University of California
Santa Barbara

Rule Representation in Explicit Categorization

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in
Dynamical Neuroscience

by

Stella Sophia von Meer

Committee in charge:

Professor F. Gregory Ashby, Chair
Professor Jean Carlson
Professor Barry Giesbrecht
Professor Jeff Moehlis

March 2020

The Dissertation of Stella Sophia von Meer is approved.

Professor Jean Carlson

Professor Barry Giesbrecht

Professor Jeff Moehlis

Professor F. Gregory Ashby, Committee Chair

December 2019

Acknowledgements

I would like to thank Greg Ashby for giving me the opportunity to pursue this degree. I am grateful for Dane with all your love, encouragement, and support. Thanks to my siblings, Diana, Cosima, Laura, Jan, and Basse, for their love, support, and believing in me. Thanks to my parents for giving me life, and for my mother nurturing my early interest in math and science. I want to thank everyone in the Ashbylab: Vivian, Lauren, Yiwen, Luke, Jeff, and Paul, for your help, insight, and encouragement. Thanks to Bridget for providing me with financial opportunity. Thanks to nature for gifting me the mind from which this work emanated, and particular gratitude to the ocean for offering a sanctuary. And of course thanks to all my wonderful friends, who fill my life with joyful moments, wisdom, and who continue to believe in my creative power. Dedicated to all that have made my journey possible, with gratitude.

The way of truth is along the path of intellectual sincerity.

Henry Smith Pritchett

Curriculum Vitæ

Stella Sophia von Meer

Education

2019	Ph.D. in Dynamical Neuroscience with emphasis in Neural Engineering (Expected), University of California, Santa Barbara.
2010	M.S. in Medical Neuroscience, Charite Universitaetsmedizin Berlin, Germany.
2007	B.S. Physics (Cum Laude), Loyola University New Orleans, Louisiana.

Publications

1. Ashby, F. G., Valentin, V. V., von Meer, S. S. (2015). Differential effects of dopamine-directed treatments on cognition. Neuropsychiatric disease and treatment, 11, 1859.

2. von Meer, S. S., Valentin, V. V., Ashby, F.G. (in review). Simultaneous Reward and Category Learning

3. von Meer, S. S., Soto, F.A., Ashby, F. G. (in preparation). Exploring Robustness of an *a priori* estimated Rule-Based Category Learning Network using TMS.

4. von Meer, S. S. Ashby, F. G. (in preparation). Rule Representation during Explicit Categorization.

Abstract

Rule Representation in Explicit Categorization

by

Stella Sophia von Meer

One defining characteristic of human behavior is the ability to select an appropriate action in an entirely novel situation. How abstract rules are represented in the brain, and how these representations operate on internal models of the world to generate flexible rapidly optimized behaviors remains an open question in neuroscience as well as computer science. The hallmark of quickly optimized flexibility inherent to explicit behavioral rules has lead to the assumption that these rules are based on high level abstractions. In the recent past, behavioral measurements in the human psychophysics literature were linked to predictions at the processing level and the convenient mathematical construct of a decision criterion has precipitated in several cognitive process models. However, the assumption that higher order decision processes involve comparisons to some internal criterion is not trivial and was investigated. This thesis provides evidence to falsify the criterion as a processing component in human decision making and offers fundamental insights to reverse engineer decision processes in the brain. Cognitive flexibility and rule guided behavior appear to rely on phylogenetically advanced extrapolation processes that are mediated by dynamic feed-forward and feed-back circuits, which continually update internal and external information to support goal directed behavior.

Contents

Chapter 1

Introduction

1.1 Statement of Significance

The representation of explicit abstract rules in the human brain remains an enigma that continues to be explored from many differing perspectives including single-unit recordings, electroencephalography, functional magnetic resonance imaging, neuropharmacology, transcranial magnetic stimulation, near infrared spectroscopy, and mathematical psychology. The computational rules underlying the neurobiology that supports rule-guided behavior remain elusive. Most cognitive models viewed through the lens of Marr′s [1] levels of analysis are computational rather than algorithmic, and theories merely scratch the surface of an accurate implementation. Countless popular cognitive models are based on the notion of a criterion, despite the lack of neurobiological evidence that support such computations in decision making. While criterion based models have a high degree of utility in the realm of data analysis, these models may be less informative at the process level where the criterion may not find a specific biological correlate. Is the criterion an epiphenomenon or does it hold implementational credence? This thesis contributes evidence towards a more complete picture of explicit rule representation and

challenges the criterion relic inherent to many cognitive theories.

1.2 General Background

One defining characteristic of the human being is the ability to select an appropriate action in an entirely novel situation. This remarkable cognitive ability appears to rely on phylogenetically advanced extrapolation processes. Categorization behavior represents an instance where available perceptual and cognitive information is integrated to generate an appropriate response. Behavioral rules are grossly classified as implicit and explicit. Implicit behavioral rules are inflexibly optimized to one particular data set via time consuming trial-by-trial reinforcement learning. While explicit rules appear to operate flexibly on any new data space with incredibly small optimization times. The hallmark of quickly optimized flexibility inherent to explicit behavioral rules has lead to the assumption that these rules are based on high level abstractions. However, it remains a mystery how explicit rules are encoded and learned. For instance, the discovery of the optimal abstract rule occurs suddenly, as if all the available information converges to one realization at one instance, commonly known as the *AHA! moment*. How explicit rules are represented in the brain, and how these representations operate on internal models of the world to generate flexible rapidly optimized behaviors remains an open question that continues to boggle minds of scientists in the fields of neuroscience as well as computer science.

