[42] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[43] R. S. Sutton and A. G. Barto, "Time-derivative models of pavlovian reinforcement," in *Learning and Computational Neuroscience*, J. W. Moore and M. Gabriel, Eds., pp. 497–537, MIT Press, Cambridge, Mass, USA, 1990.

[44] J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan, "Temporal difference models and reward-related learning in the human brain," *Neuron*, vol. 38, no. 2, pp. 329–337, 2003.

[45] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Mass, USA, 1998.

[46] R. Stuart Sutton, *Temporal credit assignment in reinforcement learning [Ph.D. thesis]*, University of Massachusetts Amherst, Amherst, Mass, USA, 1984.

[47] P. Dayan and B. W. Balleine, "Reward, motivation, and reinforcement learning," *Neuron*, vol. 36, no. 2, pp. 285–298, 2002.

[48] R. C. O'Reilly, S. A. Herd, and W. M. Pauli, "Computational models of cognitive control," *Current Opinion in Neurobiology*, vol. 20, no. 2, pp. 257–261, 2010.

[49] C. H. Chatham, S. A. Herd, A. M. Brant et al., "From an executive network to executive control: a computational model of the n-back task," *Journal of Cognitive Neuroscience*, vol. 23, no. 11, pp. 3598–3619, 2011.

[50] D. Joel, Y. Niv, and E. Ruppin, "Actor-critic models of the basal ganglia: new anatomical and computational perspectives," *Neural Networks*, vol. 15, no. 4-6, pp. 535–547, 2002.

[51] W.-T. Fu and J. R. Anderson, "Solving the credit assignment problem: explicit and implicit learning of action sequences with probabilistic outcomes," *Psychological Research*, vol. 72, no. 3, pp. 321–330, 2008.

[52] J. D. Wallis, "Orbitofrontal cortex and its contribution to decision-making," *Annual Review of Neuroscience*, vol. 30, pp. 31–56, 2007.

[53] M. P. Noonan, N. Kolling, M. E. Walton, and M. F. S. Rushworth, "Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement," *European Journal of Neuroscience*, vol. 35, no. 7, pp. 997–1010, 2012.

[54] P. L. Croxson, M. E. Walton, J. X. O'Reilly, T. E. J. Behrens, and M. F. S. Rushworth, "Effort-based cost-benefit valuation and the human brain," *The Journal of Neuroscience*, vol. 29, no. 14, pp. 4531–4541, 2009.

[55] S. W. Kennerley and M. E. Walton, "Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies," *Behavioral Neuroscience*, vol. 125, no. 3, pp. 297–317, 2011.

[56] M. J. Frank and E. D. Claus, "Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal," *Psychological Review*, vol. 113, no. 2, pp. 300–326, 2006.

[57] J. M. Hyman, L. Ma, E. Balaguer-Ballester, D. Durstewitz, and J. K. Seamans, "Contextual encoding by ensembles of medial prefrontal cortex neurons," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 13, pp. 5086–5091, 2012.

[58] J. J. Day, J. L. Jones, R. M. Wightman, and R. M. Carelli, "Phasic nucleus accumbens dopamine release encodes effort- and delay-related costs," *Biological Psychiatry*, vol. 68, no. 3, pp. 306–309, 2010.

[59] J. Duncan, M. Schramm, R. Thompson, and I. Dumontheil, "Task rules, working memory, and fluid intelligence," *Psychonomic Bulletin & Review*, vol. 19, no. 5, pp. 864–8870, 2012.

[60] J. R. Anderson, *The Architecture of Cognition*, Harvard University Press, Cambridge, Mass, USA, 1983.

[61] P. M. Fitts and M. I. Posner, *Human Performance*, Belmont, Mass, USA, 1967.

[62] P. Redgrave and K. Gurney, "The short-latency dopamine signal: a role in discovering novel actions?" *Nature Reviews Neuroscience*, vol. 7, no. 12, pp. 967–975, 2006.

[63] G. Fernández, H. Weyerts, M. Schrader-Bölsche et al., "Successful verbal encoding into episodic memory engages the posterior hippocampus: a parametrically analyzed functional magnetic resonance imaging study," *The Journal of Neuroscience*, vol. 18, no. 5, pp. 1841–1847, 1998.

[64] D. Kumaran, J. J. Summerfield, D. Hassabis, and E. A. Maguire, "Tracking the emergence of conceptual knowledge during human decision making," *Neuron*, vol. 63, no. 6, pp. 889–901, 2009.

[65] D. C. Noelle and G. W. Cottrell, "A connectionist model of instruction following, pages," in *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, J. D. Moore and J. F. Lehman, Eds., pp. 369–374, Lawrence Erlbaum Associates, Mahwah, NJ, USA, January 1995.

[66] G. Biele, J. Rieskamp, and R. Gonzalez, "Computational models for the combination of advice and individual learning," *Cognitive Science*, vol. 33, no. 2, pp. 206–242, 2009.

[67] B. B. Doll, W. J. Jacobs, A. G. Sanfey, and M. J. Frank, "Instructional control of reinforcement learning: a behavioral and neurocomputational investigation," *Brain Research*, vol. 1299, pp. 74–94, 2009.

[68] J. Li, M. R. Delgado, and E. A. Phelps, "How instructed knowledge modulates the neural systems of reward learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 1, pp. 55–60, 2011.

[69] M. M. Walsh and J. R. Anderson, "Modulation of the feedback-related negativity by instruction and experience," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 47, pp. 19048–19053, 2011.

[70] T. T. Rogers and J. L. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach*, MIT Press, Cambridge, Mass, USA, 2004.

[71] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

[72] F. Bacchus, "AIPS '00 planning competition: the fifth international conference on Artificial Intelligence Planning and Scheduling systems," *AI Magazine*, vol. 22, no. 3, pp. 47–56, 2001.

[73] E. D. Sacerdoti, "Planning in a hierarchy of abstraction spaces," *Artificial Intelligence*, vol. 5, no. 2, pp. 115–135, 1974.

[74] M. B. Do and S. Kambhampati, "Planning as constraint satisfaction: solving the planning graph by compiling it into CSP," *Artificial Intelligence*, vol. 132, no. 2, pp. 151–182, 2001.

[75] P. Gregory, D. Long, and M. Fox, "Constraint based planning with composable substate graphs," in *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI '10)*, H. Coelho, R. Studer, and M. Wooldridge, Eds., IOS Press, 2010.

[76] A. L. Blum and M. L. Furst, "Fast planning through planning graph analysis," *Artificial Intelligence*, vol. 90, no. 1-2, pp. 281–300, 1997.

[77] E. Fink and M. M. Veloso, "Formalizing the prodigy planning algorithm," Tech. Rep. 1-1-1996, 1996.

[78] P. Dayan, "Bilinearity, rules, and prefrontal cortex," *Frontiers in Computational Neuroscience*, vol. 1, no. 1, pp. 1–14, 2007.

[79] S. Thrun and L. Pratt, "Learning to learn: introduction and overview," in *Learning To Learn*, S. Thrun and L. Pratt, Eds., Springer, New York, NY, USA, 1998.

[80] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine Learning*, vol. 28, no. 1, pp. 7–39, 1997.

[81] G. Konidaris and A. Barto, "Building portable options: Skill tran sfer in reinforcement learning," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, M. M. Veloso, Ed., pp. 895–900, 2006.

[82] K. Ferguson and S. Mahadevan, "Proto-transfer learning in markov decision processes using spectral methods," in *Proceedings of the Workshop on Structural Knowledge Transfer for Machine Learning (ICML '06)*, 2006.

[83] R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk, "Recurrent processing during object recognition," *Frontiers in Psychology*, vol. 4, article 124, 2013.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 308176, 10 pages
http://dx.doi.org/10.1155/2013/308176

Research Article

Augmenting Weak Semantic Cognitive Maps with an "Abstractness" Dimension

Alexei V. Samsonovich and Giorgio A. Ascoli

Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive MS 2A1, Fairfax, VA 22030-4444, USA

Correspondence should be addressed to Alexei V. Samsonovich; asamsono@gmu.edu

Received 14 December 2012; Accepted 29 April 2013

Academic Editor: Rajan Bhattacharyya

The emergent consensus on dimensional models of sentiment, appraisal, emotions, and values is on the semantics of the principal dimensions, typically interpreted as valence, arousal, and dominance. The notion of weak semantic maps was introduced recently as distribution of representations in abstract spaces that are not derived from human judgments, psychometrics, or any other a priori information about their semantics. Instead, they are defined entirely by binary semantic relations among representations, such as synonymy and antonymy. An interesting question concerns the ability of the antonymy-based semantic maps to capture all "universal" semantic dimensions. The present work shows that those narrow weak semantic maps are not complete in this sense and can be augmented with other semantic relations. Specifically, including hyponym-hypernym relations yields a new semantic dimension of the map labeled here "abstractness" (or ontological generality) that is not reducible to any dimensions represented by antonym pairs or to traditional affective space dimensions. It is expected that including other semantic relations (e.g., meronymy/holonymy) will also result in the addition of new semantic dimensions to the map. These findings have broad implications for automated quantitative evaluation of the meaning of text and may shed light on the nature of human subjective experience.

1. Introduction

The idea of representing semantics geometrically is increasingly popular. Many mainstream approaches use vector space models, in which concepts, words, documents, and so forth are associated with vectors in an abstract multidimensional vector space. Other approaches use manifolds of more complex topology and geometry. In either case, the resultant space or manifold together with its allocated representations is called a *semantic space* or a *semantic (cognitive) map*. Examples include spaces constructed with Latent Semantic Analysis (LSA) [1] and Latent Dirichlet Allocation (LDA) [2], as well as many related techniques, for example, ConceptNet [3, 4]. Other examples of techniques include Multi-Dimensional Scaling (MDS) [5], including Isomap [6], and related manifold-learning techniques [7], Gardenfors' conceptual spaces [8], very popular in the past models of self-organizing feature maps, and more.

The majority of these approaches are based on the idea of a dissimilarity metrics, which is to capture semantic dissimilarity between representations (words, documents, concepts, etc.) with a geometrical distance between associated space elements (points or vectors). In other words, the metrics that determines the allocation of representations in space is a function of their semantic dissimilarity. In this case, two representations allocated at close points in space must have similar semantics and vice versa: two representations with similar semantics must be close to each other in space. Conversely, representations unrelated to each other must be separated by significant distance.

We introduced the term "weak semantic cognitive mapping" to denote an alternative approach, exploited here, which is not based on dissimilarity [9–11]. The idea is not to separate all different meanings from each other (like in MDS), nor to allocate them based on their individual semantic characteristics given a priori (as in LSA), but rather to arrange them in space based on their mutual semantic relations. The notion of weak semantic cognitive maps was originally introduced in a narrow sense, where these relations were limited to synonymy and antonymy only [9–11]. In a more general sense, as discussed below, weak semantic cognitive maps may capture other binary semantic relations as well, including hypernymy-hyponymy, holonymy-meronymy, troponymy, causality, and dependence.

While the understanding of dissimilarity as the basis of antonymy is widespread, many examples of the dictionary antonym pairs used in our analysis suggest that dissimilarity and antonymy are distinct notions. Most unrelated words may be considered dissimilar (e.g., "apple" and "inequality"), yet do not constitute antonym pairs. In contrast, antonym pairs include words that are related to each other and in a certain sense are similar to each other in their meaning and usage, for example, king and queen, major and minor, and ascent and descent. It appears that most antonym pairs (at least in the dictionaries that we used) are consistent with the notion of "opposite" rather than "dissimilar."

More generally, the method of weak semantic mapping is essentially different from most vector-space-based approaches including LSA, LDA, MDS, and ConceptNet [1–4], primarily because there is no a priori attribution of semantic features to representations in the constructive definition of the map. Only relations, but not semantic features, are given as input. As a result, semantic dimensions of the map that are not predefined to emerge naturally, starting from a randomly generated initial distribution of words in an abstract space with no a priori given semantics and following the strategy to pull synonyms together and antonyms apart [10, 11] (see Section 2: Methods). In contrast to LSA, principal component analysis is used here to reveal the main emergent semantic dimensions at the final stage only. The advantage of the antonymy-based weak semantic cognitive map compared to "strong" maps based on dissimilarity metrics is that its dimensions have clearly identifiable semantics (naturally given by the corresponding pairs of antonyms) that are domain-independent. For example, the notion of "good versus bad" that corresponds to the first principal component applies to all domains of human knowledge.

Interestingly, semantics of the emergent dimensions of antonym-based weak semantic cognitive maps are closely related to those of another broad category of "dimensional models" of affects [12] that attempt to capture human emotions, feelings, affects, appraisals, sentiments, and attitudes. Examples range from original classical models such as Osgood's semantic differential [13], Russell's circumplex [14], and Plutchik's wheel [15] to many more recent derivative integrated frameworks, like PAD (pleasure, arousal, and dominance) [16], ANEW (Affective Norms for English Words) [17], EPA (evaluation, potency, and arousal) [18], and a recent 3D model linking emotions to main neurotransmitters [19]. These dimensional models are usually derived from human experimental studies involving psychometrics or introspective judgment evaluated on the Likert scale [20]. While these models provide the most common bases for opinion mining or sentiment analysis [21], the weak semantic map is more complete in the sense that (i) it assigns values to all words, not only to emotionally meaningful words, (ii) it measures semantics associated with all antonym pairs, not only emotionally meaningful antonym pairs, and therefore is applicable to all domains of knowledge, and (iii) its dimensions are orthogonal and independent of each other. The combination of these features makes weak semantic maps extremely valuable for numerous applications.

It is surprising that the well-known dimensions of the semantic differential, PAD, EPA, and related models can be recognized in the main principal components (PC) of the above cited weak semantic map, where PC1 is related to valence, PC2 to arousal, and PC3 to dominance [11]. (This correspondence is approximate, because the principal components have zero correlations with each other, while the variables of, e.g., ANEW are strongly correlated.) For example, "love" and "joy" have top values of valence in the affective database ANEW and also top values of PC1 of weak semantic cognitive map. Words like "anger" and "excitement" have top values of arousal in the affective database ANEW and also top values of PC2 in weak semantic cognitive map. This correspondence is consistent in weak semantic maps constructed based on different corpora in several major languages [11]. The observation is unexpected, because the weak semantic map is not derived from any semantic features of words given a priori, and is not explicitly related to emotions and feelings by its construction. In fact, any pair of antonyms defines a map dimension, including antonym pairs that are not associated with affects, for example, "abstract-specific." It is also surprising that the weak semantic map is low-dimensional: the number of PCs that account for 95% of the variance of the multidimensional distribution typically varies from 4 to 6, depending on the corpus [11].

How complete is the weak semantic map narrowly defined only by antonym pairs? Certainly at least some semantic differences cannot be captured by antonymy relations, because not all concepts have antonyms (e.g., the number 921714083). Here we address a different question: whether all universal semantic dimensions can be captured by antonymy relations. For example, it may seem obvious that causality cannot be captured by antonymy. However, the issue is nontrivial, as there are many examples of causally related antonyms (e.g., attack-defend, begin-end, send-receive, and even cause-effect). Thus, two logical possibilities stand.

(1) Antonym-based semantic maps separate representations along all semantic dimensions that make sense for all domains of knowledge. Thus, if there is a semantic characteristic X that makes sense for all domains of knowledge such that some concepts can be characterized as having more X than others, then there is a direction on the narrow weak semantic map along which those concepts are separated based on their value of X.

(2) The alternative: there is at least one general semantic characteristic X defined for all domains that is ignored by the antonym-based weak semantic map. In other words, the variance in X measured across all concepts is not accounted by the map coordinates of concepts, and vice versa, no significant part of the variance of the map can be accounted by X.

Here we argue for (2), quantifying the notion of "abstractness" (or ontological generality) as an example of X. Our

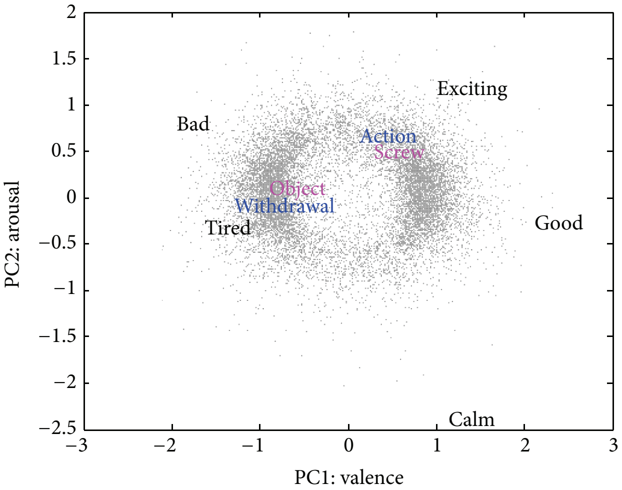


Figure 1: A sample from the antonymy-based weak semantic cognitive map constructed by Samsonovich and Ascoli [11]. Grey dots show all 15,783 words from the MS Word English dictionary. Similar results were obtained by WordNet. Words shown in color are examples of hypernym-hyponym pairs: “action-withdrawal” and “object-screw.” Selected examples illustrate that there is no clear separation of hypernyms and hyponyms on the map.

technical definition of “abstractness” is based on hyponym-hypernym relations among words.

Before presenting results of the computational study, we briefly discuss the hypothesis at an intuitive level. While “abstract-specific” is a pair of antonyms, which corresponds to a direction on the narrow weak semantic map, the two antonyms “abstract” and “specific” themselves have approximately the same measure of “abstractness” (the X value) associated with them. Intuitively, this observation must hold for most antonym pairs, because antonyms pairs do not typically constitute a hypernym-hyponym couple. Therefore, it is unlikely that there is a hyperplane on the map that separates more abstract from more specific words. Therefore, we do not expect to find a dimension of the map based on synonyms and antonyms that could separate words by “abstractness” (see Figure 1). In contrast, there is a hyperplane (PC1 = 0) that separates “good” and “bad” words and a hyperplane (PC2 = 0) that separates “calming” and “exciting” words. That is to say, “good words” tend to be synonyms of the word “good,” but “abstract words” are not synonyms of the word “abstract” or of each other.

2. Methods

2.1. Weak Semantic Cognitive Mapping. The general idea of semantic cognitive mapping is to allocate representations (e.g., words) in an abstract space based on their semantics. This paradigm is common for a large number of techniques overviewed in Introduction. While most studies in semantic cognitive mapping are based on the notion of a dissimilarity metrics and/or on a set of semantic features given a priori, weak semantic mapping ignores dissimilarity as well as any individually predefined semantics.

The algorithm for antonymy-based weak semantic mapping is described in our previous work [11]. The semantic space is created by minimization of the “energy” of the entire distribution of words on the map, starting from a random distribution. Then, the emergent semantics of the map dimensions are defined by the entire distribution of representations on the map and typically are best characterized by the pairs of antonyms that are separated by the greatest distance along the given dimension. The main semantic dimensions are defined by the principal components of the emergent distribution of words on the map. Semantics associated with the first three PCs can be characterized as “good” versus “bad” (PC1), “calming, easy” versus “exciting, hard” (PC2), and “free, open” versus “dominated, closed” (PC3) [11]. When limited to affects, these semantics approximately correspond to the three PAD dimensions: pleasure, arousal, and dominance.

More precisely, the narrow weak semantic cognitive map is a distribution of words in an abstract vector space (with no semantics preassociated with its elements or dimensions) that minimizes the following energy function [11]:

$$H(\mathbf{x}) = -\frac{1}{2}\sum_{i,j=1}^{N} W_{ij}\mathbf{x}_i \cdot \mathbf{x}_j + \frac{1}{4}\sum_{i=1}^{N}\left|\mathbf{x}_i\right|^4, \quad \mathbf{x} \in \Re^N \otimes \Re^D. \quad (1)$$

Here $\mathbf{x}_i$ is a D-vector representing the ith word (out of N). The W_{ij} entries of the symmetric relation matrix equal +1 for pairs of synonyms, −1 for pairs of antonyms, and zero otherwise. D is set to any integer (e.g., 100) that is substantially greater than the number of resulting significant principal components of the distribution, which typically ranges from 4 to 6 and determines the dimensionality of the map. In this case the choice of D does not change the outcome. The energy function (1) follows the principle of parsimony: it is the simplest analytical expression that creates balanced forces of desired signs between synonyms and antonyms, preserves symmetries of semantic relations, and increases indefinitely at the infinity, keeping the resultant distribution localized near the origin of coordinates.

The procedure is that the initial coordinates of all words are sampled by a random number generator. Then the energy (1) minimization process starts that pulls synonym vectors together and antonym vectors apart. Then principal component analysis is used to reveal the main emergent semantic dimensions of the optimized map [10, 11]. Thus, the initial space coordinates are not associated with any semantics a priori: instead, words are allocated randomly in an abstract multidimensional space. In contrast, the starting point of traditional techniques based on LSA [1, 22] is a feature space, where dimensions have definite semantics a priori.

The representative weak semantic map shown in Figure 1 includes $N = 15{,}783$ words and was constructed based on the dictionary of English synonyms and antonyms available as part of Microsoft Word (MS Word) [11]. A similar map was also constructed using WordNet in the same work [11] and is also used in this study, together with maps constructed in [11] for other languages. Figure 1 represents the first two PCs of the distribution of words on the map constructed using the English MS Word thesaurus. The axes of the map are defined by the PCs. Selected words shown on the map in black at

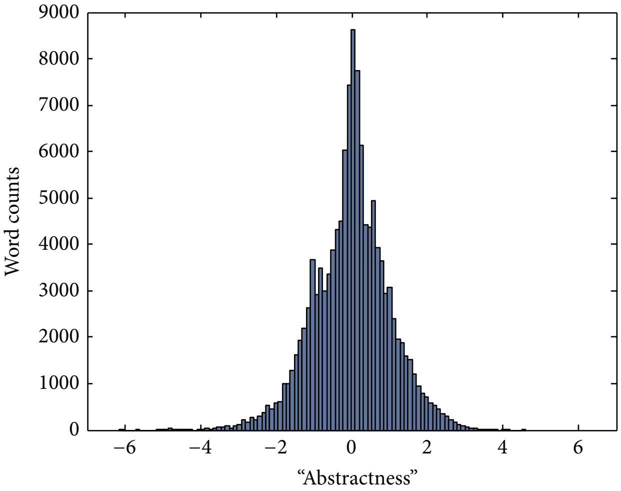


Figure 2: Distribution histogram of the 124,408 WordNet 3.0 words along the "abstractness" dimension.

Table 1: The tails of the list of 124,408 words sorted by "abstractness" in descending order.

The beginning of the list	The end of the list
Entity	Chain wrench
Physical entity	Francis turbine
Psychological feature	Tricolor television tube
Auditory communication	Tricolor tube
Unmake	Tricolour television tube
Cognition	Tricolour tube
Knowledge	Edmontonia
Noesis	*Coelophysis*
Natural phenomenon	*Deinocheirus*
Ability	*Struthiomimus*
Social event	*Deinonychus*
Craniate	Dromaeosaur
Vertebrate	*Mononychus olecranus*
Higher cognitive process	*Oviraptorid*
Physiological property	*Superslasher*
Mammal	*Utahraptor*
Mammalian	*Velociraptor*

their map locations characterize the semantics of the map. The two hypernym-hyponym pairs, "object-screw" (shown in pink) and "action-withdrawal" (in blue), illustrate the map inability to capture the "abstractness" dimension, confirmed quantitatively by correlation analysis in the next section. It should be pointed out here that the negative valence of "object" can be attributed to the meaning of the verb "object" that is merged with the noun "object" on this string-based semantic map.

2.2. Measuring the "Abstractness" of Words. Here we refer to the "abstractness" of a concept as its ontological generality. The WordNet database contains information that allows us to arrange English words on a line according to their "abstractness" (or ontological generality). This information is contained in the hyponym-hypernym relations among words. The goal is to separate hypernym-hyponym pairs in one dimension tentatively labeled "abstractness," so that each hyponym has a lower "abstractness" value compared to its hypernyms. Given a consistent hierarchy, a solution would be, for example, to interpret the order of a word in the hierarchy as a measure of its "abstractness." Unfortunately, the system of hyponym-hypernym relations among words available in WordNet is internally inconsistent: it has numerous loops and conflicting links. Therefore, we use an optimization approach analogous to the antonymy-based weak semantic mapping based on (1). The underlying idea is to give each word i its "abstractness" coordinate x_i in such a way that the overall correlation between the difference in word "abstractness" coordinates x and the reciprocal hypernym-hyponym relations of the two words is maximized. Unfortunately, an energy function similar to H (1) cannot be used here, because the symmetry of hypernym-hyponym relations is different from the symmetry of antonym and synonym relations. Nevertheless, we showed in previous work [23] that the goal can be achieved by using the following definition of word "abstractness" values $\{x\}$:

$$\vec{x} = \underset{\mathbb{R}^n}{\operatorname{argmin}} \left[\sum_{i,j=1}^{n} W_{ij} \left(x_i - x_j - 1 \right)^2 + \mu \sum_{i=1}^{n} x_i^2 \right], \quad (2)$$

where n is the number of words, μ is a regularization parameter, and $W_{ij} = 1$ if the word i is a hypernym of the word j and zero otherwise. Here the first sum is taken over all ordered hyponym-hypernym pairs.

The publicly available WordNet 3.0 database (http://wordnet.princeton.edu/) was used in this study. The hypernym-hyponym relations among $n = 124{,}408$ English words were extracted from the database as a connected graph defining the matrix W, which was used to compute the energy function (2). Optimization was carried out with standard MATLAB functions, as described in [23].

3. Results

3.1. Measuring Correlations of Augmented Map Dimensions. The one-dimensional semantic map of "abstractness" was computed as described in Section 2. The resultant distribution of 124,408 WordNet words in one dimension is shown in Figure 2. The two ends of the sorted list of words along their "abstractness" are given in Table 1.

This map was then combined with several antonymy-based weak semantic maps that are previously constructed [11]. The "abstractness" map was merged with any given narrow weak semantic map as the following. First, the set of words was limited to those that are common for both maps. Then, the augmented map was defined as a direct sum of the two vector spaces; that is, the "abstractness" dimension was added as a new word coordinate.