Reverse engineering decision making processes in the brain has lead to the development of several cognitive models. The majority of these models generate an output (response) after a comparison to some criterion. Dominant models include stochastic diffusion type models [2] and accumulator/counter models [3]. Most of these cognitive models differ in their underlying processing assumptions, however, one striking similar-

ity is that a response is generated only once a particular criterion has been reached or breached. The criterion discussed here is different from systematic tendencies in decision making, e.g. decision bias [4] or decision criteria [5]. Rather the criterion here references a processing component in cognitive models. The assumption that higher order information processes involve comparisons to some criterion is not trivial and may find its roots in signal processing. Abstractly, decision making can be dissected into perceptual and decisional processes. These have been studied extensively using one of the most influential frameworks: signal detection theory (SDT), which separately estimates perceptual sensitivity and a decision criterion [6]. SDT has a long tradition in the field of psychology and has been applied to almost all tasks in which two stimuli must be discriminated, including yes/no tasks, rating tasks, and forced-choice tasks [7, 8]. Twenty years after SDT was formalized, it was generalized to multidimensional perceptual spaces by general recognition theory (GRT; [9]). Both GRT and SDT assume the perceptual effect of a stimulus is subject to variability resulting from external and internal measuring noise. Therefore, the perceptual effect is associated with a multivariate probability density in perceptual space [10], although it can occur anywhere within that space.

GRT can account for multidimensional perceptual experiences and provides several other advantages over SDT: 1. in its most general form no distributional assumptions are made about the perceptual effects, 2. dimensional interactions in the perceptual space are rigorously defined (e.g. varieties of perceptual independence; [9]), and 3. the decision criterion in GRT generalizes from a point in unidimensional perceptual space to a continuous curve in two-dimensional space or a hyperplane in n-dimensional perceptual space [11]. Each response region is deterministically associated with a particular response. One prominent special case of GRT assumes the perceptual distributions to be Gaussian, and therefore the most popular form of (unidimensional) SDT is contained in GRT. Kadlec and Townsend [11] developed a rigorous set of mathematical relations

between this Gaussian version of GRT and SDT, and provide decision trees to test for varieties of perceptual independence (e.g. perceptual independence, sampling independence, perceptual separability, decisional separability, and marginal response invariance). Since the development of GRT, it has been applied to a variety of perceptual phenomena [12] frequently under the Gaussian assumption [13, 14, 15], although exceptions certainly exist [9, 16].

An excellent overview of GRT can be found in Ashby and Soto [12]. Briefly, in GRT a stimulus is defined as a collection of components or features, with each represented by a dimension in perceptual space. The perceptual effect of a particular stimulus at a particular instance in time is a random sample from a multivariate joint probability density function in perceptual space. From the relation of the probability density to the marginal projections onto each perceptual axis, it can be deduced whether the perceptual dimensions are perceived integrally or separably. Perceptual processes are related to similarity judgments that separate the perceptual space into recognition regions [13]. It is assumed that recognition regions are partitioned into response regions according to some decision boundary that can be defined in several ways [17, 18, 19]. In the absence of bias, the decision boundary is assumed to be equivalent to an optimal partition [20] that is represented by the likelihood ratio [19], and the mapping from recognition regions to response regions is deterministic (i.e. one-to-one; [21]). Importantly, the mapping between the physical stimulus components and the dimensions of the hypothetical perceptual space is likely monotonic, but not linear, e.g. as assumed by Fechner's law, or Stevens law [22]. Furthermore, the form of this ordered relationship is assumed to be preserved under the dimensional stretching or shrinking that is used to model cognitive top-down control operations such as attention [13].

GRT provides methods to test three fundamental properties of perception: perceptual independence (statistical independence between perceptual components of one stimulus,

i.e. distribution axes are parallel to the perceptual axes), perceptual separability (perceptual invariance of one component relative to another for a group of stimuli), and decisional separability (decision bounds that are orthogonal to the perceptual axes). Under decisional separability, the decision boundary intersects the relevant perceptual axis orthogonally and can simply be described by the intercept or *criterion* value. In this case, the decision depends only on information on the relevant perceptual axis (i.e. the marginal probability density). Note, the decision bound that operates on stimulus space is not equivalent to the decision boundary in perceptual space. Even though monotonicity may be preserved when mapping representations from one space onto the other, the covariance matrices associated with representation in either space are likely different and so the shape of the boundary is subject to change [19]. In many applications of GRT the decision bound represents a fundamental construct.

According to Marr's levels of analysis, both SDT and GRT are computational models [1] that have multiple algorithmic implementations, and despite the success of both theories, the GRT decision boundary or SDT criterion may not have actual psychological meaning. Nonetheless, the idea of a decision boundary has shaped our interpretation of neurobiological data in a non-trivial way (e.g. rule neurons [23, 24]. Furthermore, SDT and GRT are both static models that were developed with respect to accuracy, and as such required no further processing assumptions [9]. The fact that computational-level models are associated with multiple process interpretations becomes evident with the possibilities in which a decision boundary can be estimated. For instance, the decision rule that maximizes accuracy could be in the form of a likelihood ratio, discriminant function, or Boolean algebra.