The resultant augmented maps were used to compute the correlation between "abstractness" and other map

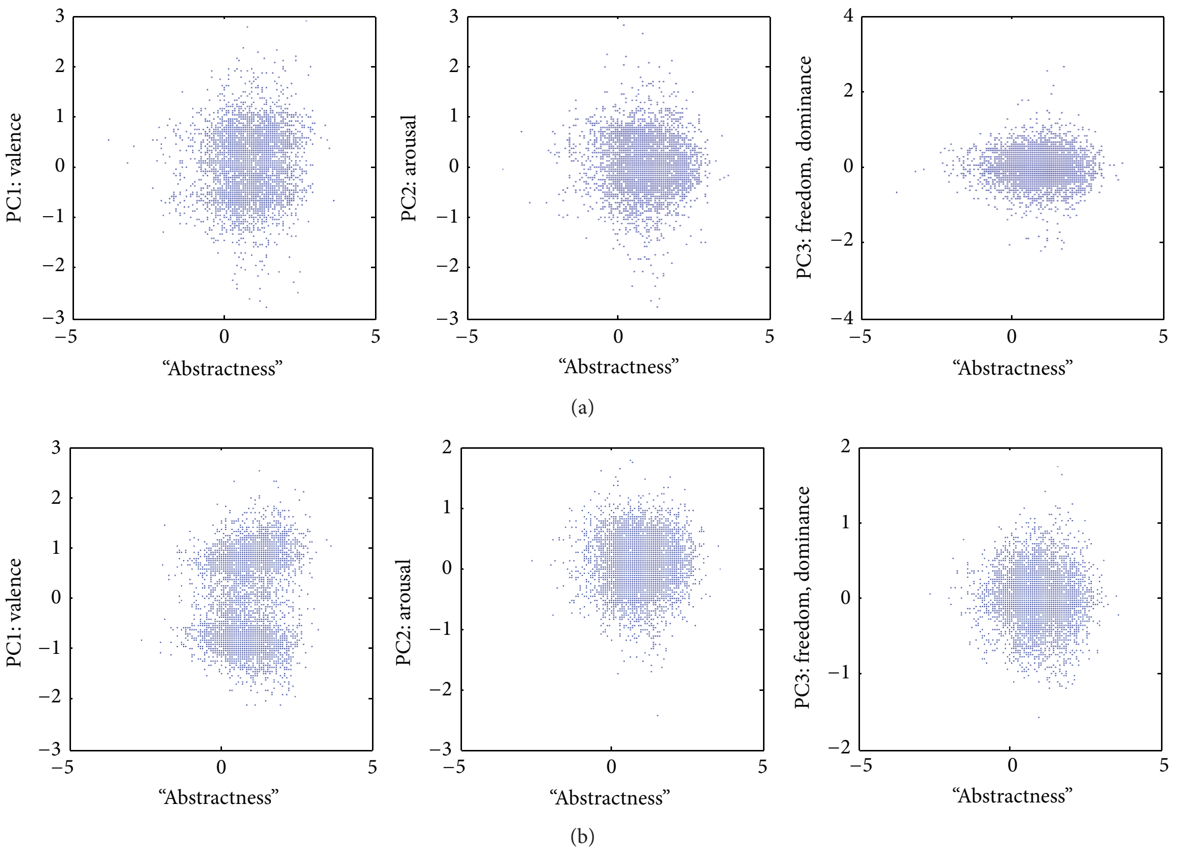


Figure 3: Correlations of "abstractness" with principal components of the antonymy-based weak semantic cognitive maps. (a) The map constructed using WordNet 3.0; (b) the map constructed using the Microsoft Word thesaurus.

Table 2: Pearson correlation coefficient R and the corresponding accounted variance (R^2) of "abstractness" with PC1: valence, PC2: arousal, and PC3: freedom/dominance, measured in four augmented maps constructed based on WordNet 3.0 and the MS Word English, French, and German thesauri.

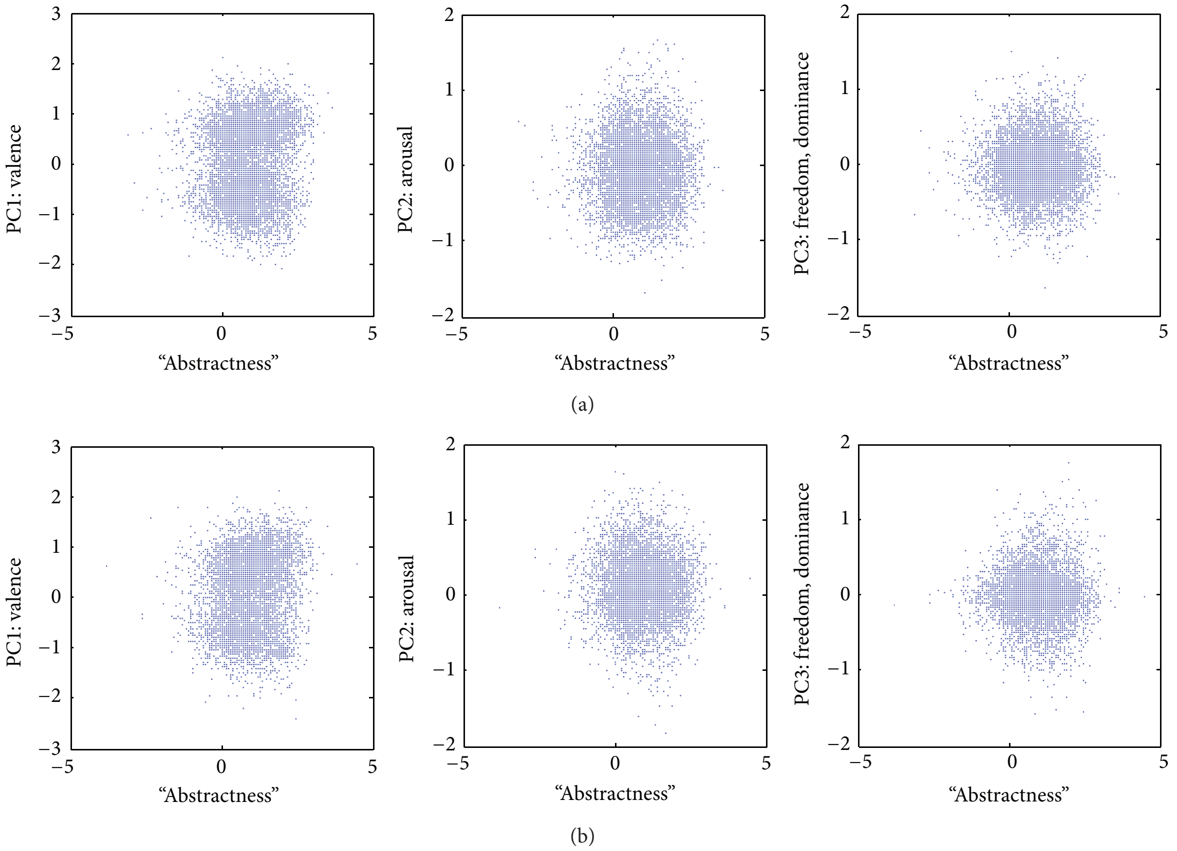
	PC1: valence		PC2: arousal		PC3: freedom, dominance	
	R	R^2	R	R^2	R	R^2
WordNet	0.09	0.8%	−0.07	0.5%	−0.01	0% (NS)
MS Word English	0.12	1.4%	0.01	0% (NS)	−0.03	0.1% (NS)
MS Word French	0.11	1.2%	0.02	0% (NS)	0.01	0% (NS)
MS Word German	0.14	2.0%	−0.02	0% (NS)	0	0% (NS)

dimensions. The main question was how, if at all is the new "abstractness" dimension related to the principal components of the antonymy-based weak semantic map? Figure 3 illustrates the scatterplots of word "abstractness" values derived from WordNet with the dimensions of narrow weak semantic maps derived from WordNet data (Figure 3(a)) and from MS Word (Figure 3(b)). The Pearson correlation coefficient R and the corresponding accounted variance R^2 are given in Table 2 for each PC.

Similar results were obtained for augmented weak semantic maps in other languages (constructed based on the MS Word thesaurus as described in [11]): French (Figure 4(a)) and German (Figure 4(b)). Automated Google translation was used to merge maps in different languages.

In all cases "abstractness" is only positively correlated with valence ($P < 10^{-8}$ in all corpora), while none of the correlation coefficients with the other two dimensions (arousal and freedom) are statistically significant in a consistent way across corpora. Even in the case of valence, the correlation coefficient remains small (Table 2). This finding is further addressed in Section 4.

Overall, the results (Figures 3 and 4) show that the new "abstractness" dimension is practically orthogonal to the narrowly defined weak semantic map dimensions. Indeed,



(a)

(b)

Figure 4: Correlations of "abstractness" with principal components of the antonymy-based weak semantic cognitive maps in other languages. (a) The map constructed using the French dictionary of MS Word. (b) The map constructed using the German dictionary of MS Word.

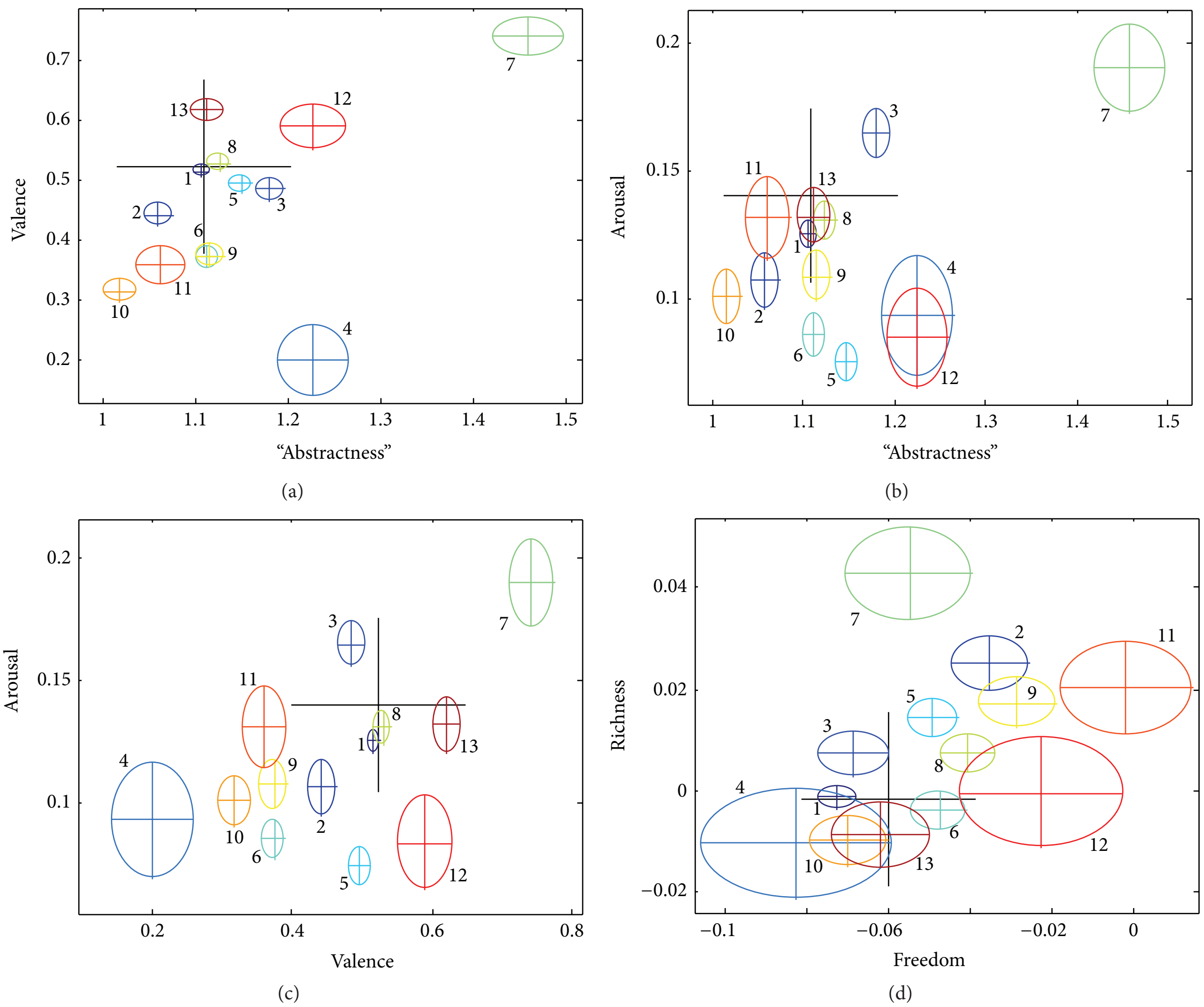
in most cases the correlation is not significant. In the minority of the cases where the correlation is statistically significant, the correlation coefficient is sufficiently small as to become marginal. Specifically, little information is lost by disregarding the fraction of the variance of the distribution of words on the weak semantic map accounted by the word "abstractness" or, vice versa, the fraction of the variance in the word "abstractness" accounted by the weak semantic map dimensions (Table 2).

In conclusion, the previous weak semantic map dimensions do not account for a substantial fraction of variance in "abstractness," and word "abstractness" values do not account for a substantial fraction of variance in the distribution of words on antonymy-based weak semantic maps.

3.2. Examples of Document Mapping with the Augmented Semantic Map. Traditionally, only the valence dimension is used in sentiment analysis. At the same time, other dimensions including "abstractness" are frequently indicated as useful (e.g., [24]). We previously applied the weak semantic map to analysis of Medline abstracts [25]. As an extension of that study, we now applied the augmented semantic map to analyze various kinds of documents.

Using the MS Word English narrow weak semantic map merged with the WordNet-based "abstractness" map, this part of the study asked the following key research questions: how informative is the new dimension compared to familiar dimensions at the document level? Specifically, how well are different kinds of documents separated from each other on the augmented map compared to the narrow weak semantic map? How capable is the new "abstractness" dimension compared to antonymy-based dimensions in terms of document separation? Being aware of more advanced methods of sentiment analysis [21, 26], here we adopted the simplest "bag of words" method (computing the "center of mass" of words in the document, not to be confused with LSA). This parsimonious choice is justified because at this point we are interested in assessing the value of the new dimension compared to familiar dimensions of the narrow weak semantic map, rather than achieving practically significant results.

For each document, the average augmented map coordinates of all words were computed, together with the standard error in each dimension. The results are represented in Figure 5 by crossed ovals, with the center of the cross representing the average and the size of the oval representing the standard error (i.e., the standard deviation divided by the square root of the number of identified words). The large black crosses in each panel represent the average of all words in the dictionary weighted by their overall usage



(a)

(b)

(c)

(d)

Figure 5: Representations of 13 documents (details in the text) on the augmented semantic map. The Pearson correlation coefficient R and the corresponding P value were computed for each panel. None of the correlations are significant. (a) Valence versus "abstractness," $R = 0.54$, $P = 0.06$. (b) Arousal versus "abstractness," $R = 0.50$, $P = 0.09$. (c) Arousal versus valence, $R = 0.54$, $P = 0.057$. (d) Richness (PC4) versus freedom (PC3), $R = 0.46$, $P = 0.12$.

frequency, not limited to materials of this study and derived as in [11]. Colors and numbers of ovals in Figure 5 correspond to RGB values and item numbers given in the following list of corpora:

(1) Project Gutenberg's A Text-Book of Astronomy, by George C. Comstock (http://www.gutenberg.org/files/34834/34834-0.txt), 9626 words, rgb = (0, 0, 6);

(2) Martha Stewart Living Radio Thanksgivings Hotline Recipes 2011 (http://www.hunt4freebies.com/free-martha-stewart-thanksgiving-recipes-ebook-download), 2091 words, rgb = (0, 0, 9);

(3) Al Qaida Inspire Magazine Issue 9 (http://www.en.wikipedia.org/wiki/Inspire_(magazine)), 2555 words, rgb = (0, 2, 10);

(4) A suicide blog (http://www.tumblr.com/tagged/suicideblog), 387 words, rgb = (0, 5, 10);

(5) 152 Shakespeare sonnets [27], 4170 words, rgb = (0, 8, 10);

(6) The Hitchhiker's Guide to the Galaxy, by Douglas Adams (http://www.paulyhart.blogspot.com/2011/10/hitchhikers-guide-to-galaxy-text_28.html), 4187 words, rgb = (1, 10, 9);

(7) 10 abstracts of award-winning NSF grant proposals (downloaded from http://www.nsf.gov/awardsearch), 585 words, rgb = (4, 10, 6);

(8) 196 reviews of the film "Iron Man", 2008 (http://www.mrqe.com/movie_reviews/iron-man-m100052975/), 3902 words, rgb = (8, 10, 2);

(9) 170 reviews of the film "Superhero Movie", 2008 (http://www.mrqe.com/movie_reviews/superhero-movie-m100071304/), 2204 words, rgb = (10, 9, 0);

(10) 160 reviews of the film "Prom Night", 2008 (http://www.mrqe.com/movie_reviews/prom-night-m100076394/), 2114 words, rgb = (10, 6, 0);

(11) 47 anecdotes of/about famous scientists (retrieved from http://jcdverha.home.xs4all.nl/scijokes/10.html), 919 words, rgb = (10, 3, 0);

(12) transcript of Obama's speech at the DNC on September 6, 2012 (http://www.foxnews.com/politics/2012/09/06/transcript-obama-speech-at-dnc), 491 words, rgb = (10, 0, 0);

(13) "Topological strings and their physical applications," by Andrew Neitzke and Cumrun Vafa (http://www.arxiv.org/abs/hep-th/0410178v2), 1909 words, rgb = (7, 0, 0).

The selected documents are mostly well separated in 3 dimensions, including valence (PC1), arousal (PC2), and "abstractness" (Figure 5). At the same time, the ovals more frequently overlap on the plane freedom-richness (PC3-PC4). Visually, "abstractness" is approximately as efficient as valence (PC1) in its ability to separate documents and appears to be more efficient than other dimensions; however, the oval separation on the valence-arousal projection (Figure 5(c)) looks slightly better than on the valence-"abstractness" projection (Figure 5(a)). This observation suggests that disregarding "abstractness" may not significantly affect the quality of results, while disregarding valence would substantially impair the quality of document separation (e.g., on the "abstractness-"arousal plane, Obama's speech overlaps substantially with the suicide blog, while valence separates the two documents significantly).

Differences between the above 13 documents along these 5 dimensions were quantified with analysis of variance. Specifically, the MANOVA P value was 0.027, suggesting that all five semantic dimensions are mutually independent in characterizing the selected 13 corpora. Moreover, in order to compare how informative different semantic dimensions are relative to each other, two sets of characteristics were computed (Table 3), namely, (i) the ANOVA P values to reject the null hypothesis that all 13 corpora have the same mean in each selected semantic dimension and (ii) the MANOVA P values to reject the null hypothesis that the means of all 13 corpora belong to a low-dimensional hyperplane within the space of all but one semantic dimensions.

These results can be interpreted as follows. The lower the P value for ANOVA is, the more informative the selected semantic dimension is. On the contrary, the lower the P value for MANOVA is, the less informative the selected semantic dimension is, because MANOVA was computed in the space of all semantic dimensions except the one selected. Therefore, results represented in Table 3 indicate that "abstractness" (dimension 0) is nearly as informative as valence (dimension 1) and could be more informative than arousal (dimension 2, based on ANOVA only), freedom (dimension 3), and richness (dimension 4). More data are needed to verify this interpretation.

4. Discussion

Statistical analysis indicates that "abstractness" is positively (if marginally) correlated with valence consistently across corpora, which is not the case with other semantic dimensions. On the one hand, the amount of variance in the distributions of words that can be attributed to interaction between valence and "abstractness" is not substantial (only 2% of variance or less); therefore, the two dimensions can be considered orthogonal for practical purposes. On the other hand, the consistent significance of this negligibly small correlation across datasets and languages indicates that there may be a universal factor responsible for it. This factor could be the usage frequency of words that affects the probability of word selection for dictionaries. Stated simply, abstract positive words and specific negative words are used more frequently than abstract negative words and specific positive words. Specifically, our previous study [11] showed that the mean valence (normalized to unitary standard deviation) of all words weighted by their usage frequency is significantly positive (0.50 using frequency data from a database of Australian newspapers and 0.59 using frequency data from the British National Corpus). Using the results in the present study, the mean normalized "abstractness" is between 0.99 (weighted by "Australian" frequency) and 1.39 (weighted by "British" frequency). An equivalent explanation is that abstract words and positive words are both used more frequently than specific words and negative words. Specifically, the correlation with frequency is small but significantly positive both for valence (0.064 Australian, 0.061 British) and for "abstractness" (0.036 Australian, 0.019 British). This interpretation is consistent with data at the level of documents (Figure 5(a)), where the correlation coefficient is even higher, yet not significant (not shown). Another potential source of correlation is the selection of words for inclusion in dictionaries. It seems, however, counterintuitive that the overall picture should be affected by marginal inclusions of rare words. Nevertheless, it would be interesting to check elsewhere how the correlation changes across sets of words found in various types of documents.

The method of weak semantic mapping is an alternative to other vector-space-based approaches including LSA, LDA, MDS, and ConceptNet [1–4], primarily because (i) no semantic features of words are given as input and (ii) the abstract space of the map has no semantics associated a priori with its dimensions. It is therefore not surprising that emergent semantic features (dimensions) in weak semantic mapping are substantially different from emergent semantic dimensions obtained by LSA and related techniques: the latter are typically domain specific and harder to interpret [22].

From another perspective, it is interesting that emergent semantic dimensions of a weak semantic map are so familiar. All generally accepted dimensional models of sentiment, appraisal, emotions and values, attitudes, feelings, and so forth converge on semantics of their principal dimensions, typically interpreted as valence, arousal, and dominance [12–14, 16–18]. Antonymy-based weak semantic mapping appears to be consistent with this emergent consensus [9–11], despite the stark difference in methodologies (human

Table 3: ANOVA and MANOVA P values for selected semantic dimensions characterizing the means of the 13 corpora. Dimensions are numbered as follows: 0, "abstractness", 1, PC1 (valence), 2, PC2 (arousal), 3, PC3 (freedom/dominance), and 4, PC4 (richness).

Semantic Dimension	0	1	2	3	4
One dimension, ANOVA	$1.2e-36$	$5.9e-57$	$3.1e-15$	$2.1e-7$	$6.2e-11$
All but one, MANOVA	0.018	0.040	0.041	$5.1e-7$	$1.4e-7$

judgment or psychometrics versus automated calculations based on subject-independent data). The number of semantic dimensions, or factors, used in the literature varies from 2 to 7, which roughly corresponds to the variability in the number of significant principal components of the narrow weak semantic map [11]. Why do antonyms relating to the "dimensional models" of affect, and not others, make for good PCs? This interesting question remains open and should be addressed by future studies.

The present study unambiguously demonstrates the inability of narrow weak semantic maps to capture all universal semantic dimensions. Here we presented one dimension, "abstractness," that is not captured by "antonymy-" defined weak semantic maps. This is due to the fact that, in general, hypernym-hyponym pairs are not antonym pairs and vice versa. Therefore, hypernym-hyponym relations cannot be captured with the map defined by antonym relations, and the map needs to be augmented. The example of "abstractness" that we found is probably not unique: we expect a similar outcome for the holonym-meronym relation, which will be addressed elsewhere. Our previous results indicated that antonym relations are essential for weak semantic mapping, while synonym relations are not [28].

Thus, the present work shows that narrow weak semantic maps (and related dimensional models of emotions) are not complete in this sense and need to be augmented by including other kinds of semantic relations in their definition. A question remains open as to whether any augmented semantic map may be considered complete—or there will always be new semantic dimensions that can be added to the map. We speculate that there exists a complete finite-dimensional weak semantic map. Moreover, the number of its dimensions can be relatively small. This is because the number of distinct semantic relationships in natural language is limited, as is the number of primary categories [29], or the number of primary semantic elements of metalanguage known as semantic primes [31, 32]. This notion of "completeness," however, may only be applicable to a limited scope, for example, all existing natural languages.

We found that hyponym-hypernym relations induce a new semantic dimension on the weak semantic map that is not reducible to any dimensions represented by antonym pairs or to the traditional PAD or EPA dimensions. Its tentative labeling as "abstractness" or ontological generality, however, remains speculative. In any case, it is not our ambition here to define the notion of "abstractness" or to establish a precise connection between the real notion of abstractness and our new "abstractness" dimension, a topic that should be addressed elsewhere.

Findings of this study have broad implications for automated quantitative evaluation of the meaning of text, including semantic search, opinion mining, sentiment analysis, and mood sensing, as exemplified in Figure 5 and Table 3. While multidimensional approaches in opinion mining are nowadays popular, the problem is finding good multidimensional ranking of all words in the dictionary. Traditional bootstrapping methods (e.g., based on cooccurrence of words) to extend the ranking of positivity from a small subset of words to all words may not work, for example, for "abstractness." The approach presented here should be useful for such applications.

Finally, we speculate that this approach may shed light on the nature of human subjective experience [30] by revealing fundamental semantics of qualia as PCs of the weak semantic cognitive map. In addition, we suggest other connections of our findings, for example, to semantic primes [31, 32].

Acknowledgments

The authors are grateful to Mr. Thomas Sheehan for help with extraction of the relation data from WordNet. Part of Giorgio A. Ascoli's time was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract no. D10PC20021. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US Government.

References

[1] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5228–5235, 2004.

[3] C. Havasi, R. Speer, and J. Alonso, "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '07)*, Borovets, Bulgaria, 2007.

[4] E. Cambria, T. Mazzocco, and A. Hussain, "Application of multidimensional scaling and artificial neural networks for biologically inspired opinion mining," *Biologically Inspired Cognitive Architectures*, vol. 4, pp. 41–53, 2013.

[5] R. F. Cox and M. A. Cox, *Multidimensional Scaling*, Chapman & Hall, 1994.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[7] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, "Spectral methods for dimensionality reduction," in *Semisupervised Learning*, O. Chapelle, B. Schoelkopf, and A. Zien, Eds., pp. 293–308, MIT Press, Cambridge, Mass, USA, 2006.

[8] P. Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, MIT Press, Cambridge, Mass, USA, 2004.

[9] A. V. Samsonovich, R. F. Goldin, and G. A. Ascoli, "Toward a semantic general theory of everything," *Complexity*, vol. 15, no. 4, pp. 12–18, 2010.

[10] A. V. Samsonovich and G. A. Ascoli, "Cognitive map dimensions of the human value system extracted from natural language," in *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms. Proceedings of the AGI Workshop 2006*, B. Goertzel and P. Wang, Eds., vol. 157 of *Frontiers in Artificial Intelligence and Applications*, pp. 111–124, IOS Press, Amsterdam, The Netherlands, 2007.

[11] A. V. Samsonovich and G. A. Ascoli, "Principal semantic components of language and the measurement of meaning," *PLoS ONE*, vol. 5, no. 6, Article ID e10921, 2010.

[12] E. Hudlicka, "Guidelines for designing computational models of emotions," *International Journal of Synthetic Emotions*, no. 1, pp. 26–79, 2011.

[13] C. E. Osgood, G. Suci, and P. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, Urbana, Ill, USA, 1957.

[14] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[15] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, no. 4-5, pp. 529–553, 1982.

[16] A. Mehrabian, *Nonverbal Communication*, 2007.

[17] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "Emotion and motivation: measuring affective perception," *Journal of Clinical Neurophysiology*, vol. 15, no. 5, pp. 397–408, 1998.

[18] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-Cultural Universals of Affective Meaning*, University of Illinois Press, Urbana, Ill, USA, 1975.

[19] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Medical Hypotheses*, vol. 78, no. 2, pp. 341–348, 2012.

[20] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[21] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[22] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds., *2007Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ, USA.

[23] A. V. Samsonovich, "A metric scale for "Abstractness" of the word meaning," in *Intelligent Techniques for Web Personalization and Recommender Systems: AAAI Technical Report WS-12-09*, D. Jannach, S. S. Anand, B. Mobasher, and A. Kobsa, Eds., pp. 48–52, The AAAI Press, Menlo Park, Calif, USA, 2012.

[24] D. McNamara, Y. Ozuru, A. Greasser, and M. Louwerse, "Validating coh-metrix," in *Proceedings of the 28th Annual Conference of the cognitive Science Society*, R. Sun and N. Miyake, Eds., Erlbaum, Mahwah, NJ, USA, 2006.

[25] A. V. Samsonovich and G. A. Ascoli, "Computing semantics of preference with a semantic cognitive map of natural language: application to mood sensing from text," in *Multidisciplinary Workshop on Advances in Preference Handling, Papers from the 2008 AAAI Workshop, AAAI Technical Report WS-08-09*, J. Chomicki, V. Conitzer, U. Junker, and P. Perny, Eds., pp. 91–96, AAAI Press, Menlo Park, Calif, USA, July 2008.

[26] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: an empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[27] W. Shakespeare, *A Lover's Complaint*, The Electronic Text Center, University of Virginia, 1609/2006.

[28] A. V. Samsonovich and C. P. Sherrill, "Comparative study of self-organizing semantic cognitive maps derived from natural language," in *Proceedings of the 29th Annual Cognitive Science Society*, D. S. McNamara and J. G. Trafton, Eds., p. 1848, Cognitive Science Society, Austin, Tex, USA, 2007.

[29] I. Kant, *Critique of Pure Reason*, vol. A 51/B 75, Norman Kemp Smith, St. Martins, NY, USA, 1781/1965.

[30] G. A. Ascoli and A. V. Samsonovich, "Science of the conscious mind," *The Biological Bulletin*, vol. 215, no. 3, pp. 204–215, 2008.

[31] A. Wierzbicka, *Semantics: Primes and Universals*, Oxford University Press, 1996.

[32] C. Goddard and A. Wierzbicka, "Semantics and cognition," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 2, pp. 125–135, 2011.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 294878, 19 pages
http://dx.doi.org/10.1155/2013/294878

Research Article

Hippocampal Anatomy Supports the Use of Context in Object Recognition: A Computational Model

Patrick Greene,[1] Mike Howard,[2] Rajan Bhattacharyya,[2] and Jean-Marc Fellous[1,3]

[1] *Graduate Program in Applied Mathematics, University of Arizona, Tucson, AZ 8572, USA*
[2] *HRL Laboratories, LLC, Malibu, CA 90265, USA*
[3] *Department of Psychology, University of Arizona, Tucson, AZ 8572, USA*

Correspondence should be addressed to Mike Howard; mdhoward@hrl.com

Received 21 December 2012; Revised 26 March 2013; Accepted 4 May 2013

Academic Editor: Giorgio Ascoli

The human hippocampus receives distinct signals via the lateral entorhinal cortex, typically associated with object features, and the medial entorhinal cortex, associated with spatial or contextual information. The existence of these distinct types of information calls for some means by which they can be managed in an appropriate way, by integrating them or keeping them separate as required to improve recognition. We hypothesize that several anatomical features of the hippocampus, including differentiation in connectivity between the superior/inferior blades of DG and the distal/proximal regions of CA3 and CA1, work together to play this information managing role. We construct a set of neural network models with these features and compare their recognition performance when given noisy or partial versions of contexts and their associated objects. We found that the anterior and posterior regions of the hippocampus naturally require different ratios of object and context input for optimal performance, due to the greater number of objects versus contexts. Additionally, we found that having separate processing regions in DG significantly aided recognition in situations where object inputs were degraded. However, split processing in both DG and CA3 resulted in performance tradeoffs, though the actual hippocampus may have ways of mitigating such losses.

1. Introduction

We make sense of the world by comparing our immediate sensations with memories of similar situations. A very basic type of situation is an encounter with objects in a context. For example, objects such as a salt shaker, a glass, and a sink are expected in a kitchen. Even if these objects are encountered in an office, they suggest a kitchen-like function to the area (e.g., it is a kitchenette—not a work cubicle). In other words, the objects evoke the context in which they have been experienced in the past, and the context evokes objects that have been experienced there. The hippocampus, which is essential for the storage and retrieval of memories, is likely to play a central role in this associational process.

In rats, the hippocampus is oriented along a dorsal-ventral axis, while in primates this axis becomes an anterior-posterior axis. In both species, signals reach the hippocampus via the entorhinal cortex (EC layers II and III), which can be divided into lateral and medial portions (denoted LEC and MEC, resp.). Both the LEC and MEC can be further subdivided into caudolateral and rostromedial bands, with the caudolateral bands projecting mainly to the posterior half of the hippocampus and the rostromedial bands projecting mainly to the anterior half [1]. Within the hippocampus, these entorhinal projections reach the dentate gyrus (DG) and CA3 via the perforant path, as well as CA1. Because of the low probability of activation of its neurons, DG is thought to be responsible for producing a sparse representation of a given input which has minimal overlap with other input patterns, thereby reducing interference [2]; however the role of DG in memory is still in question [3–5]. DG projects to CA3 via the mossy fibers, a set of very strong but sparse connections. In addition to receiving inputs from DG and EC, CA3 also has many recurrent connections which are believed to serve a pattern completion purpose, allowing details lost in the sparse DG representation to be recovered in CA3 via recurrent activity and the help of EC perforant path inputs [6, 7]. The proximal region of CA3 (relative to DG) then projects to the distal portion of CA1, while the distal region of CA3 projects to the proximal portion of CA1 [8]. These connections occur in both the anterior and posterior sections of the hippocampus, with each having its own relatively independent (except in the intermediate area between anterior and posterior) DG, CA3, and CA1 subareas.

CA1 receives input from EC, with the distal portion of CA1 receiving input from LEC and proximal CA1 receiving MEC input. CA1 is essential for proper hippocampus function, since CA1 lesions result in anterograde amnesia [9]. The function of CA1 is not fully known however, although several ideas have been suggested based on theoretical [6, 7] or experimental considerations [10, 11]. We propose below a novel role for the distal and proximal areas of CA1. Each of these CA1 regions then sends output to other parts of the brain via two main pathways. The first is via the subiculum (where CA1 proximal connects to the distal part of subiculum and vice versa for CA1 distal) and to EC layers V and VI. The second pathway is via the fornix, which projects to the mammillary bodies and the thalamus.

LEC receives input mainly from perirhinal cortex and MEC receives most of its inputs from parahippocampal cortex (or postrhinal cortex in rats) which receives highly processed sensory information [12]. In this paper, we will refer to information about both the surrounding environment and spatial position within this environment, carried by the MEC, as the "context," and the information carried by LEC as the "object," which may include relational and configural information about objects [13]. It has been shown that in rats, MEC neurons display highly specific spatial grid fields, whereas LEC neurons have only weak spatial specificity [14]. This supports the notion that spatial environmental information arrives at the hippocampus primarily through MEC, whereas nonspatial information (what we call object information) is conveyed through LEC [10, 14]. Note that although our definition of context is based on the physical environment, other equally valid definitions are possible. For example, in a word list memorization task, context can refer either to the list in which a word appears (if there are multiple lists) or to a "processing context" that describes the actions done during the processing of the word, such as counting the number of vowels. It can also refer to a "temporal context" that describes, for example, whether a word was learned later or earlier during a session [15]. In the temporal context model (TCM) [12] and context maintenance and retrieval (CMR) framework [13], context is defined as an internally maintained pattern of activity different from the one corresponding to perception of the item itself. This context, consisting of background information about the object, changes over time and becomes associated with other coactive patterns.

The most obvious use of this incoming object and context information would be to associate and store object and context memories in hippocampus. However, while the necessity of hippocampus for spatial context recognition and navigation is well documented in rats [16, 17], various studies on the role of the rat hippocampus in object recognition have returned surprisingly mixed results. Several studies have found that novel object recognition in rats is impaired following hippocampal damage [18], temporary inactivation of the dorsal region [19], or attenuation of LEC inputs to the dorsal region [10]. These experimental results suggest that detailed information about the world may indeed be represented within the dorsal hippocampus and may be dissociable from contexts, while other studies have concluded that only contextual information is stored in hippocampus [20, 21], or that the hippocampus is not required for intact spontaneous object recognition memory [22]. Analysis of neural spike data during an object recognition memory task in rats showed that hippocampal pyramidal cells primarily encode information about object location but also encode object identity as a secondary dimension [23]. Manns suggested that objects were represented mainly as points of interest on the hippocampal cognitive map, and that this map might aid the rat in recognizing encounters with particular objects [23].

In humans, the question of where memory for objects is stored is still debated, although patients such as H.M. and K.C. who have had bilateral hippocampus removals demonstrate that the hippocampus is required for the formation of new object memories and recall of most short- and medium-term memories (those formed within the last several years) [24, 25]. It is known that the human hippocampus is active during object-type recall [26]. Specifically, during successful memorization of word lists, there is significantly more activation of the posterior hippocampus than the anterior hippocampus [27]. A greater degree of posterior activation is also seen during the encoding of novel pictures [28]. However, the posterior region often responds to spatial tasks as well, particularly those concerning local spatial detail (see [29] for a review of differences in spatial and other types of processing between the anterior and posterior regions). In this study we assume that both specific object and context representations exist and are stored as memories within the hippocampus. While both regions seem to process spatial contextual information, only the posterior region has been strongly implicated in object memory as well. We therefore hypothesize that the anterior region of the primate hippocampus is primarily processing contextual information, while the posterior region is relatively more object oriented. The models that we develop in this study have explicit object recognition as a main feature and should therefore mainly be considered models of the primate hippocampus because of the evidence for explicit object representations in this case. We will discuss how our models can be related to the rat hippocampus in Section 4.

In summary, we assume that object and context memory are mainly stored in the posterior and anterior regions of hippocampus, respectively. Recall, however, that the posterior region also receives input from the caudolateral band of the MEC (which carries contextual information), and the anterior region receives input from the rostromedial band of the LEC (which carries object information). These connections raise the question of the purpose of having both object and context information reach the posterior and anterior subdivisions of the hippocampus. Recent reconsolidation experiments have shown that spatial contextual information plays a significant role in object retrieval and encoding [30, 31]. We propose that the MEC connections to the posterior stream mentioned above are vital for this. The experiments we describe next explain why context plays such a pivotal role in memory. We provide evidence that elements of hippocampal anatomy such as differentiation between the blades of DG and functional separation of the distal and proximal regions of CA1 may work together to improve the selective use of context information in object recognition, and that this can in turn improve memory performance in certain situations.

Overall, we attempt to formulate a coherent explanation for the role of several distinct anatomical features of the hippocampus and how they work together. This explanation centers on the idea that some of these anatomical differences may have evolved in order to deal with the two intrinsically different types of information that enter the hippocampus through LEC and MEC. These two types of information are "object" information (specific items within an environment, e.g., a spoon) and contextual information (the environment itself—generally less numerous than objects and related to general classes of objects, e.g., the kitchen).

Our hypothesis is that the anatomical features of the hippocampus can help manage the flow of these two types of information better than an undifferentiated hippocampus could—that they allow these two types of information to come together only in areas where it is beneficial and keep them apart otherwise. The question we are addressing in this paper is the following: can these anatomical features actually improve performance by playing the information managing role that we have proposed? We determine this by testing on a number of basic memorization tasks and find that the models with these features do indeed perform better than the baseline model on some of the tasks.

Why would we want to examine this question? There has been a large amount of work done on the theoretical aspects of how the hippocampus stores generic inputs and what role each of the main subregions (DG, CA3, and CA1) may play. In recent years, however, anatomical studies have demonstrated that there is a high degree of differentiation in terms of connectivity along multiple axes of the hippocampus (posterior-anterior and distal-proximal) and within each of the subregions. At the same time, experimental studies have shown that this differentiation has actual consequences for the memorization ability of different regions, and the studies above have shown that context plays an important role in object memorization. Thus, it is important to consider how these new findings fit into the theoretical picture of how the hippocampus works. We can no longer just consider the hippocampus or its subregions as single blocks (CA1, CA3, . . .) nor consider all inputs as homogeneous if we are to have any hope of explaining existing behavioral data at the neural network level. We come at the question of how the anatomical data can explain the new experimental data with two important ideas that we believe have not been adequately expressed up to now: (1) that the anatomical features mentioned above play an information managing role whose existence only becomes necessary once we start to consider at least two different types of information converging in the hippocampus and (2) that the roles of these individual features only make sense when looking at their interaction with everything else; for example, differentiation within DG on its own would be less useful for managing information if the rest of the upstream regions like CA1 did not also have features (like the proximal-distal distinction in our model) that make use of how DG partitions this information.

2. Methods

2.1. Model Structure and Connectivity. We use an expanded version of a model of the hippocampus developed by O'Reilly et al. [32]. The original model is a basic hippocampus consisting of a single input (EC layers II and III), a DG, CA3, and CA1 layer and a single output (EC layers V and VI). This model includes recursive connections within CA3 and DG to CA3 connections that are 10 times stronger than the EC to CA3 connections to mimic the sparse but powerful mossy fiber synapses. The smallest computational element is a "unit," which simulates a small population of neurons in a rate-coded fashion [33]. We will use the term neuron synonymously with unit in the rest of the paper. The network is trained using the Leabra algorithm, which is based on the generalized recirculation algorithm. Unlike the original model, we do not pretrain the EC → CA1 → output connection. In addition, we did not model an explicit EC output layer; we simply have an output layer. Further details of the original model can be found elsewhere [6, 34].

Our model explicitly separates the posterior and anterior halves of the hippocampus, so that the network has two CA3 regions, two DG regions, and two CA1 regions, each in the posterior and anterior poles. EC is split into lateral and medial regions (LEC and MEC, resp.), with LEC connected to all three layers on both the posterior and anterior sides to simulate the outputs of the caudolateral and rostromedial bands, respectively, and similarly for MEC. As supported by the neuroanatomy, CA3 proximal (in relation to DG) connects to CA1 distal and CA3 distal connects to CA1 proximal [8]. In order to model this distal/proximal connectivity distinction, we split each of the two CA1 regions into half again, to give four separate CA1 regions (two on the posterior side and two on the anterior side). Each CA1 receives input from the ipsilateral CA3 along with either LEC input (if it is distal) or MEC input (if it is proximal). This network will be referred to as the "Baseline" network (Figure 1).

We model inhibition in each layer as a competitive k-winner-take-all process, where only the top k most active neurons send their outputs to the next layer. Thus we can set the activity level in each region to approximately that seen in experimental results, where the activity level refers to the percentage of active neurons at any given time. EC, DG, CA3, and CA1 have experimental activity levels of 7%, 1%, 2.5%, and 2.5%, respectively [34]. In our model, these levels are set to 25%, 1.5%, 2.3%, and 2.5%, respectively. The discrepancy in EC (both LEC and MEC) is because it is serving as our input layer and does no computation; EC is just large enough to hold training patterns with 25% of the units active. The LEC and MEC layers each consist of 64 neurons. The DG, CA3, and CA1 layers on the posterior side consist of 800, 256, and 800 neurons, respectively (the distal and proximal regions of CA1 have 400 neurons each). The same numbers apply on the anterior side.

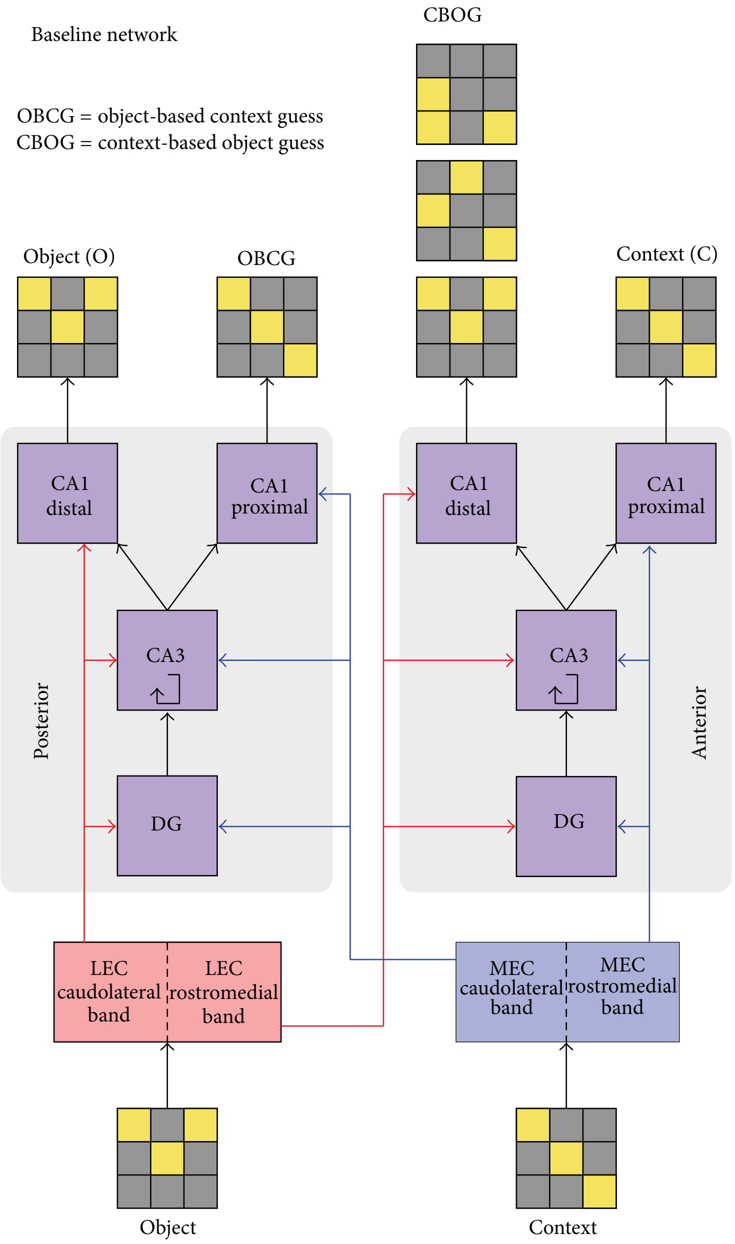


Figure 1: Layer and connectivity diagram of the Baseline network. Matrices representing an object and a context are the inputs to the network. The outputs are an object (O), an object-based context guess (OBCG), a context-based object guess (CBOG), and a context (C). The OBCG output is the context that the input object is associated with during training, and the CBOG output is the set of objects that were associated with the input context during training.

As discussed above, the LEC primarily carries object information while the MEC carries spatial contextual information. Hence in our model we conceptualize the LEC inputs as "objects" and MEC inputs as "context." In assigning roles to the output layers corresponding to the distal and proximal CA1 regions, we first note that these two regions lie on largely separate output pathways: CA3 proximal connects mainly to CA1 distal and CA1 distal connects mainly to the proximal part of the subiculum, which in turn projects back to the LEC [8, 35]. On the other hand, CA3 distal connects mainly to CA1 proximal and CA1 proximal connects to the distal part of the subiculum, which in turn projects back to the MEC [8, 35]. If these pathways were both carrying the same type of information, there would be no need for such a wiring scheme to keep them separate. Since our model only contains two types of information, object and context, we assume that one of these pathways is carrying object information and the other is carrying context.

On the posterior side of hippocampus we are mainly focused on its object processing capabilities; hence we assume that the relevant outputs must be largely dependent on using object-type information from LEC. We hypothesize that these two outputs are an object guess and an object-based context guess. The object guess pathway does standard object recognition by taking the input object, matching it to the closest object in memory, and giving the best match as its output. The object-based context guess pathway uses the object input to generate the context that the object is associated with: if one gives it the object "swing set," it returns "playground," if one gives it "refrigerator," it returns "kitchen," and so forth. We emphasize that not every neuron in the given regions is doing these operations or using only one type of information to do them. But, to the extent that we have neurons that are encoding nonspatial information in these regions, we predict that there will be more of them (or alternatively, that the degree to which they are sensitive to spatial information will be lower) in the distal region of CA1 compared to the proximal region. Experimental results by Henriksen et al. provide support for this, showing that the strongest spatial modulation occurs in the proximal part of CA1, and that distal CA1 cells are less spatially tuned [36].

On the anterior side of the hippocampus, since we focus on its contextual processing capabilities, we require that its outputs be largely dependent on using context-type information from MEC. We hypothesize that these two outputs are a context guess and a context-based object guess. The context guess pathway matches the input context to the closest context in memory, and the context-based object guess uses the input context to generate a list of the set of objects associated with the given context. For example, given the context input "playground," it would output the object list "swings, sandbox, slide."

The final question is which of the distal or proximal CA1 regions is playing each of these roles. It is known that MEC projects preferentially to the proximal region of CA1, while LEC projects preferentially to the distal region [37]. Assuming that the purpose of the two CA1 streams is to keep object and context-type information largely separate, it seems unlikely that object information from LEC would then be projected to the context stream at CA1, and similarly for MEC inputs and the object stream. Thus, on the posterior side, we conclude that the object guess is output by distal CA1 and the object-based context guess is output by proximal CA1. Similarly, on the anterior side, we conclude that the context-based object guess is output by distal CA1, and the context guess is output by proximal CA1.

2.2. Model Variants. Variants of the Baseline network were designed to investigate the effect of two additional anatomical details. The first is the differentiation between the inferior and superior blades of DG. As shown in Figure 2, the DG may be functionally separated into two parts because of the different strengths of LEC and MEC connections onto the superior and inferior blades and a postulated dendritic gating mechanism [38, 39]. Both blades receive proximal dendritic MEC input via the medial perforant path (MPP) and distal dendritic LEC input via the lateral perforant path (LPP). However, the superior blade receives stronger LPP input whereas the inferior blade receives stronger MPP input. We further hypothesize that the effect of this connectivity is different depending on whether the given DG region lies in the posterior or anterior hippocampus.

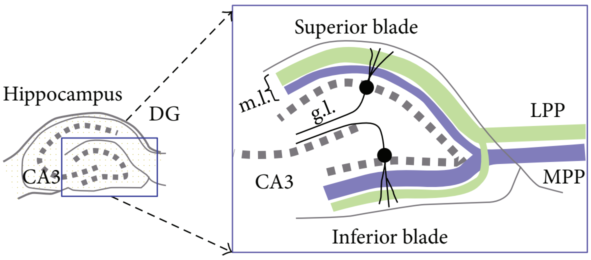


Figure 2: Connectivity of lateral perforant path (LPP) and medial perforant path (MPP) inputs to superior and inferior blade of DG. The LPP and MPP fiber lamina are thicker on the superior blade and inferior blades, respectively, resulting in higher effective synaptic weights (adapted from [38]).

In the posterior hippocampus, the object information contained in the LPP input is more relevant to its task than the context information coming from the MPP input. Thus we would expect that the DG neurons in posterior hippocampus would be biased toward (or learn to weight more heavily) the LPP inputs over the MPP inputs. However, the fact remains that the MPP inputs are more proximal to the soma and thus cannot be completely ignored. The hypothesized result of this tug-of-war (more relevant LPP input but more proximal MPP input) is that, in the superior blade where the LPP object inputs are already stronger than the MPP context inputs, LPP is able to largely control the neurons' firing. In the inferior blade where LPP inputs are weaker, they are able to achieve approximate parity with the MPP input.

In anterior hippocampus the MPP contextual inputs are both more relevant and more proximal to the soma. We hypothesize that this allows the MPP inputs to control the neurons' firing, though to a greater extent in the inferior blade than the superior blade, where LPP input cannot be totally ignored.

We model the two blades of DG as separate layers in both the anterior and posterior sides of hippocampus in order to determine their effect on performance. The model with DG layers split in this way, but with all other architecture the same as in the Baseline model, will be referred to as the "SplitDG" model (Figure 3).

The second anatomical detail we consider is differentiation between the proximal and distal regions of CA3. As mentioned in the introduction, CA3 has distal and proximal regions just as in CA1 (here distal and proximal refer to distance from DG, rather than to the location on the dendrite). These regions receive different amounts of inferior and superior blade DG input and have distinct patterns of recurrent connections [8]. The amount of recurrent versus feedforward connections is also different between the two subareas. Thus these two regions of CA3 may be performing functionally different roles. In order to determine the purpose of such a split and test whether it may confer some performance advantage, we construct a third network that has CA3 split into two layers on each of the posterior and anterior sides, in addition to the DG split described above.