However, in 2000 Ashby published a dynamic version of GRT, where the point percept from static GRT is replaced with a multivariate diffusion process that generates predictions at the level of response time (RT) [25]. Here, the drift rate is proportional to the

distance between the mean percept and the decision bound. Under the assumption that the dynamic percept has time invariant intra-trial variance-covariance matrices, the drift rate formalized a process for the *RT-distance hypothesis*, which predicts a monotonic decrease of RTs with increasing distance between the percept and the decision bound [16]. A stronger version of the RT-distance hypothesis predicts marginal invariance, that is, RTs times are symmetrically distributed about the criterion. Although the stochastic GRT model predicts violations in the RT-distance hypothesis when some infinitesimal variance-covariance matrices (of the independently sampled perceptual effects during a trial) are unequal [25], the existence of a decision boundary continues to be assumed. The question remains: does the brain represent abstract rules via a criterion?

1.3 Experimental Evidence

In the realm of categorization [26], an armamentarium of invasive animal and human neuroimaging studies including neuropsychiatric disease have established the existence of several disparate learning systems that cater to different environmental demands [27, 28, 29]. For instance, Ashby and Waldron [30] found that human categorization performance data was best explained by non-parametric rather than parametric models and introduced a neurobiological model called the *striatal pattern classifier* which is mediated by connectivity within the striatum. However, in their experiments decisional separability was violated, such that their results describe the implicit procedural system only. The representation of computational rules that support nonparametric information-integration strategies inherent to the procedural system has been characterized extensively. A neurobiological model of procedural memory describes highly specific stimulus-category label and category label-response associations that are learned via a sharp temporal precision dopaminergic reinforcement signal at cortico-striatal synapses

[31, 32, 28, 33]. One defining feature of procedural memories is the lack of generalizability. For example, specific implicit associations learned during information-integration tasks do not generalize to responses associated with a different region in space, e.g. switching response locations [34] or areas of stimulus space for which associations have not been trained, even if the decision bound that partitions the new stimulus space is a linear extension of the previously learned associations, e.g. analogical transfer [35]. In contrast, explicit memories generalize to new response goals and performance during analogical transfer that affected irrelevant stimulus dimension only was nearly perfect in a rule-based task.

The representation of abstract generalizable behavioral rules in the explicit system has not been modeled successfully at the neurobiological level. Instead, models rely on abstract computations themselves (i.e. parametric decision bound estimation, discriminant functions, and logical rules). In the current model of explicit category learning, the core process has been characterized by a hypothesis testing mechanism that is implemented via a criterion based discriminant function [36, 37]. The logical rules underlying hypothesis testing are easily verbalizable [36] and operate on perceptual dimensions independently, given that these satisfy the assumption of independence [13]. It is noteworthy that the verbal description rarely refers to a comparison to a decision criterion, rather the operational rule references response regions (e.g. *respond A if the relevant feature is large, and respond B if the relevant feature is small*). When perceptual separability is satisfied general abstractions are learned via working memory and executive attention processes that recruit a prefrontally mediated explicit decision network [38, 39, 37, 28, 40]. Given that explicit categorization relies on working memory and attention, the learned associations are subject to capacity and perceptual complexity limits [41, 42, 43, 44, 12, 26]. Furthermore, the working memory traces that support the decisions process are assumed to be unstable, which has been associated with trial-by-trial variability in the represen-

tation of the criterion [39].

An often overlooked consequence of multivariate decision bound models is that linear and quadratic bounds, once estimated, ought to result in similar accuracy and RT results. However, quadratic bounds are more difficult to learn than linear bounds. These performance differences, although long established, did not suffice to overrule the assumption of parametric classification in category learning [30]. Furthermore, Ashby and Ell [45] showed that the amount of category overlap fundamentally affected decision strategy in information-integration tasks. Specifically, more than half of the participants were best fit by the suboptimal unidimensional rule throughout the experiment, when the optimal procedural strategy would outperform the suboptimal rule by 12% accuracy but the feedback was not deterministic (high category overlap). Interestingly, monetary incentives biased towards a procedural strategy slightly increased the number of explicit strategy users. Their results suggest that category environments with moderate uncertainty such as probabilistic classification tasks provoke explicit strategies if these are nearly optimal at maximizing accuracy. This suggests that successful performance of the procedural system is sensitive to reward uncertainty and confirm the previous finding that effective procedural learning requisites consistent category-response mappings [34]. Thus, the explicit system appears to withstand reward uncertainty, such that probabilistic feedback is not sufficient to interfere with successful categorization performance. The fact that consistent category-response associations are not necessary for successful performance appear to favor a criterion model or a potentially complex alternative. However, the experiments presented in the following contribute to resolve this issue.

1.4 Experimental investigation

To my knowledge there has not been an investigation on the validity of the decision boundary as a component of the decision process. There are two candidates that could support the decision process: 1. the decision boundary, and 2. the perceptual regions partitioned by the decision boundary. Since our investigation is focused on the case in which decisional separability holds, i.e. the perceptual regions are assumed to be separated by a decision bound that is orthogonal to the relevant categorization dimension and can be fully described by the intercept, we will refer to the decision bound model as the *criterion* model. The two candidates give rise to the criterion and the direct-mapping models, which make very different and easily testable predictions at the level of accuracy and RT. Experiment 1 explores the nature of rule representation by testing predictions of the criterion model against the predictions of the direct-mapping model with a speeded unidimensional rule-based task. Experiment 2 seeks to characterize rule representation further by exploring neuroelectric correlates using the non-invasive electro-encephalography neuroimaging technique.

Chapter 2

Behavioral Rule Region Experiment

2.1 Introduction

The notion of a response criterion is ubiquitous in cognitive science, in part because of the widespread influence of signal detection theory. For example, the standard model of rule use assumes decisions are based on a comparison of the stimulus to some remembered criterion. In the field of category learning, simple rules such as *respond A if the stimulus is large on dimension X, and B if it is small* could be implemented either by comparing the stimulus to a criterion value that separates large and small values, or by mapping large values directly to an A-response and small values directly to a B-response. While the criterion approach has proven fruitful in post-experiment data analysis, its validity as process model in decision making remains unchallenged.

The criterion model (CM) builds on distance-based similarity measures that find success in the stimulus space, but may not translate into perceptual space despite the enticing term *psychological distance* [17]. The criterion represents the optimal partition between stimulus distributions, and is assumed to be placed at the estimated mid-point between stimulus category means and variances. The CM is a decisional reduction of GRT and

assumes the RT-distance hypothesis. The CM assumes a decision rule is implemented by comparing the current stimulus to an internal criterion and relies on trial-by-trial comparisons. Hence to sort large and small stimuli, the criterion model assumes people learn the value of an intermediate stimulus to which each sample stimulus is compared. The relationship between perception and decision is typically modeled by some function plus noise. The CM contains two sources of noise: perceptual noise and criterial noise. If the noise is ignored, the relationship between perception and decision is ordinal. Adding perceptual and criterial noise will decrease accuracy and increase RT for stimuli close to the decision boundary. One major limitation of the CM is that the optimal placement of the criterion depends on correctly inferring the form of the underlying category distributions [9, 19]. This is nearly impossible because the participant does not know the location of every exemplar in the distribution, nor the true parameters of the perceptual noise distributions. Furthermore, imperfect memory prevents robust a priori estimation of the expected distributions ('expected' because this is before a sufficient stimulus sample was seen) and makes noiseless memory of the criterion impossible. At first glance, the CM model appears parsimonious and well suited to describe the algorithmic level of the underlying neural machine. At second glance, the estimation procedure the CM requires reveals itself to be quite laborious and subject to several sources of error. For instance, for a single criterion, one needs to estimate the category distributions up to the second moment as well as the mid-distance point between the estimated category distributions. These limitations make the CM less attractive and intractable at the *implementational* level [1]. An alternative to the criterion model is the direct mapping model in which the perceptual regions are highlighted and the decision boundary has no psychological meaning.

The CM predicts that response accuracy will increase monotonically [46] and response time (RT) will decrease monotonically [16, 25] as the distance of the presented stimulus to

the criterion increases. These predictions were tested with a speeded unidimensional rule-based categorization task [40] for which perceptual separability and decisional separability are assumed [12]. Following Sternberg's observation that *the selection of a response requires the use of information that is in memory, the latency of the response will reveal some thing about the process by which the information is retrieved* [47], RT information is a valuable behavioral parameter that allows a glimpse at the underlying process and carries strong empirical weight in the assessment of process models. The RT results presented here falsified the predictions of the criterion model and are consistent with a direct-mapping model.

2.2 Method

2.2.1 Participants

In this experiment participants were 50 (24 males, 26 females; age range: 19-24) and 61 (28 males, 33 females; age range: 19-23) healthy UC Santa Barbara undergraduate students in likelihood and control conditions, respectively. The participants had no previous exposure to the stimuli prior to the experiment.

2.2.2 Categorization Stimuli

The stimuli were gray-scale circular sine-wave gratings that varied across trials on two dimensions: spatial frequency (cycles per degree of visual angle; CPD) and orientation (radians of counterclockwise rotation from horizontal). All stimuli subtended 7° of visual angle and were displayed against a gray background. The stimuli in each category were generated according to the randomization technique (e.g. [17]) explained in the following. First, a 100 × 100 stimulus space was defined in which perceptual salience in both

(e.g. CPD and orientation) dimensions was approximately equal. The two categories were defined by bivariate normal distributions and parameter values specified in Table 2.1. Next, 400 random samples were selected from each bivariate normal to generate two categories. To minimize distributional overlap, any outlier more than 3 standard deviations from the distribution mean (identified via *Mahalanobis* distance) were discarded and replaced during the sampling procedure. The sample values on both dimensions were linearly transformed so that the sample statistics exactly matched the population parameter values listed in Table 2.1. Specifically, each $[0, 100]$ value for spatial frequency (i.e., x) was converted to cycles per degree of visual angle (i.e., f) via $f = \frac{x}{30} + 0.25$ and each $[0, 100]$ value for orientation (i.e., y) was converted to radians of counterclockwise rotation from horizontal (i.e., o) via $o = \frac{\pi}{200}y + \frac{\pi}{9}$. During the experiment, each (f, o) coordinate pair was used to create a sine-wave grating using Brainard's [48] Psychophysics Toolbox. All participants in each condition were presented with the same stimuli. However, presentation order was randomized across participants.