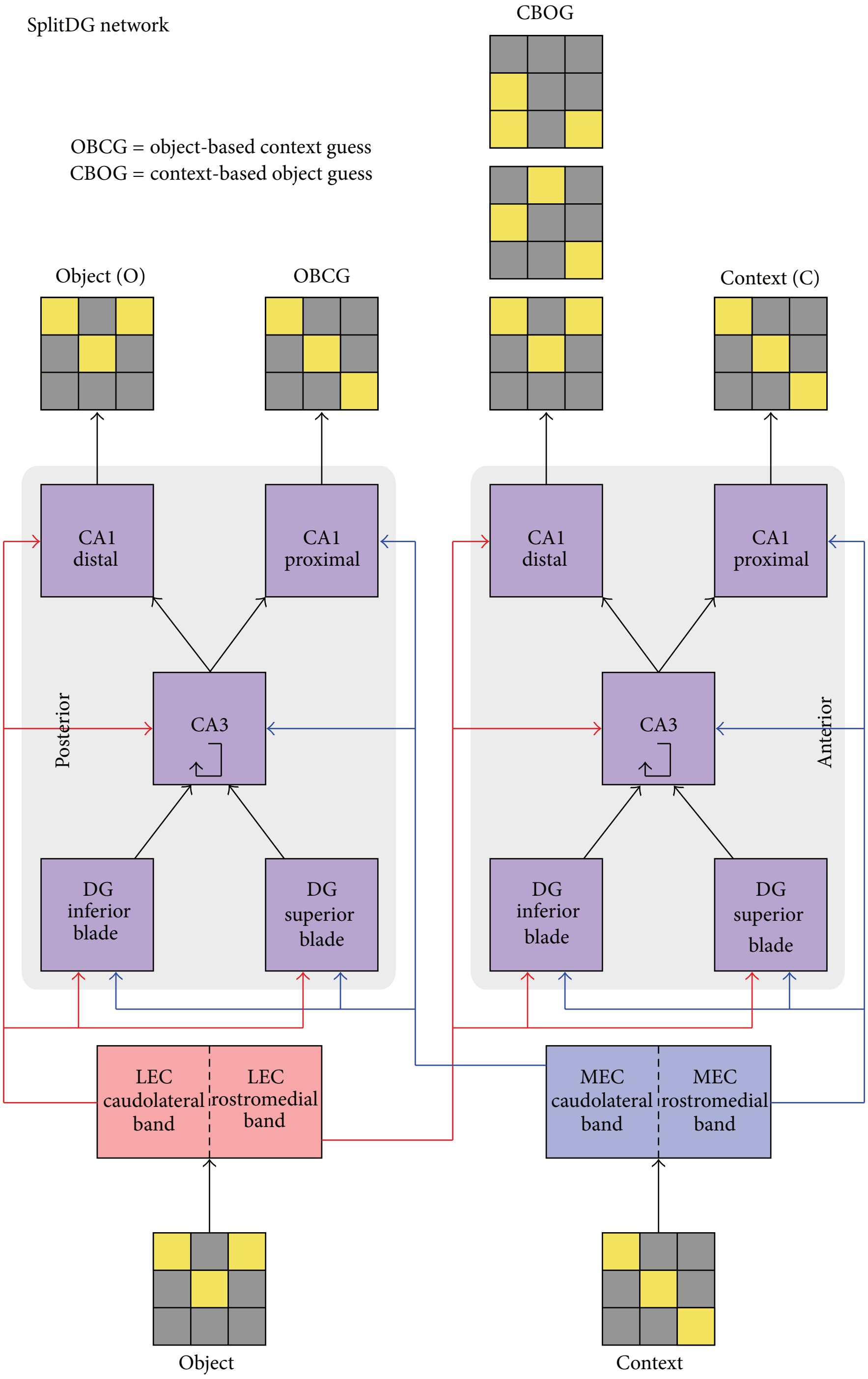


Figure 3: Layer and connectivity diagram of the SplitDG network.

Anatomically, the inferior blade of DG projects to proximal CA3, while the superior blade projects to both proximal and distal portions of CA3 [8]. As a modeling approximation we connect the inferior blade to proximal CA3 and the superior blade to distal CA3 only. Although our model does not capture the detailed connectivity of CA3, we believe it serves as a good starting point for understanding the purpose of having distinct CA3 regions. We will refer to this network as the "AllSplit" network (Figure 4).

2.3. The "+" Networks. We constructed two additional networks, SplitDG+ and AllSplit+, for the purposes of comparison across networks with equal training set error. SplitDG+ is the same as SplitDG, except that each of the DG layers is

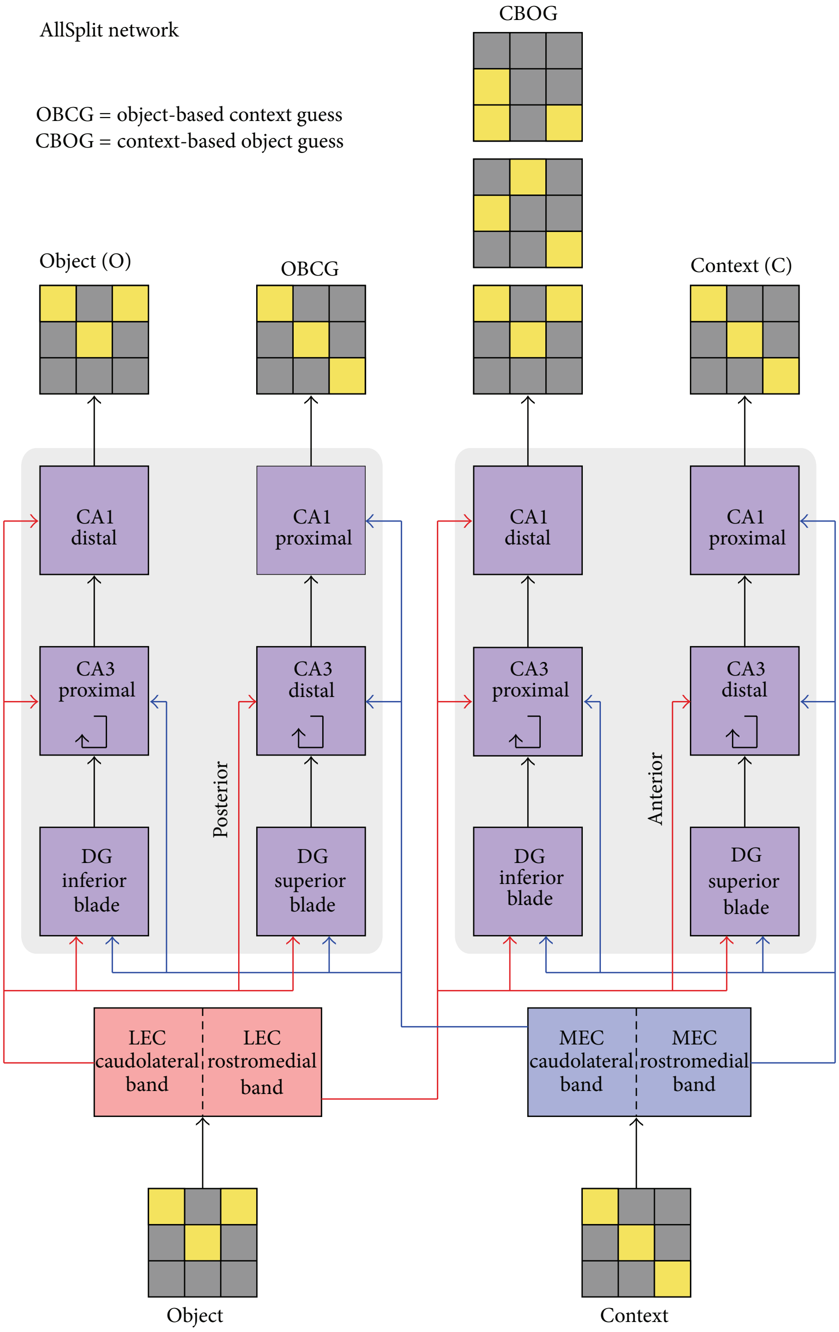
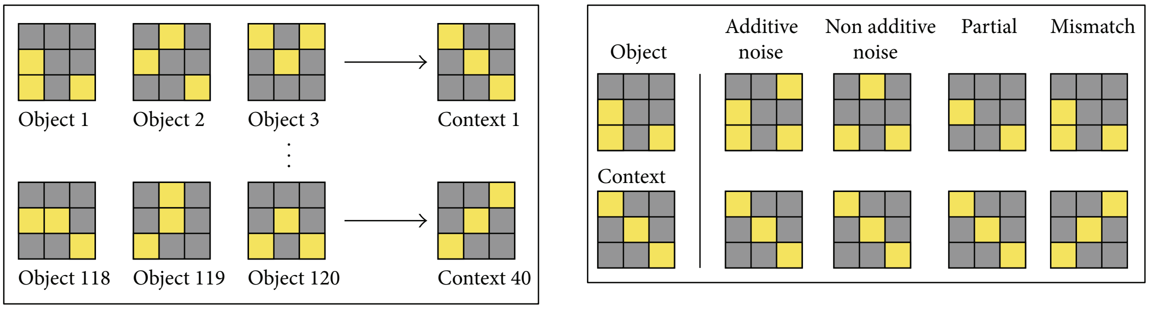


Figure 4: Layer and connectivity diagram of the AllSplit network.

doubled in size. Similarly, AllSplit+ is the same as AllSplit, except that both the CA3 and DG layers have been doubled in size. The relevance of these networks is addressed in more detail in the discussion.

2.4. Training and Test Sets. The training set consists of object patterns and context patterns (Figure 5). Each object is a random 8 × 8 matrix of zeros and ones, consisting of 16 ones (active units) and 48 zeros (inactive units). Contexts are constructed the same way. There are 120 unique objects and 40 unique contexts (3 unique objects per context).

The output layers of the network are referred to as "object" (O), "object-based context guess" (OBCG), "context-based object guess" (CBOG), and "context" (C). The correct output for the object output layer (used as a training signal and ground truth for the error metric) is the object matrix for the input object. For the OBCG layer, the correct output is the context matrix associated with the given object input. For the CBOG layer, the correct output is the three object matrices for the three objects associated with the given context. Finally, for the context output layer, the correct output is the context matrix for the input context.



(a) Training set

(b) Test set

Figure 5: Training and test sets. The training set consists of 120 objects and 40 contexts, with 3 objects per context. The test sets are the same as the training set, except with either noise added (additive or nonadditive noise), part of the pattern missing (partial cue), or an object and context mismatch.

The network is trained for 20 epochs, where each epoch consists of presenting all 120 object-context pairs in a random order and applying the Leabra weight update algorithm after each presentation. Twenty epochs were chosen as the stopping point because all networks' training error had stabilized at close to their minimum value by this time.

After training, the networks' weights are frozen, and the networks' performance is measured using four test sets: additive noise, nonadditive noise, partial cue, and context mismatch (Figure 5). In additive noise tests, objects or contexts have some of the zeros in their matrix replaced by ones, simulating additional active units. In non-additive noise tests, for each zero that is replaced by a one, a one from the original pattern is replaced by a zero, so that the total number of active units remains the same. In partial cue tests, some of the ones in the original object or context pattern are replaced by zeros, resulting in a fewer number of active units overall. In the context mismatch test, an object is paired with a different context from the one it was associated with during training. The level of difficulty of each test depends on the number of units that are changed from the original pattern, which we denote by percentages in the figures.

Many experimental or real-life situations can be interpreted in terms of these simple tests or a combination of them. For example, if the object we are memorizing is a man's face, we recognize who he is even if he has grown a mustache (additive noise), is wearing a hat (non-additive noise, since it adds something but also covers his hair, which is one of his original features), or is partially turned away from us (partial cue). In addition, we recognize him even if we see the same man in a different context (mismatch), although this may be a somewhat more subtle issue than the previous ones, which we will discuss further.

3. Results

3.1. Setting the Crossconnection Weights for the Baseline Model. We will refer to the connections from LEC to the anterior side of hippocampus and from MEC to the posterior side as "crossconnections," since they bring object information into the context-dominated anterior side and context information into the object-dominated posterior side, respectively. The first task was to determine how the relative amount of crossconnection and noncrossconnection input affects the error rate of the Baseline network and use this to maximize its performance. Since the OBCG and CBOG output layers are used in different situations from the O and C layers, we test them accordingly on a different set of tasks. The O and C layers were tested on a set with mixed additive and non-additive noise introduced to object and context (15% noise in each layer) and a set where both object and context were incomplete (40% complete each). The OBCG layers were tested when object and context were mismatched, with noise (30%) in context only, and partial (40%) in context only. For the CBOG layer, the mismatch test was the same, but the noise and partial tests were in the object input only (30% object noise and 40% partial object) rather the context. The results can be seen in Figure 6.

To determine the optimal LEC and MEC weights for each output stream, we plot each output layer's average error over the set of relevant tests as a function of the crossconnection input it receives. This is shown in Figure 7. We use this as a guide to set the relative weights of the crossconnections for all the networks to levels which optimize their performance on the sample tests. Note that for networks such as SplitDG or AllSplit which have split layers, we optimize the crossconnection strengths for these layers independently, while for the Baseline network, we must average the optimal connection strengths over the two output types. For example, since the O output does best with a multiplier of 3 while OBCG does best with a multiplier of 0, we end up with the Baseline network having a relative weight multiplier of 1.5 for the MEC to dorsal side crossconnections. For the AllSplit network, we do not need to make this compromise and can directly use a multiplier of 3 for the MEC inputs into the DG and CA3 areas which feed into O and use a small multiplier close to 0 for the DG and CA3 areas which feed into OBCG. The SplitDG network has the same weighting for crossconnections to DG and CA1 as the AllSplit network and the same weighting to CA3 as the Baseline network, since it only has a single CA3 which the O and OBCG streams must share. These results show that there is unlikely to be a single

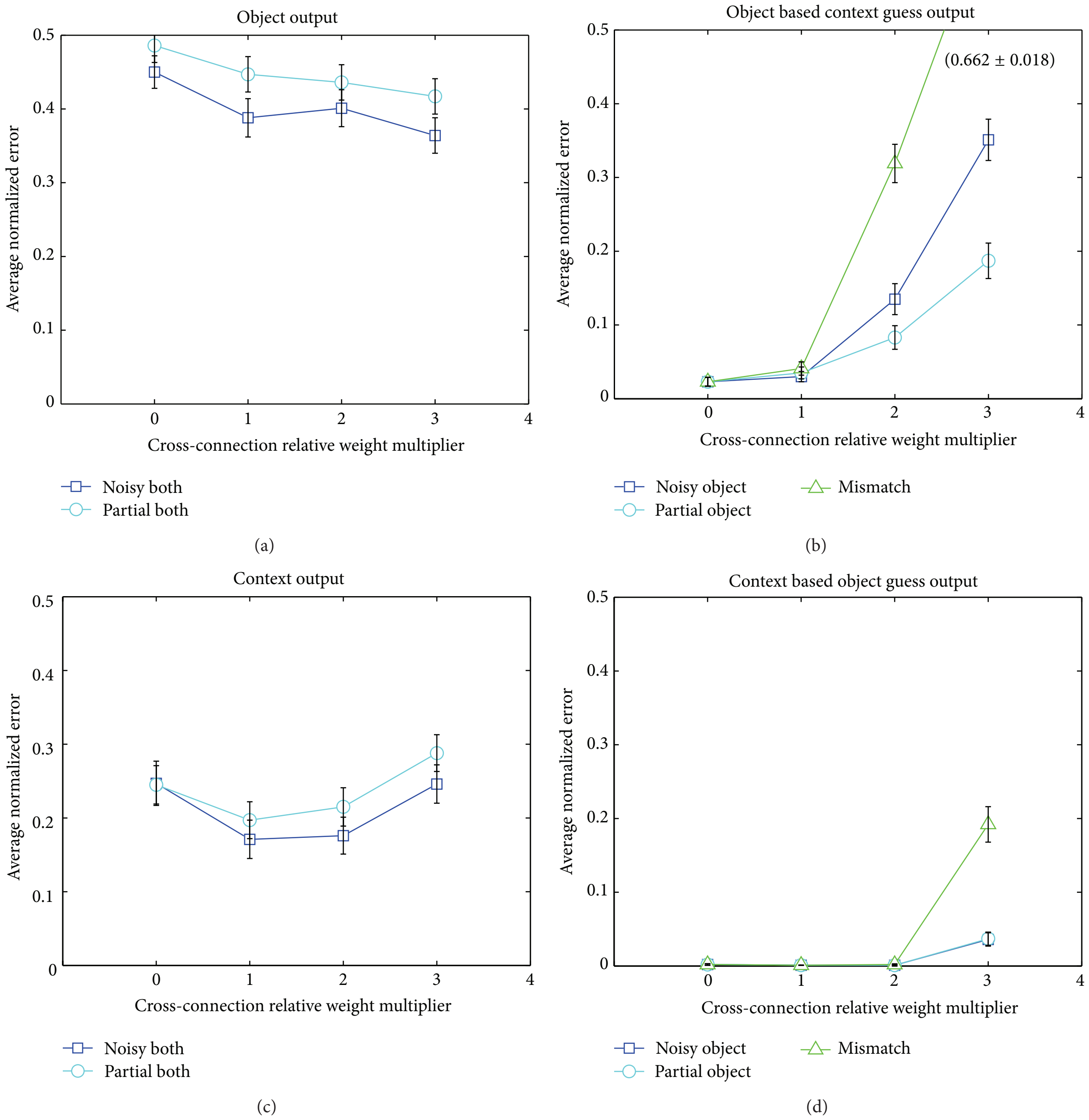


FIGURE 6: Error for each of the four output layers of the Baseline network on various sample tasks, as a function of crossconnection weighting. Crossconnection input refers to LEC input to anterior hippocampus and MEC input to posterior hippocampus. Higher relative weight multiplier values mean stronger MEC input to posterior and stronger LEC input to anterior streams. (a) Object output error on noisy and partial cue tests (where both object and context are noisy or partial, resp.) as a function of crossconnection strength. (b) OBCG output error on noisy and partial cue tests (here the noise and partial are only in the context) as a function of crossconnection strength. (c) Same as A, except the error is measured at the context output layer. (d) Same as B, except only the object is noisy or partial, and the error is measured at the CBOG output layer. Error bars are standard errors of the mean.

set of crossconnection weights that optimizes performance for the various output layers across a range of different tasks. The flexibility provided by having different DG and CA3 layers that can take different levels of crossconnection input provides an advantage and may be one of the reasons why this anatomical differentiation exists in the hippocampus.

3.2. Training Error. Having fixed the crossconnection weights in all networks to values that minimize the error over the sample test sets, we now compare the networks. First we measure the error on the training set after 20 epochs, when the error has reached its asymptotic minimum. Figure 8 shows the average error for each of the five networks,

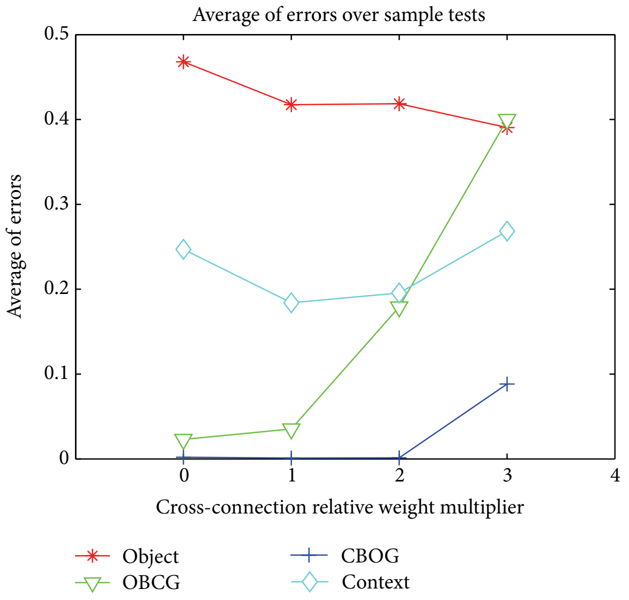


FIGURE 7: Average of errors over sample tests for each output layer, as a function of crossconnection strength.

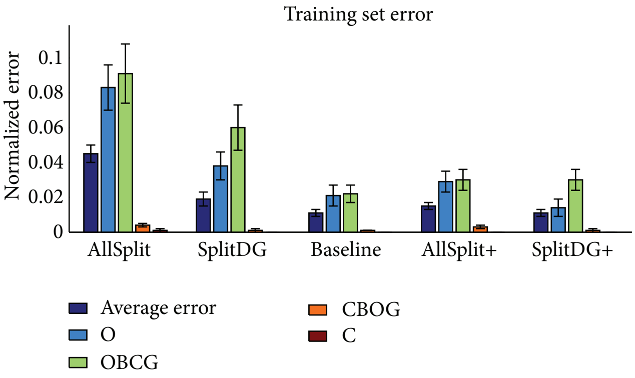


FIGURE 8: Error on the training set for each of the five networks after 20 epochs of training.

along with the error on each of the four outputs individually. The networks can be divided into two categories for further comparison: those which have the same number of neurons, consisting of AllSplit, SplitDG, and Baseline and those which have the same initial training set error, consisting of Baseline, AllSplit+, and SplitDG+. This illustrates the fact that differences in layer size may play an important role in the networks' basic memorization ability. When a layer is split, each of the halves can specialize more efficiently on the task, for example, pattern completing an object or converting an object to a context guess. On the other hand, it must hold the same number of object or context memories despite being half the size, resulting in more memorization errors. Figure 8 shows two possible outcomes of this tradeoff: for the context and CBOG streams, there is no difference in training error before and after splitting the CA3 and DG layers which lie on those streams (compare C and CBOG error between Baseline, SplitDG, and AllSplit). This is due to the fact that these layers only need to store 40 context memories, so even when they are split in half they have no difficulty memorizing them all. However, for the object and OBCG streams, splitting their respective DG or CA3 layers results in a significant increase in training error (compare O and OBCG error between the same three networks). In this case they need to memorize 120 objects, and a CA3 or DG layer half the size is not sufficient. The results of the "+" networks show that this is no longer a problem if we simply have more neurons to start with. The question of whether it is more appropriate to compare Baseline with AllSplit+ and SplitDG+ (since they start off with the same training set error) or to compare Baseline with AllSplit and SplitDG (since they have the same number of neurons) depends on which situation is more likely to reflect biological reality and will be addressed further in the discussion. In all subsequent tests we include the results for each of the five networks.

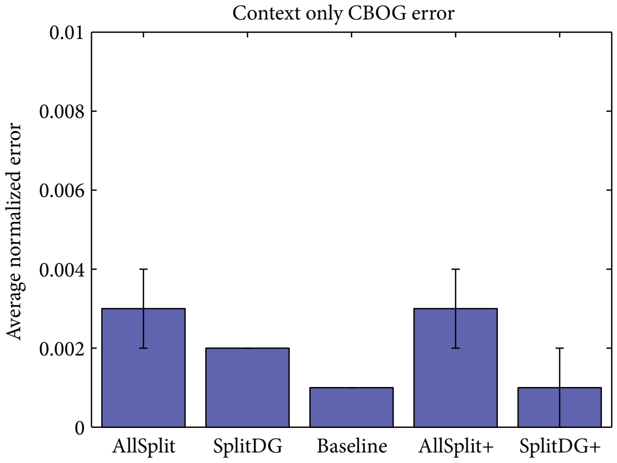
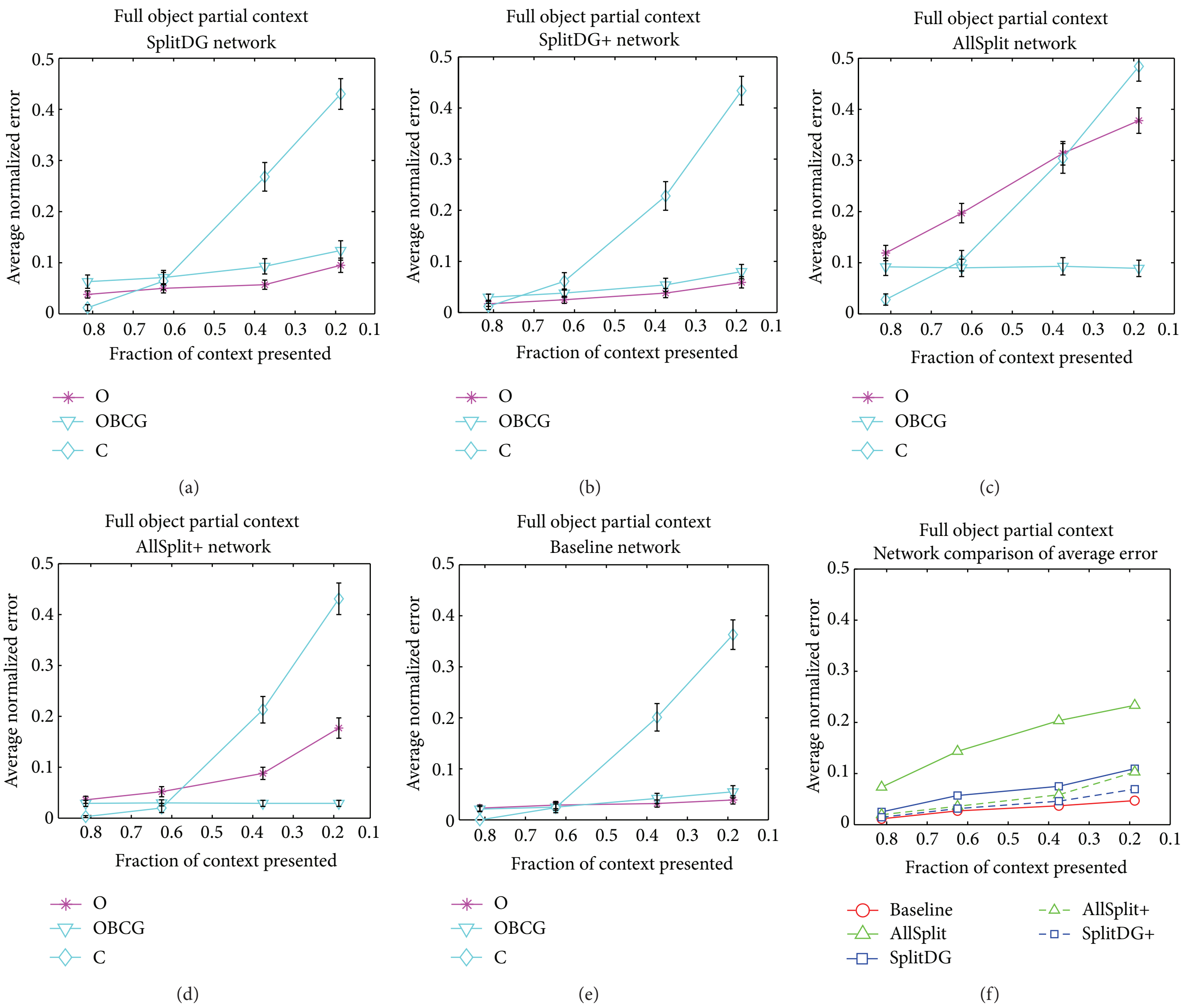


FIGURE 9: Error on the context-based object guess (CBOG) output when given only the context as input.

3.3. Test Sets. We seek to determine how, and in what situations, contextual information can be used by the hippocampus to aid in object recognition and recall (and similarly how object information can aid context recognition), and what role differentiation within DG and CA3 may play in using this information. To answer these questions, we have constructed three primary networks with varying degrees of differentiation in the DG and CA3 layers and will test the ability of each of these networks to recognize objects and contexts under various conditions of degraded inputs.

A common and simple test of human memory is to have a subject memorize a list of words or set of objects, then recall them given a cue. We would like to determine if our network is capable of giving this object output even without the object input. We simulate this task in our networks by presenting a context (the cue—which would consist of the room and the experimenter) and use the CBOG output to get a list of the objects which have been memorized in the given context. Figure 9 shows that the CBOG stream performs well in this task. There is little difference between networks here since all



(a)

(b)

(c)

(d)

(e)

(f)

Figure 10: Error for each of the networks' O, C, and OBCG layers when a partial context and full object were given as input. (a) SplitDG, (b) SplitDG+, (c) AllSplit, (d) AllSplit+, (e) Baseline, and (f) average error across the object output and the lowest of the two context outputs (C or OBCG) for each network, as a function of percentage of context input presented.

use the same crossconnection strength into the anterior side, where CBOG is located.