Table 2.1: Mean and Variance of Experimental Stimulus Distributions

Category	Stim N	μ_f	μ_o	σ_f^2	σ_o^2	cov
A_{cons}	360	27.5	50	40	251	0
B_{cons}	360	72.5	50	40	251	0
A_{disp}	40	72.5	50	1	251	0
B_{disp}	40	27.5	50	1	251	0

In the control condition each category was generated from sampling only one bivariate normal, such that all stimuli were region consistent. While in the likelihood condition each category consisted of 90% region consistent (*cons*) and 10% region disparate (*disp*) stimuli. I chose these proportions to generate a sufficient sample size (40 samples) for the disparate stimuli and rule out a memorization strategy while keeping the likelihood

of stimulus membership region consistent (i.e. an ideal observer would respond A if the stimulus falls into the *A region*). The stimulus distributions along with the equal likelihood contours (concentric ellipses) are shown in Figure 2.1. The likelihood ratio between the region consistent and region disparate stimuli equals 1.988, while the likelihood ratio at the distribution means equals 2.321. The marginal stimulus distribution densities along the relevant CPD dimension and their response likelihoods are shown in Figure 2.2. The figure shows that the stimulus space can be divided into a decision space with corresponding A- and B-regions, and the vertical line on the x-axis at $x = 50$ synonymous with the decision criterion denotes the equal likelihood-response-contour between the two distributions separated by $d' = 2.4338$.

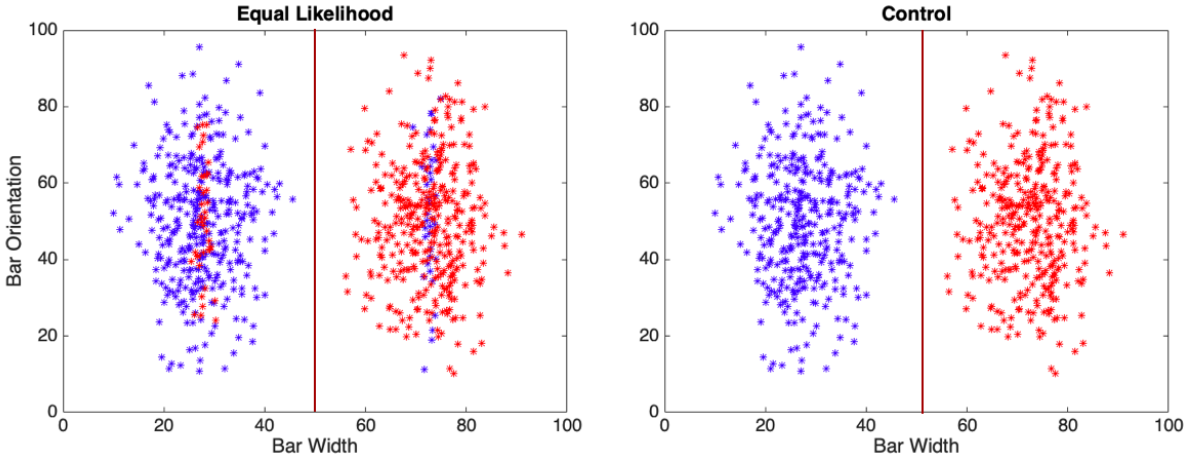

Figure 2.1: Stimulus distributions for likelihood and control conditions with equal probability category contours located at Bar Width coordinate 50.

2.2.3 Procedure

The experiment took place in a dimly lit room that can hold up to eight participants. Each participant sat in front of a computer screen with a keyboard for responding, and was provided with headphones for individual auditory feedback. Participants were informed that they would be categorizing novel circular sine-wave gratings (stimulus)

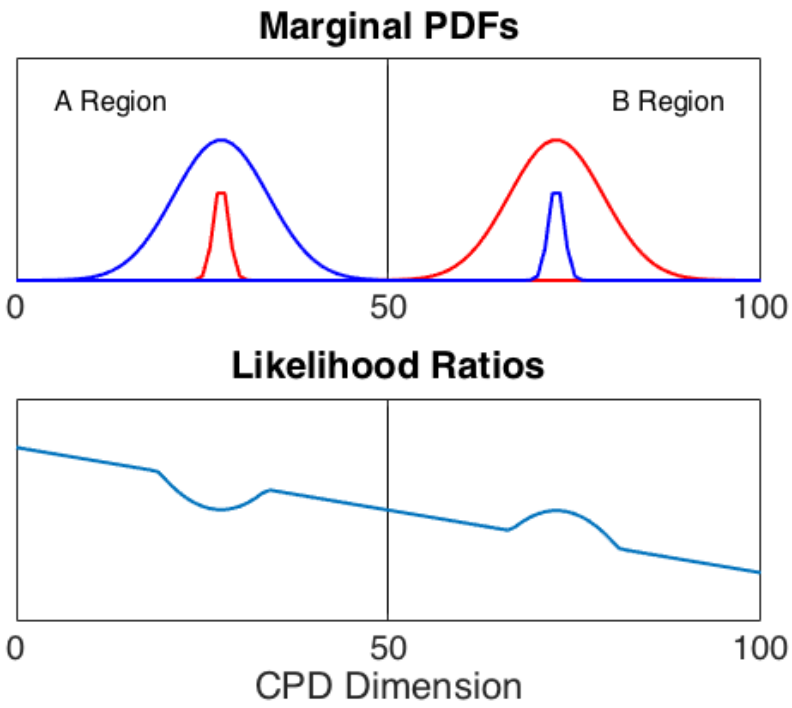

Figure 2.2: Marginal stimulus distribution densities (top) and their response log-likelihoods (bottom).

belonging either to category A or B, and that category membership would become apparent through feedback. Each trial began with crosshairs centered on a gray background shown for 300ms. Followed by the stimulus centered on the same gray background at a visual angel of approximately 7°. The stimulus remained on screen until a response was recorded. The response keys were labeled "A" and "B" on the keyboard letters "d" and "k". Participants were told that the maximum time to respond was 2 seconds and that feedback was given together with a point total that were displayed for 800ms. All incorrect categorizations were followed by a low-pitched sound and a red *-1*, whereas all correct categorizations were followed by a cash-register sound ("ka-ching") and a green *+1*. The feedback points were shown in the center of the screen and a total score was shown below in white (*Score: #*). Participants were instructed to respond as quickly as possible without sacrificing accuracy.

Since participants were asked to categorize up to 800 trials, we assigned a goal score of 800 points to ensure participants were motivated to maintain high accuracy and short response times throughout the experiment. Previous research in our lab has shown that participants are motivated to maintain high accuracy when given a goal score at the beginning of the experiment together with the information that once that goal score was reached participants were free to leave the experiment. Furthermore, to ensure stable performance parameters, short-term goals were provided: 5 correct responses in a row would result in *+2* points accompanied by the cash-register sound on the fifth correct response. Failure to respond within 2 seconds after the onset of stimulus presentation resulted in the auditory error feedback of a buzzer paired with the text "Too slow!" and pressing a key unrelated to the experimental set-up led to the same auditory error feedback paired with the text "Incorrect key pressed!" both were shown for 1 second in the center of the screen. Completion times and completion trial were subject to variation due to the fact that achieving the goal score would terminate the experiment at that trial.

2.2.4 RT Analysis Methods

Since RT distributions tend to be positively skewed and subject to large inter-subject variability [49], several non-parametric analyses were chosen to omit strong assumptions about the RT distributions.

Vincentizing

The *Vincent* averaging technique allows for group averaging without distorting the underlying functional form [50]. RTs for each subject were organized in ascending order, quantiles were calculated per participant, and then aggregated into group quantiles. In

this way, information about the shape at the level of the individual subject was preserved in the group average. It is noteworthy that in this technique, midpoints between quantiles were estimated, such that the resulting quantiles were not in one-to-one correspondence to the percentiles, instead the percentiles described the quantile midpoints.

RT Smoothing

To gather a companion picture to accuracy as a function of the distance to the category boundary, RT distributions were estimated with respect to the equal likelihood-contour using the Parzen kernel estimator [51]. First, all RT samples, assumed to be *i.i.d.*, were collected across all participants separately for control and likelihood conditions. Second, the samples in each vector were sorted according to their associated x-coordinate (the relevant categorization dimension), such that the resulting vector contained RT samples ordered with respect to increasing x-values. Thirdly, the shape of this distribution was estimated using the following Gaussian kernel estimator:

$$\hat{f}_x(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h_n}\right), \tag{2.1}$$

where K is the standard Gaussian density:

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2} z^2, \tag{2.2}$$

where n represents the total number of observations, and $h_n > 0$ represents the smoothing parameter equal to the standard deviation of the Gaussian distribution. For any value t, the estimate of the density at that point is given by the Gaussian density centered at that point, so that each RT observation in the sample contributes an amount to the estimate that is proportional to the height of the Gaussian at the point of the estimate. In other

words, this method is a more sophisticated version of the relative frequency histogram, where the estimate is proportional to the sum of the height of the Gaussian densities at that point. The choice of the Gaussian width parameter is arbitrary. However, there exists a trade-off between smoothness and functional form of the estimated RT function.

2.3 Results

All subsequent analyses are focused on the last 500 trials as these guaranteed stable performance parameters (accuracy and RT) and enough data points for robust RT analyses.

2.3.1 Decision Bound Regression

Decision bound regression was performed to determine whether participants employed decision bounds alternative to the equal likelihood-response-contour shown in the top panel of Figure 2.2. All but two participants in each condition were best fit by a unidimensional linear classifier near the optimal equal likelihood-response-contour shown in Figure 2.3. Responses from the divergent four participants that were best fit by the general linear classifier were analyzed in more detail. Scatter plots revealed that these participants did not employ an alternative strategy to the simple one-dimensional rule, instead the distribution of errors was not balanced with respect to the equal likelihood-response-contour and consequentially favored the general linear classifier over the unidimensional bound during model fitting. This result was likely to occur, given that participants were not error-less in 500 consecutive trials. Importantly, the RTs that these four subjects produced were within or below the group average for their condition.