Next we consider the case where the context, rather than the object, is missing to various degrees. This test will help us determine the degree to which relying on contextual input to recognize objects is disadvantageous when the context is degraded. Figure 10 shows the individual performance of the output layers O, OBCG, and C as a function of how much of the context is given for the various networks, illustrating the effect of having increased MEC inputs into the object stream. Because the AllSplit network's object stream uses a relatively large amount of context information, partial context input has a greater adverse effect on the AllSplit network's O output than it does on the Baseline network's O output. The same is true for SplitDG and its "+" counterpart. Thus we do not expect the AllSplit network to do well compared to the Baseline network in this situation, and Figure 10(e), which gives the average error for each network by taking the average of the error from the O output and the best context output (either C or OBCG), confirms this. The "+" networks do relatively better since their larger CA3 sizes allow the partial context-object mix within the object stream to be pattern completed to a higher degree. This figure also shows the advantage (for all the networks) of having an OBCG output when context is difficult to discern. When the fraction of context drops below 60%, the networks can rely on OBCG for their context guess rather than the context stream output C.

The analogous situation on the object side is to present a partial object and a full context. This test helps us determine how well the various network architectures can utilize context to aid object recognition. At first glance it seems that we ought to make use of the CBOG output to generate an object guess

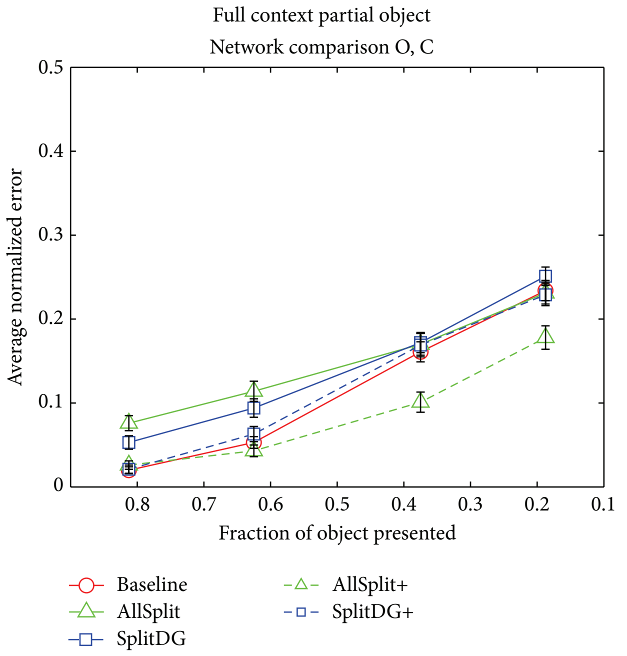


Figure 11: Average error on the O and C output layers when full contexts and partial objects were given as input.

using the clean context, just as we used the OBCG layer in the partial context case above. However, the problem is that the CBOG layer activates multiple possible objects rather than a single object, and thus we would need a way of picking the correct object out of this list. Cortical areas outside of hippocampus could conceivably accomplish this by picking the closest match either to the original input or the O output; however, since we restrict our model to the hippocampus proper, we have not attempted to implement such a scheme and instead use the O output as our exclusive object guess. We consider this issue further in the discussion. Figure 11 shows that, when the object is partially given, the increased amount of context information that the AllSplit network uses via the MEC to posterior crossconnections becomes an advantage rather than a liability, as it now has an error rate similar to that of the Baseline network. When the initial training set memorization disadvantage is accounted for under the AllSplit+ network, a consistent advantage for all partial conditions is seen. Surprisingly, neither SplitDG nor SplitDG+ is able to do better than the Baseline network, suggesting that some degree of heterogeneity within CA3 is necessary to take advantage of the additional context information.

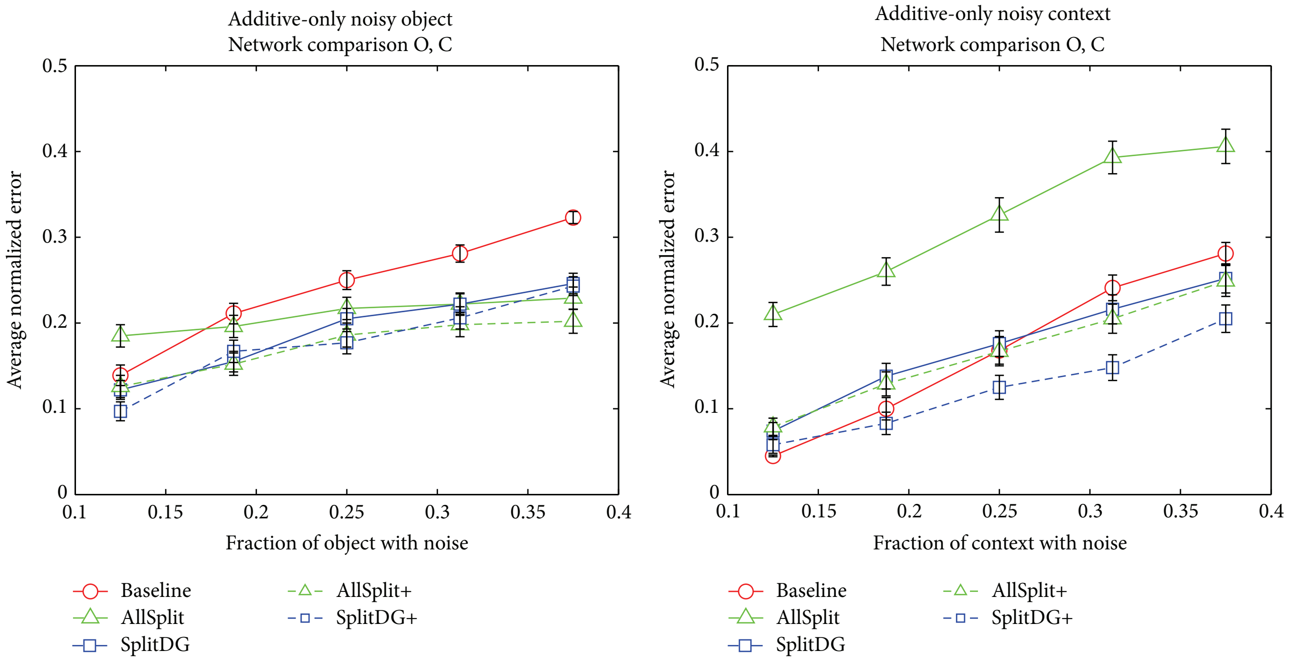
Figure 12 illustrates the effect of having additive-only noise in the object or context input layers. These tests are of the same nature as the partial input tests done previously and are designed to determine if there is any difference in how the networks deal with noise, and whether this allows more or less effective use of the crossconnection inputs. As with the partial object case, the AllSplit network performs well with object noise by using the additional context information available to its object stream to help it guess the object. In this case, the SplitDG and SplitDG+ networks also do better than the Baseline network and about the same as their AllSplit counterparts, though slightly worse in high noise situations. When the noise is in the context input, AllSplit does worse since it must deal with additional noise in its object representation. The larger DG and CA3 areas of the SplitDG and "+" networks clearly help with this task and bring performance on par with or even better than the Baseline network (in the case of SplitDG+), indicating that even if the context input is highly noisy, a large CA3 can extract enough additional context information to aid in object identification.

Figure 13 shows the results of the non-additive noise task. As in the additive-only task, the split networks perform better than the Baseline network when the object is noisy, with the AllSplit network performing better than SplitDG. When the context is noisy, the pattern is reversed, although SplitDG does just as well as the Baseline network.

4. Discussion

4.1. Anterior-Posterior Crossconnections. The results in Figure 6 suggest that a split network provides performance advantages compared to the Baseline network. Each output layer requires a different object to context input ratio in order to perform optimally on the relevant tasks. The object output layer gives the network's best guess as to what the actual object is, meaning it needs to perform well in low to medium noise and partial situations where either the object or context input (or both) is degraded. Surprisingly, additional contextual information is helpful even when that context is as noisy/incomplete as the object. This can be thought of as providing a "bigger picture" for the network to look at, and thus making it more likely that it can find some relevant clue which it can use to decipher the entire input. For example, suppose one is looking at a photograph of a person taken from a side angle so it is difficult to determine who it is (partial cue). If a wider-angle photo is now given which includes some of the person's body or clothing (partial context), this information gives a clue as to who the person is, even if the full context is unavailable. The same idea applies to noisy objects and contexts.

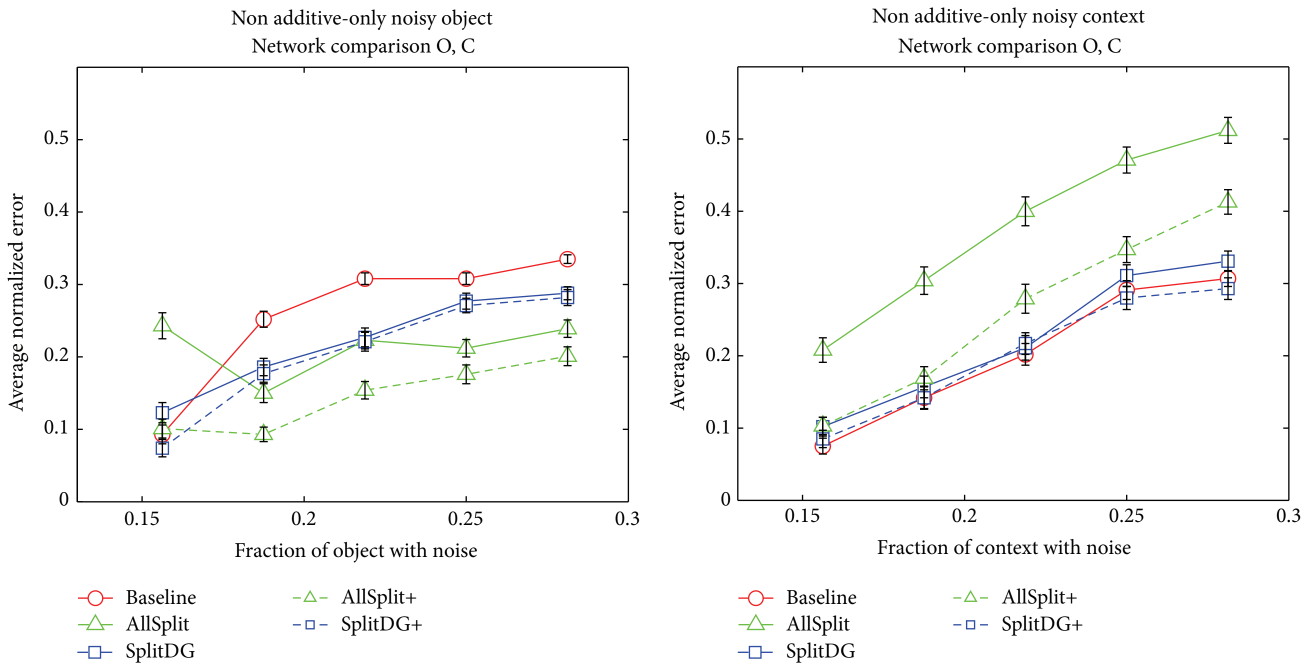
However, since each context contains several possible objects, the context input gives less information than the object input, and therefore its value (as far as the object output is concerned) decreases rapidly to zero with the amount of signal degradation. It is not a case when more information is beneficial regardless of how noisy it is. At some point, the error introduced by the noise outweighs the value of having additional information. If the object is presented noiselessly, then additional contextual information is not very useful, particularly if it itself contains noise. For the CA3 size used in our AllSplit network, this point of zero benefit occurs approximately when the context begins to have more noise or be more incomplete than the object. This is why, in the "partial context" and "noisy context" tests, we see the AllSplit network perform rather poorly with its relatively large amount of context input into the object stream (via the strong MEC connection). As we would expect, the more degraded the input context compared to the input object, the worse the AllSplit network performance. On the other hand,



(a)

(b)

Figure 12: Additive-only noise tests. (a) Error across networks, averaged over the O and C output layers, when noisy objects and noiseless contexts were presented as input. (b) Error across networks, averaged over the O and C output layers, when noiseless objects and noisy contexts were presented as input.



(a)

(b)

Figure 13: Nonadditive noise tests. (a) Error across networks, averaged over the O and C output layers, when noisy objects and noiseless contexts were presented as input. (b) Error across networks, averaged over the O and C output layers, when noiseless objects and noisy contexts were presented as input.

when the input context is less degraded than the input object, as in the "partial object" and "noisy object" tests, the AllSplit performance increases above that of the other networks. Again, because the context inputs have less absolute predictive value than the object inputs (for the object output layer) to begin with, the beneficial effect of noiseless context is less than the detrimental effect of degraded context, and as a result the noiseless context benefit does not come into play until object noise/partial levels are slightly higher. However, the beneficial effects can clearly be seen at moderate object noise levels, and for low noise levels the error is near the training threshold.

In all the networks, in the case of the context being particularly noisy/incomplete, the context output from the anterior stream may be too noisy for use. The hippocampal network would then turn to the object-based context guess output to deliver a context prediction, provided that the object input is relatively noiseless. Thus the OBCG layer needs to be effective in noisy/partial context and mismatch situations, which is what we test in Figure 6(b). In order to achieve good performance, the output must not use the MEC context input, since this layer will only be called on when the context is particularly noisy or incomplete. In addition, if the output relies too much on context, it begins to duplicate the functionality of the anterior context stream. Fortunately for the AllSplit network, this highly degraded or mismatched context situation in which C must be substituted with OBCG is also exactly the situation in which the object output fails; hence it may be able to conveniently rely on the OBCG layer's output to give it a reliable context to use. We have not implemented this backup functionality in our network.

The context output layer is similar to the object output layer in that it must be able to deal with noise in both object and context, and dealing with object noise is of higher priority (as it is with the object output) because the OBCG layer provides a backup in the case of high context noise. For the context layer, this means that it should have a small amount of object input relative to context input. Figure 6(c) shows that this naturally occurs thanks to the fact that there are much fewer contexts than objects, and thus the context stream is very effective at determining context even when they are noisy/incomplete. As a result additional object information is of little use to it, so the LEC to context stream input has less influence than the MEC to object stream input.

As with the context stream, the CBOG stream has fewer input-output associations to store; hence it relies less on the object input from LEC crossconnections. It is important that it depends mostly on context for the same reason that OBCG depends mostly on object, although the CBOG list may get called on even when the object input is usable, since it provides additional information that the object output cannot give. This layer provides a mechanism by which a list of objects can be recalled given only a single contextual cue. Networks consisting of only a single object and context output would not be able to model this task. One artificial feature of this output is that it is N times as large as the object output, where N is the number of objects per context (here 3). We are not implying that in the actual hippocampus, the region that distal CA1 on the anterior side projects to is N times as large or N times as active as the regions all the other CA1 areas project to. In the actual hippocampus these object outputs may come out one at a time, as the network activity has a time component in spiking networks. Since our model is strictly a rate-based connectionist model, the only way we can represent this output is as a single matrix in which all objects are represented at once. The OBCG output could also be represented this way, in the case where objects are allowed to appear in more than one context.

The temporal dynamics of context-based object retrieval in free recall situations have been given a theoretical foundation in the TCM (temporal context model) and CMR (context maintenance and retrieval) frameworks [40, 41]. Our model explicitly represents the biological structures and connections that make possible the basic multiple object to context associations (referred to as source clustering) assumed by these frameworks, but we do not attempt to provide a realization of any of the temporal aspects of memory (temporal clustering) which TCM and its generalizations also deal with, such as associations between successively presented contexts and the recency effect. However, allowing objects to be associated with more than one context (as they are in the case of the temporal context), our model could conceivably provide a starting point for a biological realization of the TCM framework. The varying internal context of TCM could be produced within our model by having objects output by CBOG feed back into the OBCG stream to produce an associated set of contexts, which would then be used as inputs into the CBOG stream to produce the next object to be recalled, in a repeated cycle.

4.2. Effects of Layer Size. There are two ways to approach the interpretation of the other test results, beginning with the training set error. The first way is to ignore the size of the network and compare only those networks that have similar amounts of error on the training set. In this view, a fair comparison would be between those networks that start out with equal amounts of knowledge on the training set, regardless of how many epochs it took them to get their error to that level or how many neurons they have. Here, splitting a layer into two separate sublayers has little to no disadvantage, because each sublayer is still large enough to do its task at the same level as the full layer. This has precedent in the cognitive psychology literature, where, for example, subjects being tested on recall of a list over time or in different contexts may be allowed as many trials as they need to memorize the list in the first place, so that all participants start out with the same low training error rate. This assumes that humans have enough neurons available to memorize the training list to whatever degree of accuracy is required, given enough time. In addition, it is known that in rats, during the course of a particular spatial task, only a small fraction of the hippocampal CA1 neurons fire during the entire duration of the task. This suggests that the hippocampus has many more neurons than necessary for any given task.

Of course, neurons cannot be added to actual test subjects, but in our test networks this provides an effective way to accomplish the same goal of reducing the error on the training set, so that all networks start with the same baseline error

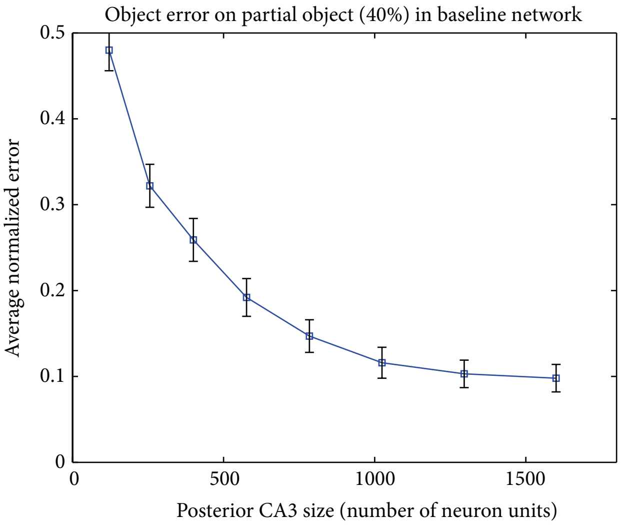


Figure 14: Error on object output layer in a 40% partial object task as a function of the size of posterior CA3 in the Baseline network.

rate. From the biological standpoint, this way of comparing networks essentially says that, in the actual brain, the memorization ability of the hippocampus for any particular task is not limited by the number of neurons, but rather the way in which they are connected. From this point of view, the basic AllSplit and SplitDG networks should be ignored, and the results of Baseline should only be compared against AllSplit+ and SplitDG+, since all three of these networks have the same error rate on the initial training set.

The second way to interpret the results is to take the neuron-limited view, where a fair comparison would be between networks which have the same number of neurons, regardless of how well they are able to store the initial training set. In this view, splitting a layer into two separate sublayers incurs the penalty of each sublayer now being half the size. Biologically, this means that neurons are costly in terms of energy required to build and maintain, and that the brain has as few neurons as possible while still being able to perform its required tasks. From this point of view, Baseline should be compared with AllSplit and SplitDG since they have the same number of neurons, and AllSplit+ and SplitDG+ should be ignored.

In the biological hippocampus, the answer probably lies somewhere between the two extremes. Figure 14 shows that increasing the size of CA3 in the Baseline network results in lower error rates, but that eventually the error stops decreasing with layer size. If the hippocampus is in the rightmost region of the graph, then it has enough neurons and there is little cost to splitting a layer, so it is best approximated by the "+" models. On the other hand, if it is near the leftmost region of the graph, it is severely neuron constrained, and splitting a layer results in a dramatic decrease in performance on each of the streams. In this case it would be better approximated by the normal (non-+) models.

Overall, the test results show that the AllSplit network is best for noisy or partial object situations and worst when given noisy or partial context. AllSplit+ has uniformly better performance as expected but follows the same general pattern as AllSplit. On the other hand, the Baseline network is relatively better at noisy or partial context situations than with noisy or partial object. Rarely is it the best network at any particular task, however, with the exception of partial context. It is most similar to the SplitDG network, which is what we expected based on its architecture. The SplitDG network has good all-around performance. Compared to Baseline, it does consistently better in noisy or partial object tests, about the same in noisy context, but noticeably worse when presented with partial context. SplitDG+ is generally about the same as SplitDG on noisy or partial object tests, but its larger DG seems to aid in the incorporation of context information when it is noisy or partial. This allows it to do significantly better than SplitDG in such tasks and puts it on par or better than Baseline. Our results thus suggest that differentiation within DG provides uniformly better performance over a nondifferentiated DG if it is large enough (SplitDG+), and generally better performance with the exception of partial context tasks if DG is size constrained (SplitDG). Additional differentiation within CA3 (AllSplit and AllSplit+) may work to further increase noisy and partial object task performance, but at the cost of the corresponding degraded context task performance.

4.3. Object Noise versus Context Noise. These results raise the question of whether it is better for the object stream to be able to deal with noisy objects (AllSplit) or noisy contexts (Baseline), where we will use the term "noise" to refer to partial cues as well. We argue that there is inherently less noise in contexts than in objects; hence dealing with object noise is more important. To make things concrete, consider the case of an animal in search of food. It has to find edible plants and insects and has to memorize a large amount of object-related information. Depending on the time of year and the time of day, the types of plants or insects it can eat and their appearance change (noise). On the other hand, the season and spatial environment are contextual cues that change slowly, and there are only a relatively small number of different contexts it must identify: its dwelling, its scavenging grounds, what season it is, and so forth. In general, the much larger number of objects in existence makes it likely that interference and noise are much more likely to occur between objects than between object and contexts, which are few in number and change only slowly over time.

The second argument is that, given some recurrent support structures, noise in context is easier for the hippocampus to deal with than noise in object. The context stream deals with context noise relatively well since the contexts are few and well memorized. Thus getting a clean context to the object stream requires only taking the context stream output (C) and feeding it back into the object stream. If the context is very noisy or absent (to the point that the context stream output is no longer useful), the output of the OBCG layer can be used instead. Thus there are two independent ways for the object stream to not have to deal with context noise, each involving only a recurrent loop.

With object noise, the situation is different. The object stream is itself responsible for determining the object; thus the only place it can turn to for additional object information is the CBOG output, which uses context to make object guesses. However, since the CBOG stream uses mainly context information, the best it can do is to give a list of possible objects that are associated with that context. Choosing one object out of this list would then require a separate calculation where the input object is compared with the CBOG output list and the best match selected. This would not be an easy task when the input object is noisy, although it would be significantly easier than the object stream's original task, which is to compare the input object to a list of 120 possible objects and choose the closest match. Thus the object noise problem can certainly be overcome with the help of additional structures, but it may be more judicious to simply use context information in the object stream from the beginning, which is exactly the solution that the AllSplit and SplitDG networks use. They then trade the object noise problem for a context noise problem, but this seems to be a much easier issue to deal with.

4.4. Mismatches. Mismatches, consisting of an object appearing in a different context from that it was learned into, are by definition rare events. If they happened frequently, the object would simply be associated with the new context and it would no longer be considered a mismatch. On the posterior side, a mismatch means that the incoming context information does not match the primary object input from LEC, thus putting it in a situation similar to having a very noisy context but noiseless object. On the anterior side, where MEC context information is primary, the incoming object input introduces uncertainty, and the situation is similar to a very noisy object but noiseless context. Due to the smaller number of inputs it needs to store and the fact that LEC input is relatively weak, mismatches have little effect on the anterior stream—if we see someone from the office at the mall, we do not have any trouble recognizing our context as the mall. On the other hand, the large amount of MEC input into the posterior stream means that a mismatched context can significantly affect object recognition—it may take us several seconds to recognize a colleague if we unexpectedly encounter them at the mall, whereas the recognition is nearly instantaneous when we see them at the office.

Any encoding and retrieval scheme which uses contextual information to recognize objects, as we believe the hippocampus does, will naturally have problems in mismatch situations. However, this is only the case if we believe that a familiar object in a different context from usual ought to still be recognized as the same familiar object. In many situations it may make sense to consider object A in context A as effectively different from object A in context B [42]. The large amount of error that a mismatch produces may be beneficial for signaling that something is wrong or unexpected and deserves our attention.

4.5. Relation to Rat Hippocampus. Our model is not explicitly a place field model, and in the way we have conceptualized it and in its current form our model better reflects the primate hippocampus. However, with some minor modifications the model would be consistent with the observation of higher-resolution place fields in dorsal compared to ventral hippocampus. We will switch to using the appropriate terminology for the rat anatomy in this discussion, so that anterior and posterior in our model are now ventral and dorsal, and the caudolateral and rostromedial bands of MEC and LEC are now dorsolateral and ventromedial, respectively.

In our model, for simplicity's sake, we make no distinction between the dorsolateral and ventromedial bands of the MEC, modeling both as carrying the same context information, albeit to different parts of hippocampus (dorsal versus ventral, resp.). However, it is known that neurons in the dorsolateral band of MEC are more spatially tuned than those in the ventromedial band [43], and thus we would expect that the dorsal hippocampus, receiving higher-resolution spatial information from the dorsolateral band, would have the tighter place fields that are seen experimentally. If we wanted to extend our model to cover this additional aspect of the anatomy, we could do this by having two different types of contextual inputs, a "local" context and a less precise "global" context which might represent the context at a larger spatial scale or contain some other nonspatial information, with the local context being carried by the dorsolateral MEC and the global context being carried by the ventromedial MEC.

Note that both the dorsal and ventral subdivisions of the hippocampus receive the nonspatial LEC inputs to some extent. However, we refer to the dorsal hippocampus as the more object-oriented layer in our model compatible with human fMRI studies and our set of sample tests (shown in Figures 6 and 7) which led us to set the relative weighting of the LEC input larger than that of the MEC input for optimal performance (and the reverse is true on the ventral hippocampus for context information). Of course, the set of "tests" that the rat hippocampus has evolved to do could be different from the basic tests that we proposed. For example, the performance on the mismatch test (where the presented object and context were not associated) was a significant factor in determining how strong the MEC to dorsal hippocampus connections should be. A strong MEC to dorsal connection results in a large amount of error on the OBCG output, and as a result those connections were kept very weak. In the rat hippocampus, however, it could be the case that it simply just does badly on mismatches because they are so rare that they do not need to be protected against with weak MEC to dorsal weighting, or it could be that in the case of mismatches, additional cortical processing is involved. In either case, the MEC to dorsal signal could well be just as strong or stronger than the LEC to dorsal signal.

In conjunction with the dorsolateral versus ventromedial band differences mentioned above, the dorsal and ventral streams of our rat-modified model would not contradict the general conception of the dorsal stream as being context oriented and more finely spatially tuned than the ventral side. In summary, the degree to which the MEC's spatial contextual information is relevant in the dorsal side of the rat hippocampus is probably much higher than that indicated in our model, where we look at objects, rather than context, as the primary information the hippocampus is storing and view context as information that can contribute to object recognition.

5. Conclusion

We constructed hippocampus models that include anatomical and functional details such as the distinction between the posterior and anterior subdivisions of the hippocampus, connections from the medial and lateral entorhinal cortex to both the posterior and anterior regions, differences between the superior and inferior blades of the dentate gyrus, and connectivity differences between distal and proximal (relative to DG) portions of CA3 and CA1. We hypothesized distinct roles for each of the CA1 areas on the proximal and distal sides and attempted to show how these anatomical details work together to increase performance on certain tasks. In particular, we showed that object and context require different treatment in terms of how much one is used to help recognize the other. This is simply due to the greater number of objects compared to the number of contexts rather than intrinsic differences in representation. In addition, we showed how the hippocampal anatomy supports the use of contextual information to help object recognition and proposed ways in which the tradeoffs inherent to this could possibly be mitigated.

Our models make several predictions that may be experimentally tested. We predict that the inferior blade of DG and proximal CA3 in the posterior region of hippocampus receives more MEC innervation, or that these neurons are more sensitive to MEC inputs, than is the case with LEC inputs into the anterior side of hippocampus. Blocking MEC input into posterior hippocampus should have a significant negative effect on object recognition when the object is noisy or only partially shown, assuming that the object was associated with a specific context, but should have only a mildly negative or even a positive effect if the context is noisy or obscured. Blocking LEC input into anterior hippocampus should have much less of an effect on context recognition in either case, assuming that there are many more objects than contexts. If the number of contexts and the number of objects are roughly equal, then we should see effects similar to those seen on the posterior stream with MEC input. Our assumptions about the two different types of information being carried along the output pathways can also be experimentally tested by comparing the information content of proximal CA1 and distal CA1 neurons. We predict that distal CA1 neurons on both the posterior and anterior sides will be more likely to carry object-type information, while proximal CA1 neurons will tend to carry primarily context-type information.

We found that the models that have only DG split (SplitDG and SplitDG+) did the best overall on our test sets, generally doing about the same as the Baseline model when the context input was degraded, and significantly better when the object input was degraded. The models with both DG and CA3 split (AllSplit and AllSplit+) did even better in noisy or incomplete object situations, but at a cost in performance on the corresponding degraded context tasks. As we mentioned in the discussion, it may be the case that degraded context situations are relatively rare compared to degraded object situations, and thus the performance tradeoff of the AllSplit networks may in fact be optimal. However, it is probably also the case that the hippocampus does not make as severe a tradeoff as we have in our models, where CA3 is either completely unified or completely split. For instance, both regions of CA3 in the actual hippocampus receive superior blade input from DG, rather than just the distal region. In our model, the superior blade on the posterior side of hippocampus carries mainly LEC object information, so including this feature may change the ratio of object to context information within proximal CA3 in favor of object information and thereby reduce some of the deleterious effects of noisy context that we observed in the AllSplit network. The two regions of CA3 also communicate to an extent, although they have different connectivity patterns in terms of the proportion of projections they send within CA3 and onward to CA1. Exactly how these differences affect hippocampal function remains a topic for future research.

To date, much of the computational literature on the hippocampus has either focused on only object memorization or only spatial context memorization and has not attempted to identify how these different types of information may mutually support each other within the hippocampus or elucidate specific anatomical details within the hippocampus that may allow this to occur. On the other hand, experimental literature that addresses details such as the LEC and MEC cross-connections has often assigned them only the vague role of allowing a mixing or integration of object and context information. We have hypothesized specific ways that object and context information may be used in the posterior and anterior regions of the hippocampus, shown that the connectivity of hippocampus supports and enables these uses, and identified specific situations in which these object-context interactions have a beneficial or deleterious effect. Our results thus suggest new ways of thinking about the sort of computations that the hippocampus may do, and how it uses both object and context to perform them.

Acknowledgments

This paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract no. D10PC20021. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the US Government. Additional support is provided by the National Institutes of Health (NIH) Grant GM084905: Computational and Mathematical Modeling of Biomedical Systems and by and NSF Robust Intelligence Grant IIS 1117388 (JMF).

References

[1] M. P. Witter, G. W. Van Hoesen, and D. G. Amaral, "Topographical organization of the entorhinal projection to the dentate gyrus of the monkey," *Journal of Neuroscience*, vol. 9, no. 1, pp. 216–228, 1989.

[2] J. K. Leutgeb, S. Leutgeb, M. B. Moser, and E. I. Moser, "Pattern separation in the dentate gyrus and CA3 of the hippocampus," *Science*, vol. 315, no. 5814, pp. 961–966, 2007.

[3] C. B. Alme, R. A. Buzzetti, D. F. Marrone et al., "Hippocampal granule cells opt for early retirement," *Hippocampus*, vol. 20, no. 10, pp. 1109–1123, 2010.

[4] W. Deng, M. Mayford, and F. H. Gage, "Selection of distinct populations of dentate granule cells in response to inputs as a mechanism for pattern separation in mice," *ELife*, vol. 2, Article ID e00312, 2013.

[5] L. M. Rangel and H. Eichenbaum, "What's new is older," *ELife*, vol. 2, Article ID e00605, 2013.

[6] R. C. O'Reilly and J. W. Rudy, "Conjunctive representations in learning and memory: principles of cortical and hippocampal function," *Psychological Review*, vol. 108, no. 2, pp. 311–345, 2001.

[7] A. Treves and E. T. Rolls, "Computational analysis of the role of the hippocampus in memory," *Hippocampus*, vol. 4, no. 3, pp. 374–391, 1994.

[8] M. P. Witter, "Intrinsic and extrinsic wiring of CA3: indications for connectional heterogeneity," *Learning & Memory*, vol. 14, no. 11, pp. 705–713, 2007.

[9] S. Zola-Morgan, L. R. Squire, and D. G. Amaral, "Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus," *Journal of Neuroscience*, vol. 6, no. 10, pp. 2950–2967, 1986.

[10] M. R. Hunsaker, G. G. Mooy, J. S. Swift, and R. P. Kesner, "Dissociations of the medial and lateral perforant path projections into dorsal DG, CA3, and CA1 for spatial and nonspatial (visual object) information processing," *Behavioral Neuroscience*, vol. 121, no. 4, pp. 742–750, 2007.

[11] M. R. Hunsaker, P. M. Fieldsted, J. S. Rosenberg, and R. P. Kesner, "Dissociating the roles of dorsal and ventral CA1 for the temporal processing of spatial locations, visual objects, and odors," *Behavioral Neuroscience*, vol. 122, no. 3, pp. 643–650, 2008.

[12] R. D. Burwell, "The parahippocampal region: corticocortical connectivity," *Annals of the New York Academy of Sciences*, vol. 911, pp. 25–42, 2000.

[13] C. B. Cave and L. R. Squire, "Equivalent impairment of spatial and nonspatial memory following damage to the human hippocampus," *Hippocampus*, vol. 1, no. 3, pp. 329–340, 1991.

[14] E. L. Hargreaves, G. Rao, I. Lee, and J. J. Knierim, "Neuroscience: major dissociation between medial and lateral entorhinal input to dorsal hippocampus," *Science*, vol. 308, no. 5729, pp. 1792–1794, 2005.

[15] S. Dennis and M. S. Humphreys, "A context noise model of episodic word recognition," *Psychological Review*, vol. 108, no. 2, pp. 452–478, 2001.

[16] S. Gaskin, A. Gamliel, M. Tardif, E. Cole, and D. G. Mumby, "Incidental (unreinforced) and reinforced spatial learning in rats with ventral and dorsal lesions of the hippocampus," *Behavioural Brain Research*, vol. 202, no. 1, pp. 64–70, 2009.

[17] M. B. Moser and E. I. Moser, "Functional differentiation in the hippocampus," *Hippocampus*, vol. 8, no. 6, pp. 608–619, 1998.

[18] R. E. Clark, S. M. Zola, and L. R. Squire, "Impaired recognition memory rats after damage to the hippocampus," *Journal of Neuroscience*, vol. 20, no. 23, pp. 8853–8860, 2000.

[19] M. N. De Lima, T. Luft, R. Roesler, and N. Schröder, "Temporary inactivation reveals an essential role of the dorsal hippocampus in consolidation of object recognition memory," *Neuroscience Letters*, vol. 405, no. 1-2, pp. 142–146, 2006.

[20] O. Hardt, P. V. Migues, M. Hastings, J. Wong, and K. Nader, "PKMζ maintains 1-day- and 6-day-old long-term object location but not object identity memory in dorsal hippocampus," *Hippocampus*, vol. 20, no. 6, pp. 691–695, 2010.

[21] D. G. Mumby, S. Gaskin, M. J. Glenn, T. E. Schramek, and H. Lehmann, "Hippocampal damage and exploratory preferences in rats: memory for objects, places, and contexts," *Learning and Memory*, vol. 9, no. 2, pp. 49–57, 2002.

[22] J. A. Ainge, C. Heron-Maxwell, P. Theofilas, P. Wright, L. De Hoz, and E. R. Wood, "The role of the hippocampus in object recognition in rats: examination of the influence of task parameters and lesion size," *Behavioural Brain Research*, vol. 167, no. 1, pp. 183–195, 2006.

[23] J. R. Manns and H. Eichenbaum, "A cognitive map for object memory in the hippocampus," *Learning and Memory*, vol. 16, no. 10, pp. 616–624, 2009.

[24] R. S. Rosenbaum, S. Köhler, D. L. Schacter et al., "The case of K.C.: contributions of a memory-impaired person to memory theory," *Neuropsychologia*, vol. 43, no. 7, pp. 989–1021, 2005.

[25] S. Corkin, "Lasting consequences of bilateral medial temporal lobectomy: clinical course and experimental findings in H.M," *Seminars in Neurology*, vol. 4, no. 2, pp. 249–259, 1984.

[26] K. A. Paller and G. McCarthy, "Field potentials in the human hippocampus during the encoding and recognition of visual stimuli," *Hippocampus*, vol. 12, no. 3, pp. 415–420, 2002.

[27] G. Fernández, H. Weyerts, M. Schrader-Bölsche et al., "Successful verbal encoding into episodic memory engages the posterior hippocampus: a parametrically analyzed functional magnetic resonance imaging study," *Journal of Neuroscience*, vol. 18, no. 5, pp. 1841–1847, 1998.

[28] C. E. Stern, S. Corkin, R. G. González et al., "The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 16, pp. 8660–8665, 1996.

[29] J. Poppenk, H. Evensmoen, M. Moscovitch, and L. Nadel, "Long-axis specialization of the human hippocampus," *Trends in Cognitive Sciences*, vol. 17, no. 5, pp. 230–240, 2013.

[30] A. Hupbach, O. Hardt, R. Gomez, and L. Nadel, "The dynamics of memory: context-dependent updating," *Learning and Memory*, vol. 15, no. 8, pp. 574–579, 2008.

[31] B. Jones, E. Bukoski, L. Nadel, and J. M. Fellous, "Remaking memories: reconsolidation updates positively motivated spatial memory in rats," *Learning & Memory*, vol. 19, no. 3, pp. 91–98, 2012.

[32] R. C. O'Reilly, R. Bhattacharyya, M. D. Howard, and N. Ketz, "Complementary learning systems," *Cognitive Science*, pp. 1–20, 2011.

[33] B. Aisa, B. Mingus, and R. O'Reilly, "The emergent neural modeling system," *Neural Networks*, vol. 21, no. 8, pp. 1146–1152, 2008.

[34] R. C. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind By Simulating the Brain*, The MIT Press, Cambridge, Mass, USA, 2000.

[35] V. Cutsuridis, B. Graham, S. R. Cobb, and I. Vida, *Hippocampal Microcircuits: A Computational Modelers' Resource Book*, Springer, 2010.

[36] E. J. Henriksen, L. L. Colgin, C. A. Barnes, M. P. Witter, M. B. Moser, and E. I. Moser, "Spatial representation along the proximodistal axis of CA1," *Neuron*, vol. 68, no. 1, pp. 127–137, 2010.

[37] G. Shepherd and S. Grillner, *Handbook of Brain Microcircuits*, Oxford Univ Press, Oxford, UK, 2010.

[38] H. Hayashi and Y. Nonaka, "Cooperation and competition between lateral and medial perforant path synapses in the dentate gyrus," *Neural Networks*, vol. 24, no. 3, pp. 233–246, 2011.

[39] P. Poirazi, T. Brannon, and B. W. Mel, "Pyramidal neuron as two-layer neural network," *Neuron*, vol. 37, no. 6, pp. 989–999, 2003.

[40] P. B. Sederberg, M. W. Howard, and M. J. Kahana, "A context-based theory of recency and contiguity in free recall," *Psychological Review*, vol. 115, no. 4, pp. 893–912, 2008.

[41] S. M. Polyn, K. A. Norman, and M. J. Kahana, "A context maintenance and retrieval model of organizational processes in free recall," *Psychological Review*, vol. 116, no. 1, pp. 129–156, 2009.

[42] L. Nadel, *The Hippocampus and Context Revisited*, Hippocampal Place Fields. Oxford Scholarship Online Monographs, 2008.

[43] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 152073, 14 pages
http://dx.doi.org/10.1155/2013/152073

Research Article

The Parietal Cortex in Sensemaking: The Dissociation of Multiple Types of Spatial Information

Yanlong Sun and Hongbin Wang

University of Texas Health Science Center at Houston, 7000 Fannin, Suite 600, Houston, TX 77030, USA

Correspondence should be addressed to Yanlong Sun; yanlong.sun@uth.tmc.edu

Received 20 December 2012; Revised 13 March 2013; Accepted 28 March 2013

Academic Editor: Giorgio Ascoli

According to the data-frame theory, sensemaking is a macrocognitive process in which people try to make sense of or explain their observations by processing a number of explanatory structures called frames until the observations and frames become congruent. During the sensemaking process, the parietal cortex has been implicated in various cognitive tasks for the functions related to spatial and temporal information processing, mathematical thinking, and spatial attention. In particular, the parietal cortex plays important roles by extracting multiple representations of magnitudes at the early stages of perceptual analysis. By a series of neural network simulations, we demonstrate that the dissociation of different types of spatial information can start early with a rather similar structure (i.e., sensitivity on a common metric), but accurate representations require specific goal-directed top-down controls due to the interference in selective attention. Our results suggest that the roles of the parietal cortex rely on the hierarchical organization of multiple spatial representations and their interactions. The dissociation and interference between different types of spatial information are essentially the result of the competition at different levels of abstraction.

1. Introduction

Sensemaking is a complex cognitive activity in which people make sense of or explain their experience or observations. Sensemaking is ubiquitous in humans' everyday life. Examples of sensemaking include medical diagnosis, scientific discovery, and intelligence analysis. Though it is plausible to argue that the core of sensemaking is abduction (a reasoning process that generates and evaluates explanations for data that are sparse, noisy, and uncertain), there is no doubt that sensemaking is not a primitive neurocognitive process. Rather, sensemaking is comprised of a collection of more fundamental cognitive processes (e.g., perception, attention, learning, memory, and decision making) working together, and certainly involves a group of brain systems from posterior regions to the prefrontal cortex.

According to the data-frame theory of sensemaking, people possess a number of explanatory structures, called frames, in which people try to fit the data into a frame and fit a frame around the data, until the data and frame become congruent [1–3]. Sensemaking is called a macrocognitive process in that it involves complex data-frame interactions (e.g., frames shape, define data, data recognize, and mandate frames), and therefore requires coordinated activities from multiple cognitive processes/systems, including attention, learning, memory, reasoning, and decision making. Whereas many different types of integrative processing models exist, the data-frame theory brings clearly into focus the emergence of the explanatory structures and the opportunity of learning in terms of extracting statistical regularities from the environment [4]. Such an approach makes the theory particularly appealing when the task environment is complex and people have to make decisions in the presence of multiple cues with a great deal of uncertainty.

Figure 1 describes a counter-insurgency surveillance example (hereafter COIN-AHA problem), in which an analyst is faced with a map that records the attacks from multiple enemy groups in the past. (For detailed modeling problems, see MITRE's Technical Report, In Press). The task is to estimate which enemy group would be more likely to be responsible for the new attack at the provided location (point of interest, POI). This task is clearly a sensemaking task. In particular, since the task stimuli are presented in a spatial environment, for effective sensemaking, different

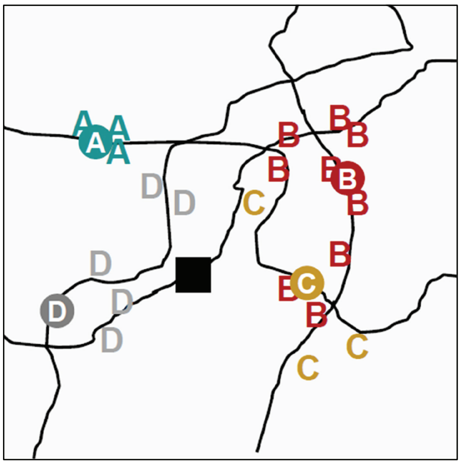


Figure 1: A typical scene in the COIN-AHA tasks. Attacks from individual enemy groups (labeled by "A", "B", "C", and "D" in different colors) are distributed along a road network. Subjects first need to estimate the radius and the center of gravity of the attacks from each enemy group. When a new attack occurs at the point of interest (POI, represented by a black square), subjects are asked to report the likelihood of each enemy group responsible for such an attack.

types of spatial properties of the environment would have to be acquired in the first place. For example, how many attacks have been carried out by each group (e.g., counting the number of objects from the visual inputs)? How large is the area that each group's attacks cover (e.g., perception of dispersion or size)? How close is the new attack to each group's active area (e.g., estimation of distance)?

Among the multiple cognitive steps in decision making, the parietal cortex (PC) has been implicated in various tasks for the functions related to spatial and temporal information processing, mathematical thinking, and spatial attention [5–8]. In the context of understanding the sensemaking processing in the COIN-AHA tasks, all these functions are certainly relevant. In this paper, we report a computational model to simulate the various functions of the parietal cortex in sensemaking (hence the PC module). In doing so, we hope to provide an integrated theory of the parietal cortex in spatial-temporal processing.

2. Value Representation in the Parietal Cortex

The central theme in modeling the parietal cortex in the COIN-AHA tasks is the estimation, representation, and integration of values based on the magnitudes of various spatial properties such as numerosity, group center, distance, and probability. Before we dive into the modeling details, we first discuss the unique role of the parietal cortex in value representation in the broader context of judgment and decision making.

First of all, in most accounts of decision theories, value representation is considered as the essential component in the decision-making process. To an extreme extent, the entire process of decision making is about value representations (e.g., [9]). Vlaev et al. [10] summarized three types of decision theories. The approach of "Type I", value-first decision making, is based on independent and absolute value scales (e.g., [11]). "Type II", comparison-based decision making with value computation, is based on comparison of values where subjective magnitude representations are context dependent (e.g., [12]). "Type III", comparison-based decision making without value computation, has no explicit psychoeconomic scales, and decisions can be reached at by binary comparison, for example, by the "priority heuristic" (e.g., [13]). Despite the different flavors, all the three types of the decision theories have to rely on some form of value representations. The difference is only on the specific forms and stages of value representation in decision making: for example, whether the value representation stably leads to a decision (Type I) or is modulated by contextual information (Type II), or whether the value is on a cardinal scale (e.g., number or magnitude-like in Types I and II) or an ordinal scale (e.g., binary comparison in Type III). Neurologically, we are interested in the neuronal correlates of values in decision making in addition to the value representation itself. It has been reported that the neuronal correlates of various types of values exist in numerous regions, such as the orbitofrontal cortex, parietal cortex, posterior cingulated cortex, dorsolateral prefrontal cortex, premotor cortex, and frontal eye fields (for reviews, see, [9, 14]). Thus, it is critical to examine different types of value representations depending on the purpose and domain of the brain function.

Compared with other brain regions, the parietal cortex plays a unique role in transforming the spatial and temporal information from the environment, such as time, distance, speed, size, and numerosity, into magnitude-like value representations [5, 15]. Neuroanatomically, the parietal cortex receives projections from multiple sensory modalities, including visual, somatosensory, and auditory. In addition, it receives inputs from the subcortical collicular pathway, which consists of the superior colliculus and the pulvinar and is thought to be closely related to spatial orienting and eye movement control [6, 16, 17]. Most significantly, the parietal cortex has been identified as part of the dorsal "where" pathway [18]. It has been indicated that there is *a common metric* of time, space, and quantity representations residing in the parietal cortex because of the need to learn about the environment through motor interactions and to encode relevant variables for action [5, 15]. This pattern of cortical connections makes the parietal cortex an ideal system for integrating and extracting spatial information from multiple modality-specific and unstable sensory channels and achieving supramodal and more stable spatial representations.

Although the encoding of values could be relevant in all stages of decision making and exhibit neuronal correlates in numerous brain regions, the value representation in the parietal cortex is unique in that it is confined by its *proximity* and *specificity* [7]. For example, fMRI studies revealed that when participants were instructed to compare number, size, and luminance, the activation of the left and right intraparietal sulci (IPS) shows a tight correlation with the behavioral-distance effect [7, 19]. Whereas hippocampal and parahippocampal regions are clearly involved in spatial cognition, they do not possess the close proximity of spatial and numerical representations as the parietal cortex does. Although frontal regions are involved in both spatial and numerical tasks, parietal activations are related to a more restricted set of cognitive processes. Such specificity probably is most evident in the comparison between the orbitofrontal cortex (OFC) and the lateral intraparietal area (LIP). In general, the values represented by LIP neurons are more subject to modulation of responses encoding the spatial properties of the visual stimuli [9]. Although there is evidence that neurons in LIP are sensitive to probabilistic classification, it seems that such a sensitivity is limited to the simple integration of visual properties (e.g., combination of shapes) [20]. In contrast, neurons in OFC represent the value of goods per se (probabilities, rewards, etc.), independently of how goods are visually presented [9]. Crucially, the bilateral horizontal segment of the intraparietal sulci (HIPS) that are consistently activated in arithmetical tasks in humans roughly coincides with the putative human ventral intraparietal area (VIP), and such an overlap between comparison processes and spatial networks in the IPS is believed to account for the behavioral interactions between representations of number and space [7]. In sum, the parietal cortex, and the IPS in particular, might be the first cortical stage that extracts visual numerical information from visual inputs [8].

Another aspect of the specificity in the parietal cortex's spatial processing comes from the selection of frames of references (FOR). While spatial representations prior to the parietal cortex are typically retinotopic, spatial representations in the parietal cortex have been transformed and are generally egocentric. In putative human homologues of macaque IPS regions, LIP represents target position in an eye-centered frame of reference and is involved in spatial updating. Ventral intraparietal (VIP) represents targets in a head-centered frame of reference, and anterior intraparietal (AIP) represents space in hand-centered coordinates [7]. According to theory of frame of reference-based maps of salience (FORMS), the parietal cortex subserves spatial representations using a range of egocentric frames of references (e.g., eye centered, hand centered, and body centered, etc.) so as to allow rapid actions [21–23]. In addition, intrinsic representations, which represent between-object relations using a world-centered frame of reference but often involve some degree of perspective taking, are also encoded in parietal cortex, especially the posterior parietal cortex [24]. Furthermore, the values encoded by the parietal cortex tend to be at the approximate level rather than exact. For example, it has been indicated that topological comparison and approximate metrics are encoded within the parietal cortex, and exact spatial metrics are encoded in hippocampus [25].

In accordance with theories described above, we have designated the PC module to be responsible for (1) extracting and representing relevant spatial information (i.e., providing relevant data from frame-matching such as radius, group center and two types of distances); (2) providing mechanisms for shifting attention during the process (i.e., defining and shaping data collection through both top-down and bottom-up modulations). Specifically, the PC module is responsible for processing the following spatial information (see Figures 1 and 2):

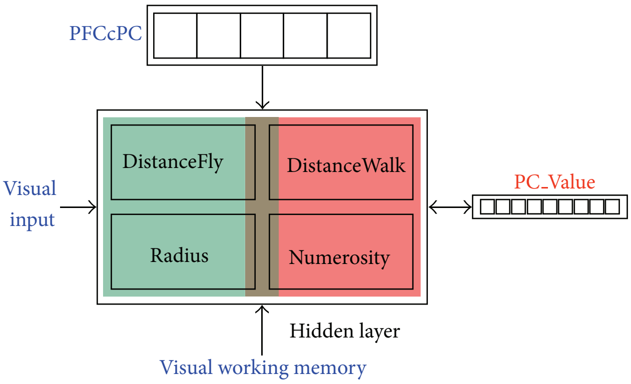


Figure 2: A schematic depiction of the metric estimates in the PC module. A common hidden layer takes inputs from both visual field and visual working memory, and outputs a magnitude value as the "PC value". There are two basic metrics being encoded on the hidden layer: numerosity (in red area, including the number of group attacks and the walking distance on a road segment) and size (in green area, including the radius of group attacks and the flying distance between the group center and POI). The overlap of red and green areas represents the possible overlapped functionality of numerosity and size. The role of PFCcPC (PFC controls PC) is to provide an "attentional prioritization" by enhancing the contrast and specialization on the hidden layer. At any time, only one of the units on PFCcPC is active and projected onto the corresponding section on the Hidden layer. As a result, that section is more active than other sections and more likely to win the competition.

(1) estimating the group center, the centroid of a cluster of attacks from a particular enemy group;

(2) estimating the dispersion ("Radius"), the two-to-one radius that spatially covers two-thirds of the attacks from a known enemy group;

(3) estimating two types of distances between the enemy group center and the point of interest (POI): "DistanceFly" represents the Euclidean distance (as crow flies) and "DistanceWalk" represents the length of the road segment (as cow walks);

(4) estimating the number of attacks from each enemy group ("Numerosity"), which will later lead to the base rate comparison (i.e., the percentage each enemy group takes in the total number of attacks).

In the following, we discuss the implementation steps and the corresponding psychoneurological justifications in detail.