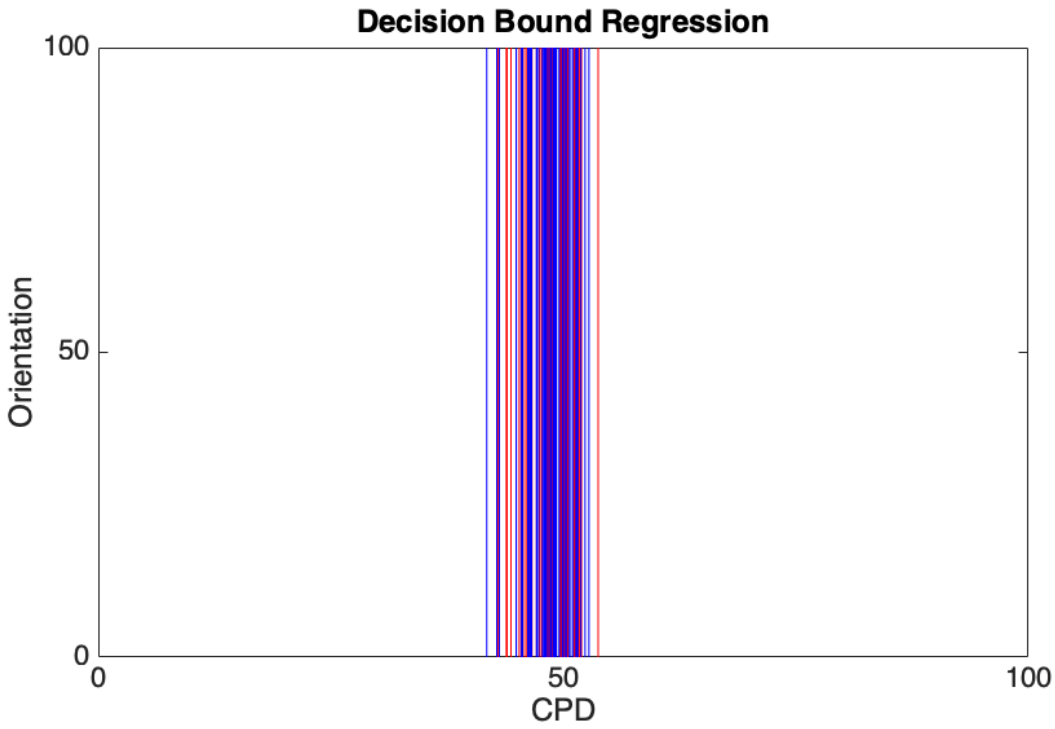

Figure 2.3: Estimated unidimensional linear decision bounds for likelihood (blue) and control (red) conditions.

2.3.2 Accuracy Analyses

Mean and minimum completion trial numbers for the control and likelihood conditions were: $M_C = 763.74, \sigma_C = 38.37; min_C = 710$, and $M_L = 777.98, \sigma_L = 27; min_L = 699$, respectively.

Figure 2.4 top panel shows the forward learning curves for proportion correct and corresponding RTs, it is evident that performance on both parameters stabilized after block 8 (200 trials). The steep transition in the forward learning curves from the first 25-trial block to the next is characteristic of rule-based learning [52]. Specifically, the steep transition suggests that participants were testing explicit hypotheses during the first few trials and converged onto a single hypothesis, such as *narrow bars belong to category A, and thick bars belong to category B*, during the ensuing trials.

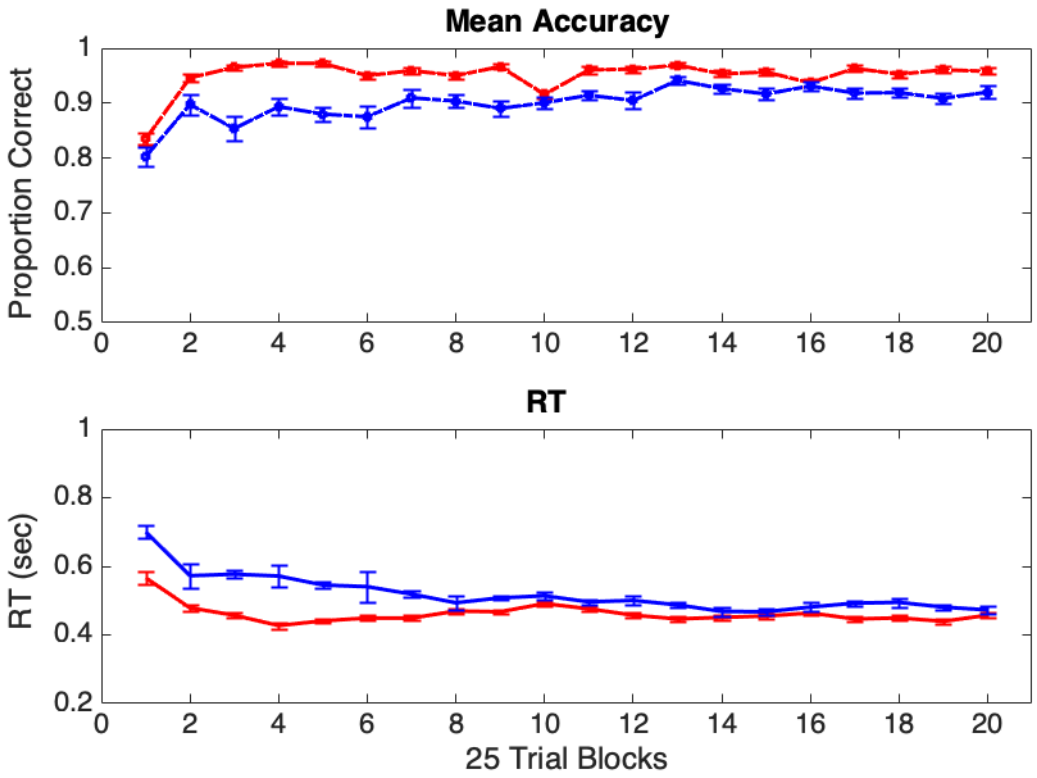

Figure 2.4: Top panel: forward learning curves for control (red) and likelihood (blue) conditions. Each point represents the mean across participants for a 25-trial block. Bottom panel: RTs matching the forward learning curves for control (red) and likelihood (blue) conditions. Each point represents the mean across participants for the median RTs per 25-trial block.