3. Inputs: Perceptual Grouping and Segmentation

The first step in modeling the parietal cortex functions in the COIN-AHA tasks is to group the various visual representations (e.g., individual attacks, POI, road) onto different input

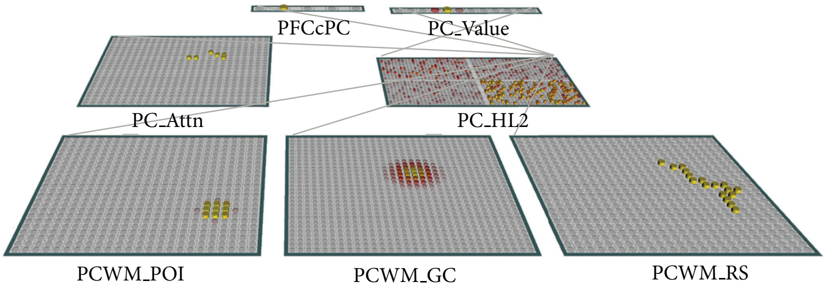


Figure 3: The parietal cortex module for computing target values on Numerosity, Radius, DistanceFly, and DistanceWalk (represented on the target layer "PC_Value"). The hidden layer "PC_HL2" is "sliced" into 4 groups, each responsible for a different type of target values. The competition on PC_HL2 is achieved by a kWTA function (k-Winners-Take-all, [26]), which is a combination of within-group inhibition (only the most active units within a group can contribute as the layer output) and entire-layer inhibition (units within a relatively weaker group are more likely to be inhibited). Such a function allows both of the dissociation and interference between different types of target computations. In the bottom-up information flow, input for both Numerosity and Radius is represented on PC_Attn (group attacks on visual field); Input for DistanceFly is based on a direct comparison of PCWM_POI (point of interest) and PCWM_GC (group center). Input for DistanceWalk is represented on PCWM_RS (road segment between POI and group center). The top-down control from layer PFCcPC represents 4 types of magnitude computation (Numerosity, Radius, DistanceFly, DistanceWalk) (the fifth unit is tentatively reserved for topological comparison). At any time, only one type of the magnitude values is available at the output level. The example shown here illustrates the case when the top-down demand is to compute Numerosity (second unit on PFCcPC), such that the bottom-right section on PC_HL2 is more likely to win over other sections in kWTA inhibition.

layers. On the one hand, our modeling focus is on the higher-order functions of value representation rather than the low-level visual processing. On the other hand, the encoding of values in the parietal cortex is heavily driven by the spatial and temporal properties of the visual inputs. To strike a balance, we made several simplifications in organizing the input layers to the PC module.

To represent the multiple attacks from enemy groups (e.g., attacks labels "A," "B," "C," and "D" in Figure 1) within the parietal cortex, our modeling approach is to represent the multiple attacks from a single enemy group as a whole on the visual input layer ("PC_Attn", see Figure 3), separated (segmented) from the attacks from other enemy groups. Then, both *numerosity* and *radius* can be computed based on PC_Attn. Next, the group center is computed as the *center of gravity* (i.e., arithmetic means of x and y coordinates of individual attacks) and represented on layer "PCWM_GC." The point of interest (POI) is represented on a separate input layer "PCWM_POI" (with lateral activations such that an object is displayed as a Gaussian bump). Then, DistanceFly (the distance "as crow flies") is computed as the Euclidian distance between the group center and POI. To compute DistanceWalk (the distance "as cow walks"), we represent the road segment between the group center and the POI on the input layer "PCWM_RS". Then, the estimation of the walking distance is in effect to estimate the length of a curved line segment, which is equivalent to numerosity counting based on the number of activated pixels on PCWM_RS, regardless the topographic distribution of individual pixels.

It is noted that in the current model, we have avoided the problem of finding the shortest path. Instead, we focus on the problem of length estimation when a road segment is explicitly provided (i.e., the walking distance). In representing the road segment as a separate visual input, our justification is that a curved line segment can be recognized and maintained as a single visual input component. Ungerleider and Bell [17] suggest that in identifying and discriminating the primitive "geons," neuronal selectivity progresses from simple line segments (in V1) to simple curves (in V2), to complex curves or combination of curves (in V4 and posterior IT cortex). In addition, it has been suggested that attention operates on object-centered as well as on location-based representations in that two connected objects (e.g., a barbell) may be represented as a single continuous object [27].

Apparently, estimating the walking distance between two objects will be affected by the curvature (the curves on the road) and connectedness (whether two objects are connected by the road). Regarding the curvature (Figure 4(a)), it has been suggested that the "sagitta" provides the best cue in accounting for the discrimination of pairs of long-duration, curved-line stimuli, over a range of one- and two-dimensional transformations, and the contour curvature was coded in terms of just two or three curvature categories, depending on curved-line orientation [28]. Regarding the connectedness (Figure 4(b)), Sun and Wang [21] found that object pairs connected or anchored to the same landmarks are easier to recall than those anchored to different landmarks. Together, these studies suggest that people are to a certain extent sensitive to the variations in curvature and connectedness. Then, by representing the road segment on a single layer, the curvature and connectedness are in effect implicitly encoded in the visual input. For example, a curvy road segment would be longer than a straight one, and two points directly connected by the same road would be closer than connected by different roads because of more curves. Then, estimating the walking distance along the road effectively becomes a task of numerosity estimation (i.e., counting the number of active units on the line segment), resulting in a nonverbal representation of magnitude and number sense housed in IPS [7, 29].

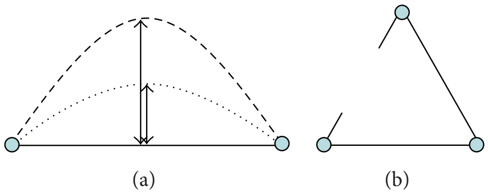


Figure 4: Distance adjustments by curvature and connectedness. By representing the road segment on a grid, the adjustment by curvature and connectedness is transformed into the task of numerosity estimation. For example, a more curvy (greater sagitta) road segment presents more active pixels on the grid (a), and a disconnected object pair requires additional routes to be connected and results in more active pixels on the grid (b).

It is noted that our method of representing various inputs to the PC module is mostly guided by the principles of *selective attention*. In particular, the PC module may receive multiple perceptual inputs in parallel from both direct visual input (layer "PC_Attn") and visual working memory ("PCWM_" layers), but the total number of input layers is limited. This is due to the consideration that when the computation of target values requires selective attention (e.g., paying attention to one particular enemy group), it generally suffers a bottleneck that poses more strict limitations on the processing capacity (e.g., [30, 31]). In addition, we also considered the constraints to display resolutions on the input layers. It appears that the superior intraparietal sulcus (SIPS) could be the candidate for providing inputs from visual working memory, with a high resolution but a limited number of slots [32]. Also note that at the current stage of modeling, the assignment of whether a particular input is directly from visual field or from visual working memory is rather arbitrary. In reality, it is likely that the assignment will be dependent on the temporal sequences of visual stimuli or on specific strategy usages by individual subjects.

Another critical issue in organizing the input layers is to consider the principles of *perceptual segmentation* (e.g., to single out a particular set of objects from others) and *attentional foveation* (e.g., multiple scans in evaluating a large number of objects or estimating the distance across a wide range of visual field). We argue that separating a single enemy group from others (e.g., group attacks on PC_Attn) and representing a cluster of spatially distributed objects as a single object (group center on PCWM_GC) are essentially the results of these principles. The guideline is that such representations can be obtained and maintained in early visual processing, especially when the different groups of objects are displayed in different colors and can be easily distinguished from each other. Strong claims have been made based on the efficient detection of groups of image elements by selective neurons that occurs in higher areas of the visual cortex [33, 34]. Using a task of transsaccadic integration (TSI) in which participants used a mouse to click on the intersection point of two successively presented bars, Prime et al. [35] found indistinguishable performance in the "Saccade" condition (bars viewed in separate fixations) and the "Fixation" condition (bars viewed in one fixation) and concluded that participants can retain and integrate orientation and location information across saccades in a common eye-centered map in occipital cortex. From the perspective of attentional foveation, it is proposed that the dorsal stream (posterior parietal and lateral premotor cortices) plays the role of serial deployment of attention over different locations of space and/or time, such that the encoding of magnitude is abstract enough to respond to both sequential and simultaneous presentations [36, 37]. Together, the parietal cortex may receive multiple visual inputs in a rather flexible fashion. During our simulations, we have indeed found that different visual input formats can result in indistinguishable performances (Figure 5).

4. Output: A "ScalarVal" Representation of Magnitude

Currently, our PC module uses a "ScalarVal" type of Leabra layers to represent a magnitude value ("PC_Value" in Figure 3) [26, 38]. (For a detailed description of the ScalarVal specification, see http://grey.colorado.edu/emergent/index.php/ScalarValLayerSpec). Such a specification encodes and decodes scalar, real-numbered values based on a coarse coded distributed representation across multiple units (e.g., a value is represented by a Gaussian bump with a fixed standard deviation). This provides a very efficient and effective way of representing scalar values [39, 40].

On a related note, there has been an ongoing debate regarding whether magnitudes are being internally represented on a linear scale or a logarithmic scale (e.g., [41]) (see Figure 6). By linear encoding, the noise (i.e., standard deviation) in the internal representation of a magnitude is tied to the specific value of the physical magnitude. Then, in comparing two magnitudes m_1 and m_2, the discriminability (i.e., the amount of the overlap between two Gaussian distributions) is determined by the Weber fraction w, and the standard deviations that are tied to the specific values of the magnitudes (with a pooled standard deviation). By logarithmic encoding, the noise in the internal representation of *any* magnitude is solely determined by the Weber fraction. Discriminability is determined by the Weber fraction w and the ratio of two magnitudes $r = m_1/m_2$, regardless of the specific values of the magnitudes (Weber's law). In our opinion, the logarithmic encoding appears to be a more appealing candidate that makes the representation of a magnitude truly abstract and with generality. It should be noted that the linear and logarithmic representations are mathematically equivalent but have different advantages during actual computation (e.g., linear models are more

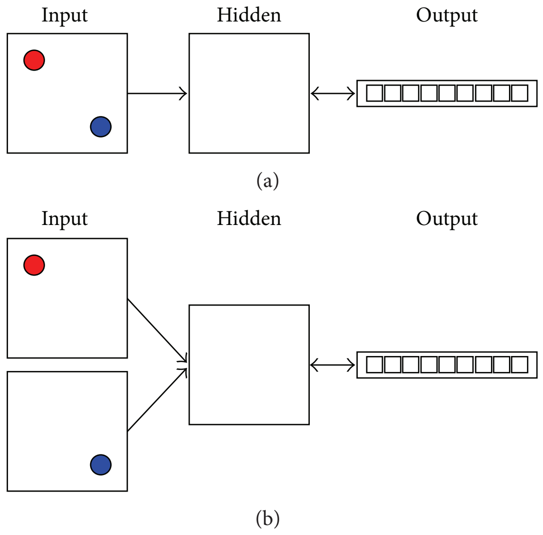


Figure 5: Two different configurations for providing the same spatial information to the PC module. (a) Two objects (e.g., a group center and a POI) are presented on the same input layer. (b) The same object pair is presented on two separate layers. When the task is to compute the Euclidean distance (flying distance) between the two objects, these two configurations yield indistinguishable performance. To compare the model performance, we computed the model-target correlation (correlation between the output values at the minus phase and the target values across trials). In both configurations, after training for 1000 epochs (20 trials in each epoch), the model can produce a model-target correlation greater than 0.95 in the last 10 epochs ($n = 200$ trials).

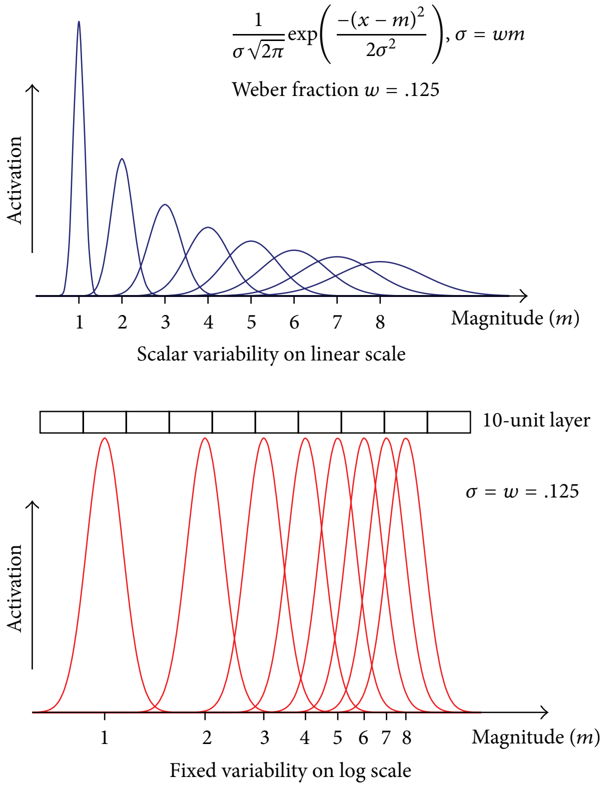


Figure 6: Comparison of the linear and logarithmic representation of magnitude. The method implemented in the PC module is analogous to the log scale representation in which we use a fixed number of units (thus fixed variance) to represent any particular level of magnitude.

convenient for addition and subtraction, and log models are more convenient for production and division). Because of the mathematical equivalence, it remains difficult to neurologically distinguish the actual representation form in the brain [42]. Nevertheless, the logarithmic representation appears to be more parsimonious in that the representation of a magnitude is independent of the range of the target values thus allowing different neurons representing different numbers to be activated in the same fashion (see Figure 6). In this regard, the default ScalarVal specification in Emergent serves our modeling purpose well.

5. Numerosity and Size on a Common Metric

The most significant aspect of the current PC module is that the computation of all types of target values (numerosity, radius, and two types of distances) largely shares a common pathway (see Figure 2). First, multiple input layers (e.g., individual group attacks, group center, POI, and road segment from direct visual and visual working memory slots) are projected onto different groups within a single hidden layer, depending on the particular task demand. This hidden layer employs a particular type of kWTA inhibition in that the winners are selected based on a combination of within-group and entire-layer inhibition (Figure 3). Its functions are analogous to those of the LIP area in that the spatial information is reencoded, *sensitively but not selectively* corresponding to the magnitude statistics from the visual environment. For example, a unit's activation may be statistically correlated with the number of active units on the input layer (i.e., sensitive to numerosity), but such a correlation on the hidden layer may not uniquely identify a number before being classified on the target layer (i.e., selectivity). In addition, the hidden layer receives a top-down signal from the layer "PFCcPC," representing a single task demand for a particular type of target values ("PFCcPC" means "prefrontal controls parietal"). At the output level, the desired target value is represented on the single target layer "PC_Value," analogous to the VIP area whose value representation *selectively* corresponds to the specific magnitude information in the visual environment. Computationally, the learning of target values occurs in two phases, an expectation-driven *minus phase* and an outcome-driven *plus phase* [26]. During the minus phase, the inputs (visual inputs plus the signal from PFCcPC) are reencoded onto the hidden layer and the target layer. During the plus phase, a teaching signal is provided on the target layer, which will provide a top-down correction by modifying the activations on the hidden layer and the corresponding weights. (For detailed descriptions of the learning rule (i.e., the extended contrastive attractor learning rule, XCAL), see

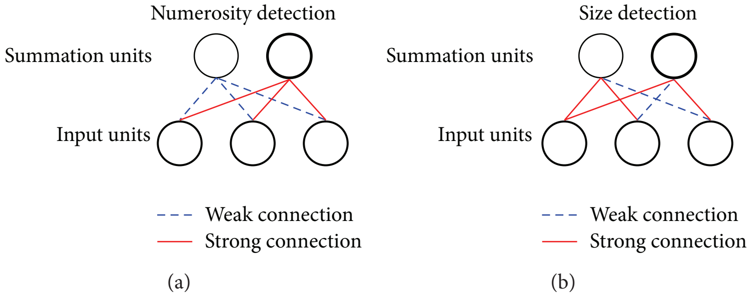


Figure 7: An illustration of numerosity detection (a) and size detection (b). In both figures, each unit on the hidden layer takes the sum of the visual input units, hence the name "*summation units*". For simplicity, we show only three units as the visual inputs and two units on the hidden layer. In representing numerosity, a summation unit takes the sum from each visual input unit *uniformly* (i.e., with equal connection weights). Thus, the activation of such a unit only responds to numerosity monotonically, and the spatial information is completely discarded. Different summation units have different connection weights from visual inputs, and their combined activation pattern is projected to the final tuned numerosity detectors that are ultimately selective to numerosity. In contrast, the summation units for encoding size (or distance between the two furthest active visual units) must receive nonuniform weights *selectively* from different spatial locations in the visual inputs (i.e., with unequal connection weights) in order to preserve the spatial information.

http://grey.colorado.edu/CompCogNeuro/index.php/CCN-Book/Learning).

The implementation of a common pathway for all types of target values is motivated by the following considerations. First of all, the recent literature suggests a "common metric" in parietal cortex responsible for the processing of all magnitude-like values such as numerosity, size, and temporal and spatial distances, namely, the temporal-spatial number line [5, 15, 43]. Second, although the projections from the input layers onto the various sections of the hidden layer are essentially in parallel, there is only one target layer. The rationale is that the perceptual stages operate in parallel but a central decision stage occurs via a serial bottleneck [44]. Third, the top-down control from PFCcPC reflects the idea that attention prioritizes stimulus processing on the basis of motivational relevance, and major sources of top-down attentional biasing have long been located principally in the dorsolateral prefrontal and posterior parietal cortices [45]. Also, the top-down connections from both PFCcPC and PC_Value to the hidden layer are consistent with the findings that the same neurons in LIP that encode values would also encode the selected actions late in the decision process [46].

Most importantly, the PC module addresses the dissociation and interference between various types of target values. Theoretically, there has been an ongoing debate regarding the interactions such as those between the processing pathways of numerosity, size, and density. On one side of the debate, it has been suggested that numerosity could only derive indirectly from texture density (e.g., [47–49]). On the other side, it has been suggested that numerosity could be an attribute "sensed directly" from the visual input, independently from texture perception [50, 51]. Most recently, Stoianov and Zorzi [52] shows that selectivity to visual numerosity emerges naturally during unsupervised learning in a hierarchical generative model of perception, invariant to area, density, and object features. This study has been cited by Ross and Burr [51] as a strong support to their theory of "visual sense of numbers."

In the PC module, among the four types of target values, there are actually only two basic types of information being extracted from the visual environment: *numerosity* and *size*. As mentioned in the previous section, estimating the walking distance on a road segment is in effect a task of counting the number of active pixels. In addition, estimating the flying distance is in effect a size or radius estimation in which the number of objects is a constant of two, regardless whether the inputs are represented on a single or separate layers (see Figure 5).

It is important to note that the dissociation and interference of numerosity and size may occur at different levels of visual analyses. We hypothesize that the key in both of dissociation and interference lies in the mechanism in which neurons selectively or uniformly sample the visual field and whether the spatial information is discarded or preserved during the sampling (see Figure 7). In the case of numerosity representation, it has been found that there were "summation units" in the parietal lobe, particularly in the LIP area, whose responses resembled the output of accumulator neurons that systematically increased or decreased with the increase of the numerosity in visual stimulus [53]. And, there were "number neurons" tuned to a preferred numerosity with "labeled-line" encoding of numerosity in the VIP area [54–57]. Thus, similar to some of the previous models on numerosity [52, 58, 59], our approach to modeling both numerosity and size estimation is to assume that the final tuned magnitude detectors on the target layer "harvest" the activations from the preceding summation units on the hidden layer. In order to selectively respond to the numerosity information, the spatial information must be discarded (e.g., the number sense of "2" arises regardless how far two objects are apart from each other). One immediate way to achieve such a dissociation is to assume the numerosity summation units samples the visual field uniformly (with approximately equal connection weights), regardless the spatial locations (see Figure 7(a)). This kind of uniform sampling has been demonstrated by [59].

On the other hand, spatial location information has to be preserved in size detection, which implies that the summation units must selectively cover different locations in their receptive fields (see Figure 7(b)).

What is interesting is how the interference between numerosity and size can arise when the spatial location information is only partially discarded or preserved. It has been found that single neurons tuned to quantity can provide information about only a restricted range of magnitudes, and only the population of selective neurons together can account for the entire range of tested stimuli [60]. Thus, it is likely that the receptive field of individual *numerosity summation units* on the hidden layer is spatially segmented and they are selective to a limited region of space, especially in a high-load condition (e.g., when the scene is crowded or subjects are distracted). Otherwise, responding to a greater range of numerosity would require finer graded activation levels thus overburden the summation neurons. As a result of the spatial segmentation, the activation of these neurons would partially carry the location information from the visual inputs. On the other hand, some of the *size summation units* may take visual inputs less selectively regarding different spatial locations. In either scenario, we would expect that the numerosity-size dissociation by the summation units is not perfect (i.e., carrying partial spatial information) and observe some neurons serving a double duty on both numerosity and size detection. We can find support to such a speculation repeatedly from both neurological (e.g., [5, 8, 54] and behavioral studies [48, 49]). Particularly, in their transcranial magnetic stimulation (TMS) experiments, Kadosh et al. [61] have found that the interference between number and size is late in the processing stream, at the point of response initiation and interaction between the stimulus attributes only in high-load conditions. And, it has been proposed that the numerosity and size estimations, and their overlaps, arise as the results of the serial deployment of attention over different locations of space and/or time via the dorsal stream (posterior parietal and lateral premotor cortices) [36, 37].

6. Topological Comparison and Representativeness

Although the main scheme in our PC module is metric estimations in a serial fashion (i.e., only one type of metrics is available at a time at the output level; see Figure 2), it should be emphasized that some bottom-up processing may indeed have occurred in a parallel fashion, resulting in behaviorally relevant representations in the process of decision making. In particular, the global nature of perceptual organization of spatial information has been described in terms of topological invariants, prior to the perception of other featural properties; that is, the processing of topological information may occur earlier than any metric estimation (e.g., [62]). Moreover, It has been suggested that posterior parietal cortex (PPC) supports topological spatial information which emphasizes the importance of proximity of local landmark cues, whereas the hippocampus supports metric spatial information which emphasizes the importance of distance between local landmark cues [25, 63, 64].

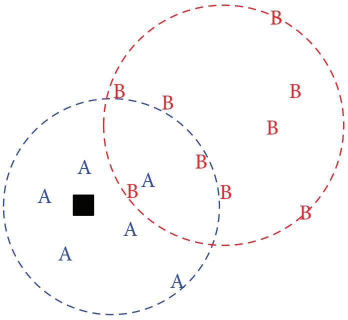
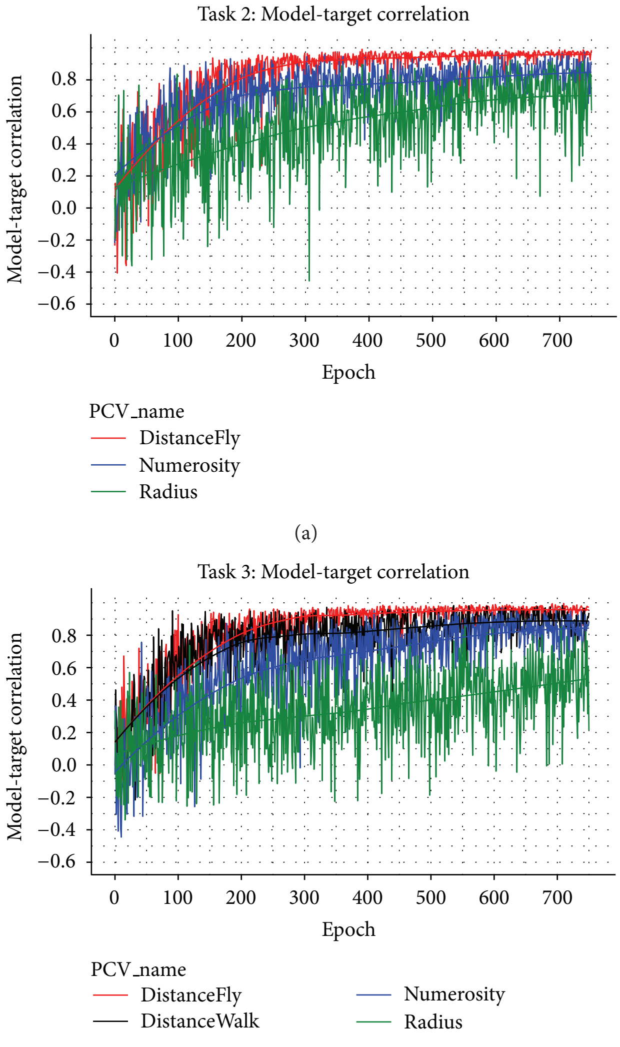


Figure 8: Topological comparison in an example trial in COIN-AHA task. The point of interest (POI, the black square) falls *inside* the region of Group A's attacks, but *outside* the region of Group B's attacks. Thus, without estimating the more abstract spatial information such as distance and numerosity, the POI might be perceived as more representative of the spatial characteristics of Group A's attacks than that of Group B.

Whereas the metric information is defined as the relationship of angles and distances between objects resulting in a continuous representation of values (e.g., radius and distances in the COIN-AHA problems), the topological relationships are represented by a connectedness relationship between objects that are invariant of metric modifications resulting in a categorical representation of values [25]. If topological comparisons indeed have occurred earlier than metric estimation, it would be very plausible that they are utilized in the decision-making process, especially as the means of shortcuts in the early stages. For example, it has been reported that expert geographers organized their thoughts and presented data to others with the topological information [65–67].

Perhaps more significantly, modeling the topological comparison would enable us to examine the *representativeness heuristic* that might arise at the level of perceptual analysis (Figure 8). It should be noted that the term of representativeness heuristic has been coined more than three decades ago [68]. Here we take a more updated interpretation described by Kahneman and Frederick [69]. According to this interpretation, both of the representativeness and availability heuristics in effect belong to the heuristic of accessibility and substitution, where an individual assesses a specified target attribute of a judgment object by substituting another property of that object—the heuristic attribute—which comes more readily to mind. Applied to the COIN-AHA tasks, it is possible that when human subjects perceive that the POI falls inside the region of Group A but outside the region of Group B, they might conclude that this particular POI is more representative of Group A's characteristics than that of Group B. Consequently, they might draw a conclusion that Group A is more likely to be responsible for the attack. That is, a decision can be made by taking a shortcut where the topological relation is used as a heuristic attribute to substitute a metric estimate of distance which only arrives later.



(b)

Figure 9: Performance of the PC module in Task 2 (a) and Task 3 (b) measured by model-target correlation. In both figures, the top-down PFCcPC-to-Hidden connections are specified as "group-one-to-one." In such a connection, each unit on PFCcPC is connected to all units in the corresponding section on the hidden layer but not other sections. For example, the unit on PFCcPC representing "numerosity" is connected with all units in the bottom-right quarter on the hidden layer (see Figure 3). Task 2 does not have a road network so that no training on DistanceWalk.

7. Module Performance

The PC module was trained within the integrated model with artificial data generated by a data generation software. (For details of the integrated model, see, [70]). At the current stage, we only focused on the training on metric estimates (e.g., numerosity, radius, flying distance, and walking distance). The training on topological comparison has not been completed thus it is omitted here. In addition, it has been found that in the COIN-AHA tasks, human subjects have mainly relied on the metric distances as the predictor [71].

Overall, we have demonstrated that the PC module can accurately extract various types of target values from the training dataset. To measure the module performance, we use the model-target correlation, which is the correlation between the minus phase activation on the target layer ("PC_Value") and the corresponding target value across trials within each epoch. Figure 9 shows the module performance in COIN-AHA Task 2 (without road network) and Task 3 (with road network). It can be seen that, in both tasks, the performances on Numerosity, DistanceFly, and DistanceWalk were very accurate (model-target correlations greater than 0.8 after 750 epochs). The only difference is that the training on Radius in Task 3 only showed a moderate performance. It is noted that the performance on DistanceFly was consistently more accurate than the performances on DistanceWalk and Numerosity. One apparent reason is due to the different levels of variances and ranges of the target values that are embedded in the input representations. For example, on a 24 by 24 grid, the maximum distance on the diagonal is $24 \times 1.414 \approx 34$, but the maximum DistanceWalk and Numerosity can be $24 \times 24 = 576$. The moderate performance on Radius can also be attributed to the variance on the input representations in which the location changes of individual units may not change the overall dispersion but can significantly affect the spatial correlations between units. That is, unlike other target values, the interactions between units can add additional noises in the model performance on Radius.

Importantly, Figure 9 shows the dissociation among various types of target values. In particular, the model-target correlations for both Numerosity and Radius in Task 2 reached approximately 0.8 after 750 epochs. Given that both target values have to be computed from the same visual input on PC_Attn, such a performance suggests an almost perfect dissociation between Numerosity and Radius (i.e., size). Crucially, this result has been obtained with the specific top-down connections from PFCcPC to the hidden layer (see Figure 3). That is, the units on PFC, each representing a unique demand, are, respectively, connected to the corresponding sections on the hidden layer. For example, the unit on PFCcPC representing "numerosity" is connected with all units in the "numerosity" section on the hidden layer, but not the units in other sections. As a result, when the demand from PFCcPC is to compute "Numerosity", the corresponding section of the hidden layer is more likely to be activated thus wins the inhibition competition over other sections. In other words, the top-down signal from PFCcPC provides a critical role in the functionality specialization on the hidden layer.

In contrast, Figure 10 shows the module performance in Task 3 with nonspecific PFCcPC-to-Hidden connections (e.g., each unit on PFCcPC is connected with all units on the hidden layer). It can be seen that dissociation had occurred to some extent, but the model-target correlation has significantly dropped for all types of target values. For example, the highest model-target correlation in Figure 10 was achieved on DistanceWalk, 0.65 (n = 100 trials), which was significantly lower than that in Figure 9, 0.86 (n = 100 trials) (comparing the two correlations by Fisher r-to-z transformation, $Z = -3.61$, two tailed $P < .001$). Since the only difference in these results was in the way PFCcPC is connected to the hidden layer, it appears that the failed dissociation between different target values is due to the lack of

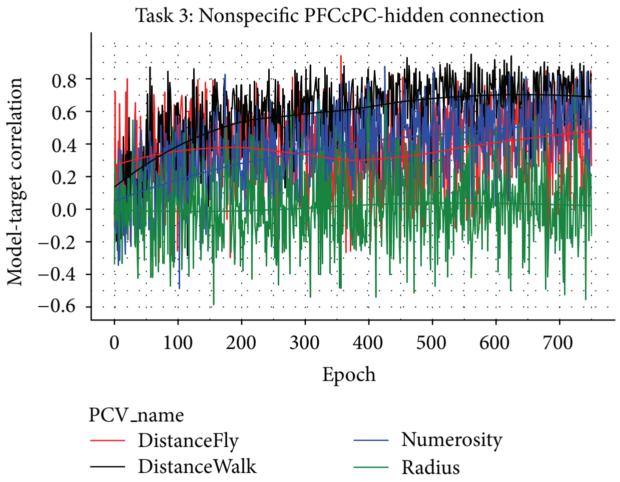


Figure 10: Performance of the PC module in task 3, with nonspecific PFCcPC-to-Hidden connections (e.g., each unit on PFCcPC is connected with all units on the hidden layer). It appears that the lack of specificity in the top-down control has caused the module's failure in dissociating between different target values.

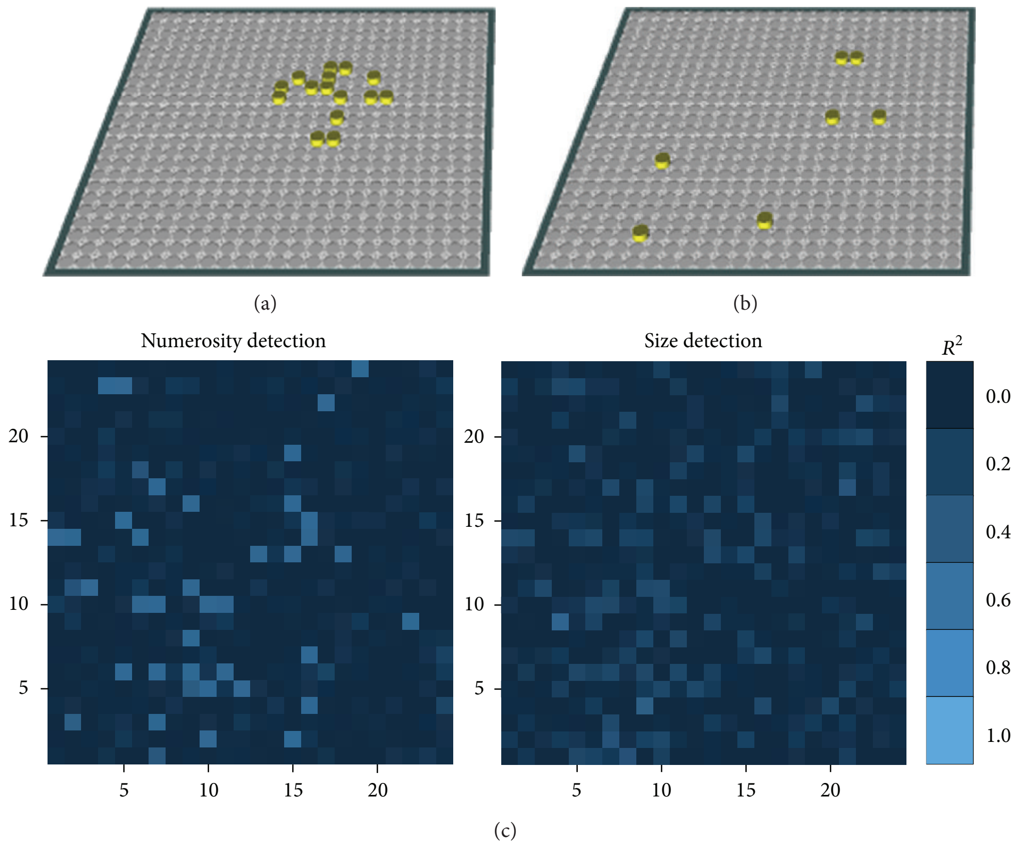
specificity in the top-down control. This finding is consistent with the current understanding in the literature on selective attention. For example, it has been suggested that the goal-directed attention can prioritize stimulus processing on the basis of motivational relevance via the dorsolateral prefrontal and posterior parietal network [45]. In our model, the top-down control was implemented by the PFCcPC-to-Hidden connections. With the specific connections, the active unit on PFCcPC is only projected onto the corresponding section on the Hidden layer. As a consequence, the units on that section are more likely to be active and better associated with the current target value since the corresponding connection weights are updated based on the activation values. Thus, by a goal-directed division of labor, different groups of units on the Hidden layer can develop associations with their own target values in a relatively independent fashion, resulting in an overall better performance.

Besides attention, expectation is considered as another top-down mechanism that mitigates the burdens of computational capacity in visual cognition, which may lie more medially in the posterior cortices as well as more ventrally in the frontal lobe [45]. In the PC module, the top-down control is not only from PFCcPC but also from the teaching signals provided on the target layer. As mentioned in the early section, it has been debated whether numerosity is a property that can be "sensed directly" from the visual input, dissociated from texture perception [51, 52]. To test this idea, we also conducted simulations with a simplified PC module with only one visual input layer and one hidden layer, and without teaching signals and top-down signals from PFCcPC (Figure 11). We find that by pure Hebbian association, units on the hidden layer can indeed show some dissociation between numerosity and size, but only to a certain extent. First, the correlation between hidden unit activation and either target value was hardly perfect (similar to the findings by [52]). For example, when computing the model-target correlations over 100 trials, a Pearson product-moment correlation coefficient of merely .254 can reach statistical significance level $P < .01$. Thus, a unit could be classified as a "numerosity neuron" without being able to selectively identify specific numbers. Second, many units showed overlapped sensitivity to both numerosity and size. That is, our finding appears to be more consistent with the proposal by Dakin et al. [48] that people's senses of number and density are intertwined (note that density = numerosity/size). Combined with the results shown above (e.g., Figures 9 and 10), it appears that perfect dissociation between numerosity and size can indeed occur, provided that there are specific goal-directed top-down controls.

8. Discussion

In this paper, we describe an integrated model of the parietal cortex for spatial-temporal information processing in sense-making. In summary, the development of the PC module suggests that, with quite similar structures, different types of environmental statistics (e.g., numerosity, size, Euclidean distance between two points, and length of curved line segment) can be extracted from visual inputs then represented as a magnitude value, supporting the proposal of a "common metric" housed in the parietal cortex (e.g., [5]).

The most significant finding from our simulations is that although early visual dissociation can occur between different types of environmental statistics, the goal-directed top-down control appears to be critical towards a complete dissociation. This finding is consistent with the current understanding in the literature on selective attention. In our model, the top-down control was implemented by the PFCcPC-to-Hidden connections. We demonstrated that high model-target correlations could be achieved only when the connections are specified with particular top-down projections. The interference without specific top-down controls can be more easily understood regarding how the kWTA inhibition mechanism would affect the dissociation between different types of environmental statistics. Crucially, different types of environmental statistics are obtained at different levels of abstraction. For example, a completely accurate estimation of numerosity (DistanceWalk and Numerosity) requires a complete spatial invariance whereas estimation of Euclidean distance (DistanceFly) or dispersion (Radius) is essentially based on spatial correlation. By the randomly initialized weights, some units may be biased towards a certain level of spatial invariance. However, with the kWTA inhibition in place, only the most active units have the chance to be updated and associated with the current target value. Thus, even when some units have shown some sensitivity to a certain type of target statistics at the early perceptual stage, such sensitivity may not be able to propagate further for the selectivity to be developed. One direct way to neurologically corroborate our simulation findings is to examine whether the task demand from the learning environment can interfere with the roles of neurons, for example, causing neurons initially sensitive to numerosity to be sensitive to size.


(a)

(b)

(c)

Figure 11: The dissociation and interference between numerosity and size. (a) A visual input with high numerosity but small patch size. (b) A visual input with low numerosity but large patch-size. (c) Without top-down control and teaching signals, units on the same hidden layer show sensitivity to either numerosity (left panel) or size (right panel), or both (the overlap of unit locations). Sensitivity is measured by R^2 in a linear regression of the activation on the respective magnitude.

Moreover, whereas our current model focuses on the interference and dissociation of different types of spatial information within the parietal cortex, it is also possible that other cortical topologies and mechanisms may contribute to the similar process. For example, it has been suggested that the mosaic organization of the superficial layers of the dorsocaudal medial entorhinal cortex (dMEC) represents a possible substrate for the modularity of the spatial map, which is an indication of early dissociation of different types of spatial information [72]. In addition, our current emphasis in modeling the parietal cortex is on the dissociation and representation of magnitude values, and a major goal is to reduce interference thus achieve high accuracy in performance. Accordingly, we have made several simplifications in modeling many of the subtasks. For example, we did not distinguish the processes of exact counting and subitizing (when the enumeration of objects is fast and accurate for sets of up to three or four items) [36]. And, we have avoided the problem of finding the shortest path in estimating the walking distance along the road.

In general, our modeling effort attempts to strike a balance between two types of preferences: whether to emphasize the mechanism of "attentional foveation" or to emphasize the mechanism of "perceptual segmentation" and "topological grouping." The former requires multiple sequential representations in the model (e.g., multiple scans in a crowded scene, exploration of all road segments between two points), and the latter makes it plausible to represent a set of stimuli as a whole (e.g., multiple objects of the same color segmented from others, a single road segment between two points). For example, the "zoom lens model" postulates that curve tracing has to be carried out in multiple passes each with a different foveation [73]. That is, in estimating the walking distance along the road, it would involve scanning multiple road segments between multiple intersections and points of interests and estimating distances according to different reference points. Neurologically, it has been posited that the parietal cortex is responsible for the transition between reference systems (e.g., [24, 74]). From behavioral studies, we have argued that the selection of reference systems (e.g., egocentric versus intrinsic) is an essential component in the internal representation of physical distances and relative locations [21, 23]. Thus, implementing the mechanisms of attentional foveation and selection of reference systems would lead to a more realistic model with the ability to identify some of the human heuristics and biases in spatial representation and reasoning. In the current model, all of the visual information is presented on the input layers at once. Thus, the model completely lacked the mechanism of attentional foveation. In addition, the selection of reference systems was rather fixed such that the model lacked the mechanism of flexibly changing the anchors of the reference system. We will further pursue these potential improvements in future research.

Acknowledgments

This paper was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) Contract no. D10PC20021 and the Office of Naval Research (ONR) Grant no. N00014-08-1-0042. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the USA Government.

References

[1] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 1: alternative perspectives," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 70–73, 2006.

[2] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 2: a macrocognitive model," *IEEE Intelligent Systems*, vol. 21, no. 5, pp. 88–92, 2006.

[3] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, "A data-frame theory of sensemaking," in *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, R. R. Hoffman, Ed., Erlbaum, Mahwah, NJ, USA, 2007.

[4] D. Kahneman and G. Klein, "Conditions for intuitive expertise: a failure to disagree," *American Psychologist*, vol. 64, no. 6, pp. 515–526, 2009.

[5] D. Bueti and V. Walsh, "The parietal cortex and the representation of time, space, number and other magnitudes," *Philosophical Transactions of the Royal Society B*, vol. 364, no. 1525, pp. 1831–1840, 2009.

[6] C. L. Colby and M. E. Goldberg, "Space and attention in parietal cortex," *Annual Review of Neuroscience*, vol. 22, pp. 319–349, 1999.

[7] E. M. Hubbard, M. Piazza, P. Pinel, and S. Dehaene, "Interactions between number and space in parietal cortex," *Nature Reviews Neuroscience*, vol. 6, no. 6, pp. 435–448, 2005.

[8] A. Nieder, "Coding of abstract quantity by "number neurons" of the primate brain," *Journal of Comparative Physiology A*, vol. 199, no. 1, pp. 1–16, 2013.

[9] M. L. Platt and C. Padoa-Schioppa, "Neuronal representations of value," in *Neuroeconomics: Decision Making and the Brain*, P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack, Eds., pp. 441–462, Elsevier, San Diego, Calif, USA, 2009.

[10] I. Vlaev, N. Chater, N. Stewart, and G. D. A. Brown, "Does the brain calculate value?" *Trends in Cognitive Sciences*, vol. 15, no. 11, pp. 546–554, 2011.

[11] P. W. Glimcher and A. Rustichini, "Neuroeconomics: the consilience of brain and decision," *Science*, vol. 306, no. 5695, pp. 447–452, 2004.

[12] M. L. Platt and P. W. Glimcher, "Neural correlates of decision variables in parietal cortex," *Nature*, vol. 400, no. 6741, pp. 233–238, 1999.

[13] E. Brandstätter, G. Gigerenzer, and R. Hertwig, "The priority heuristic: making choices without trade-offs," *Psychological Review*, vol. 113, no. 2, pp. 409–432, 2006.

[14] K. Louie and P. W. Glimcher, "Efficient coding and the neural representation of value," *Annals of the New York Academy of Sciences*, vol. 1251, no. 1, pp. 13–32, 2012.

[15] V. Walsh, "A theory of magnitude: common cortical metrics of time, space and quantity," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 483–488, 2003.

[16] J. Gottlieb, "From thought to action: the parietal cortex as a bridge between perception, action, and cognition," *Neuron*, vol. 53, no. 1, pp. 9–16, 2007.

[17] L. G. Ungerleider and A. H. Bell, "Uncovering the visual "alphabet": advances in our understanding of object perception," *Vision Research*, vol. 51, no. 7, pp. 782–799, 2011.

[18] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of Visual Behavior*, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds., MIT Press, Cambridge, Mass, USA, 1982.

[19] P. Pinel, M. Piazza, D. Le Bihan, and S. Dehaene, "Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments," *Neuron*, vol. 41, no. 6, pp. 983–993, 2004.

[20] T. Yang and M. N. Shadlen, "Probabilistic reasoning by neurons," *Nature*, vol. 447, no. 7148, pp. 1075–1080, 2007.

[21] Y. Sun and H. Wang, "Perception of space by multiple intrinsic frames of reference," *PloS ONE*, vol. 5, no. 5, Article ID e10442, 2010.

[22] F. P. Tamborello, Y. Sun, and H. Wang, "Spatial reasoning with multiple intrinsic frames of reference," *Experimental Psychology*, vol. 59, no. 1, pp. 3–10, 2012.

[23] H. Wang, T. R. Johnson, Y. Sun, and J. Zhang, "Object location memory: the interplay of multiple representations," *Memory and Cognition*, vol. 33, no. 7, pp. 1147–1159, 2005.

[24] Y. Pertzov, G. Avidan, and E. Zohary, "Multiple reference frames for saccadic planning in the human parietal cortex," *The Journal of Neuroscience*, vol. 31, no. 3, pp. 1059–1068, 2011.

[25] R. P. Kesner, "The posterior parietal cortex and long-term memory representation of spatial information," *Neurobiology of Learning and Memory*, vol. 91, no. 2, pp. 197–206, 2009.

[26] R. C. O'Reilly, Y. Munakata, M. J. Frank, and T. E. Hazy, 2012, *Computational Cognitive Neuroscience*, Wiki Book, 1st edition, http://ccnbook.colorado.edu.

[27] S. P. Tipper and M. Behrmann, "Object-centered not scene-based visual neglect," *Journal of Experimental Psychology*, vol. 22, no. 5, pp. 1261–1278, 1996.

[28] D. H. Foster and C. J. Savage, "Uniformity and asymmetry of rapid curved-line detection explained by parallel categorical coding of contour curvature," *Vision Research*, vol. 42, no. 18, pp. 2163–2175, 2002.

[29] S. Dehaene, E. Spelke, P. Pinel, R. Stanescu, and S. Tsivkin, "Sources of mathematical thinking: behavioral and brain-imaging evidence," *Science*, vol. 284, no. 5416, pp. 970–974, 1999.

[30] J. M. Wolfe, M. L. H. Võ, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and nonselective pathways," *Trends in Cognitive Sciences*, vol. 15, no. 2, pp. 77–84, 2011.

[31] W. Zhang and S. J. Luck, "Discrete fixed-resolution representations in visual working memory," *Nature*, vol. 453, no. 7192, pp. 233–235, 2008.

[32] Y. Xu and M. M. Chun, "Selecting and perceiving multiple visual objects," *Trends in Cognitive Sciences*, vol. 13, no. 4, pp. 167–174, 2009.

[33] P. R. Roelfsema, "Cortical algorithms for perceptual grouping," *Annual Review of Neuroscience*, vol. 29, pp. 203–227, 2006.

[34] P. R. Roelfsema, V. A. F. Lamme, H. Spekreijse, and H. Bosch, "Figure—ground segregation in a recurrent network architecture," *Journal of Cognitive Neuroscience*, vol. 14, no. 4, pp. 525–537, 2002.

[35] S. L. Prime, M. Niemeier, and J. D. Crawford, "Transsaccadic integration of visual features in a line intersection task," *Experimental Brain Research*, vol. 169, no. 4, pp. 532–548, 2006.

[36] M. Piazza and V. Izard, "How humans count: numerosity and the parietal cortex," *Neuroscientist*, vol. 15, no. 3, pp. 261–273, 2009.

[37] M. Piazza, A. Mechelli, C. J. Price, and B. Butterworth, "Exact and approximate judgements of visual and auditory numerosity: an fMRI study," *Brain Research*, vol. 1106, no. 1, pp. 177–188, 2006.

[38] C. R. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press, Cambridge, Mass, USA, 2000.

[39] T. E. Hazy, M. J. Frank, and R. C. O'Reilly, "Neural mechanisms of acquired phasic dopamine responses in learning," *Neuroscience and Biobehavioral Reviews*, vol. 34, no. 5, pp. 701–720, 2010.

[40] R. C. O'Reilly and M. J. Frank, "Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia," *Neural Computation*, vol. 18, no. 2, pp. 283–328, 2006.

[41] S. Dehaene, V. Izard, E. Spelke, and P. Pica, "Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures," *Science*, vol. 320, no. 5880, pp. 1217–1220, 2008.

[42] J. F. Cantlon, S. Cordes, M. E. Libertus, and E. M. Brannon, "Comment on "Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures"," *Science*, vol. 323, no. 5910, p. 38, 2009.

[43] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Developmental Science*, vol. 10, no. 1, pp. 89–96, 2007.

[44] M. Sigman and S. Dehaene, "Brain mechanisms of serial and parallel processing during dual-task performance," *The Journal of Neuroscience*, vol. 28, no. 30, pp. 7585–7598, 2008.

[45] C. Summerfield and T. Egner, "Expectation (and attention) in visual cognition," *Trends in Cognitive Sciences*, vol. 13, no. 9, pp. 403–409, 2009.

[46] K. Louie and P. W. Glimcher, "Separating value from choice: delay discounting activity in the lateral intraparietal area," *The Journal of Neuroscience*, vol. 30, no. 16, pp. 5498–5507, 2010.

[47] E. J. Anderson, S. C. Dakin et al., "The neural correlates of crowding-induced changes in appearance," *Current Biology*, vol. 22, no. 13, pp. 1199–1206, 2012.

[48] S. C. Dakin, M. S. Tibber, J. A. Greenwood, F. A. A. Kingdom, and M. J. Morgan, "A common visual metric for approximate number and density," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. 19552–19557, 2011.

[49] F. H. Durgin, "Texture density adaptation and visual number revisited," *Current Biology*, vol. 18, no. 18, pp. R855–R856, 2008.

[50] D. Burr and J. Ross, "A visual sense of number," *Current Biology*, vol. 18, no. 6, pp. 425–428, 2008.

[51] J. Ross and D. Burr, "Number, texture and crowding," *Trends in Cognitive Sciences*, vol. 16, no. 4, pp. 196–197, 2012.

[52] I. Stoianov and M. Zorzi, "Emergence of a "visual number sense" in hierarchical generative models," *Nature Neuroscience*, vol. 15, no. 2, pp. 194–196, 2012.

[53] J. D. Roitman, E. M. Brannon, and M. L. Platt, "Monotonic coding of numerosity in macaque lateral intraparietal area," *PLoS biology*, vol. 5, no. 8, p. e208, 2007.

[54] A. Nieder, "Counting on neurons: the neurobiology of numerical competence," *Nature Reviews Neuroscience*, vol. 6, no. 3, pp. 177–190, 2005.

[55] A. Nieder and E. K. Miller, "A parieto-frontal network for visual numerical information in the monkey," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 19, pp. 7457–7462, 2004.

[56] H. Sawamura, K. Shima, and J. Tanji, "Numerical representation for action in the parietal cortex of the monkey," *Nature*, vol. 415, no. 6874, pp. 918–922, 2002.

[57] H. Sawamura, K. Shima, and J. Tanji, "Deficits in action selection based on numerical information after inactivation of the posterior parietal cortex in monkeys," *Journal of Neurophysiology*, vol. 104, no. 2, pp. 902–910, 2010.

[58] S. Dehaene and J. P. Changeux, "Development of elementary numerical abilities: a neuronal model," *Journal of Cognitive Neuroscience*, vol. 5, no. 4, pp. 390–407, 1993.

[59] T. Verguts and W. Fias, "Representation of number in animals and humans: a neural model," *Journal of Cognitive Neuroscience*, vol. 16, no. 9, pp. 1493–1504, 2004.

[60] O. Tudusciuc and A. Nieder, "Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 36, pp. 14513–14518, 2007.

[61] R. C. Kadosh, K. C. Kadosh, T. Schuhmann et al., "Virtual dyscalculia induced by parietal-lobe TMS impairs automatic magnitude processing," *Current Biology*, vol. 17, no. 8, pp. 689–693, 2007.

[62] L. Chen, "The topological approach to perceptual organization," *Visual Cognition*, vol. 12, no. 4, pp. 553–637, 2005.

[63] N. J. Goodrich-Hunsaker, B. P. Howard, M. R. Hunsaker, and R. P. Kesner, "Human topological task adapted for rats: spatial information processes of the parietal cortex," *Neurobiology of Learning and Memory*, vol. 90, no. 2, pp. 389–394, 2008.

[64] N. J. Goodrich-Hunsaker, M. R. Hunsaker, and R. P. Kesner, "Dissociating the role of the parietal cortex and dorsal hippocampus for spatial information processing," *Behavioral Neuroscience*, vol. 119, no. 5, pp. 1307–1315, 2005.

[65] R. G. Golledge, "The nature of geographic knowledge," *Annals of the Association of American Geographers*, vol. 92, no. 1, pp. 1–14, 2002.

[66] M. Schneider and T. Behr, "Topological relationships between complex spatial objects," *ACM Transactions on Database Systems*, vol. 31, no. 1, pp. 39–81, 2006.

[67] D. H. Uttal, "Seeing the big picture: map use and the development of spatial cognition," *Developmental Science*, vol. 3, no. 3, pp. 247–264, 2000.

[68] A. Tversky and D. Kahneman, "Judgment under uncertainty: heuristics and biases. Biases in judgments reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[69] D. Kahneman and S. Frederick, “Representativeness revisited: attribute substitution in intuitive judgment,” in *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin, and D. Kahneman, Eds., pp. 49–81, Cambridge University Press, New York, NY, USA, 2002.

[70] S. A. Herd, T. R. Huang, T. E. Hazy, T. Kriete, and R. C. O’Reilly, “Strategic cognitive sequencing: a computational cognitive neuroscience approach,” Submitted.

[71] M. D. Ziegler, M. Howard, and A. Zaldivar, “Simulation of anchoring bias in a spatial estimation task due to cholinergic neuromodulation,” Submitted.

[72] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser, “Microstructure of a spatial map in the entorhinal cortex,” *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.

[73] C. Lefebvre, R. Dell’acqua, P. R. Roelfsema, and P. Jolicœur, “Surfing the attentional waves during visual curve tracing: evidence from the sustained posterior contralateral negativity,” *Psychophysiology*, vol. 48, no. 11, pp. 1510–1516, 2011.

[74] P. Byrne, S. Becker, and N. Burgess, “Remembering the past and imagining the future: a neural model of spatial memory and imagery,” *Psychological Review*, vol. 114, no. 2, pp. 340–375, 2007.