Next, accuracy analyses to explore performance as a function of distance to the hypothetical category boundary were conducted. In that light, accuracy across participants for each sample point along the relevant dimension was computed to maintain the highest resolution, since the perceptual granularity for *thickness* is unknown. First, trials were sorted in ascending order with respect to the x-axis (relevant category dimension). Second, the mean across participants for each x-value was computed. The results are shown in Figure 2.5. A two sample t-test shows there was no significant difference comparing proportion correct for likelihood ($M = .93$) and control ($M = .94$) conditions $[t(392) = 1.8179, p = 0.07, d = 2]$.

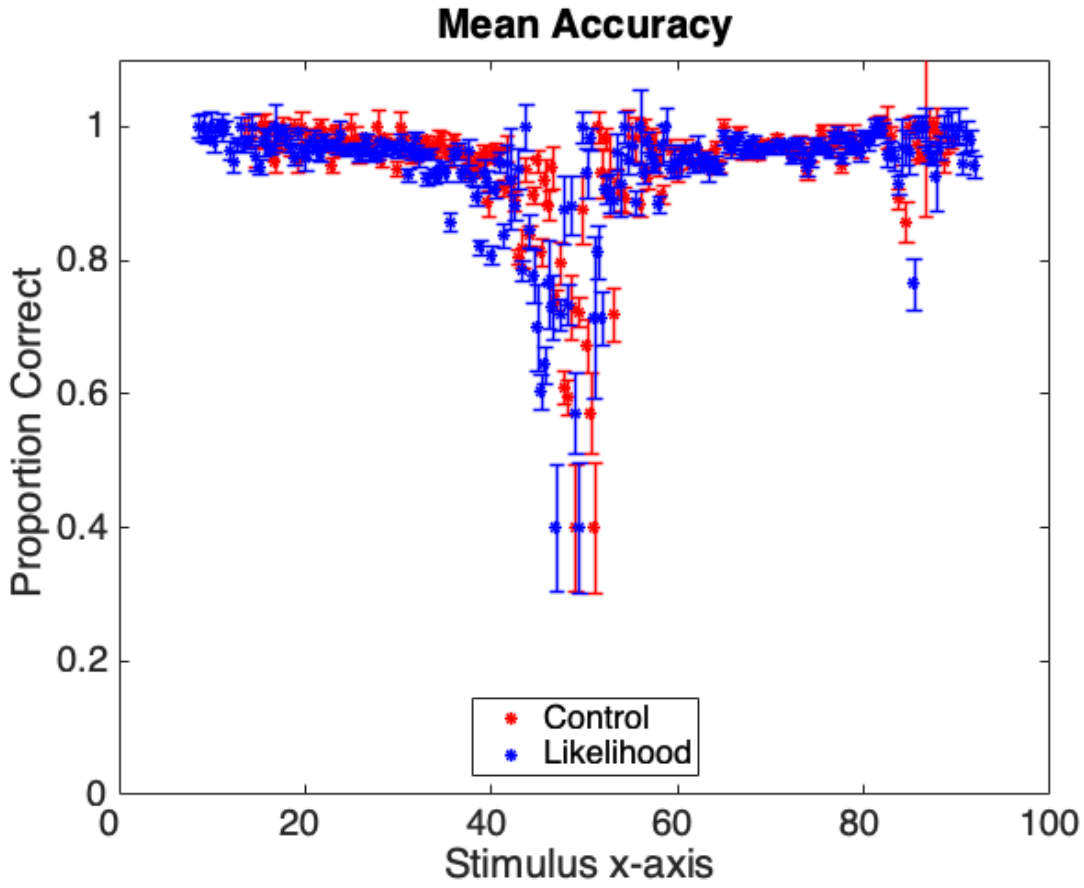

Figure 2.5: Accuracy as a function of distance to the hypothetical category boundary. Each point represents the mean across participants per x-coordinate (relevant categorization dimension).

2.3.3 RT Analyses

Vincent cumulants

To calculate the *Vincent* quantiles, quantile RTs over 25 equally spaced bins for $n_C = 30561$ and $n_L = 25050$ samples in the control and likelihood conditions were estimated. A two-sample Kolmogorov-Smirnof test indicated that the RTs from the control and likelihood conditions belonged to different underlying distributions $D(55609) = 0.015, p < 0.01$. Since the alternative hypothesis for this statistical comparison stated that RT samples from the control condition were drawn from a smaller distribution, this

result implies an ordering at the level of the cumulants. The results for the Vincent quantiles are depicted in Figure 2.6. The Figure does not show an ordering of the Vincent cumulants. However, a clear crossing of functions can be seen at 395ms. The absence of a clear ordering of the two functions rules out a speed-accuracy trade off due to any alternative influence not considered here.

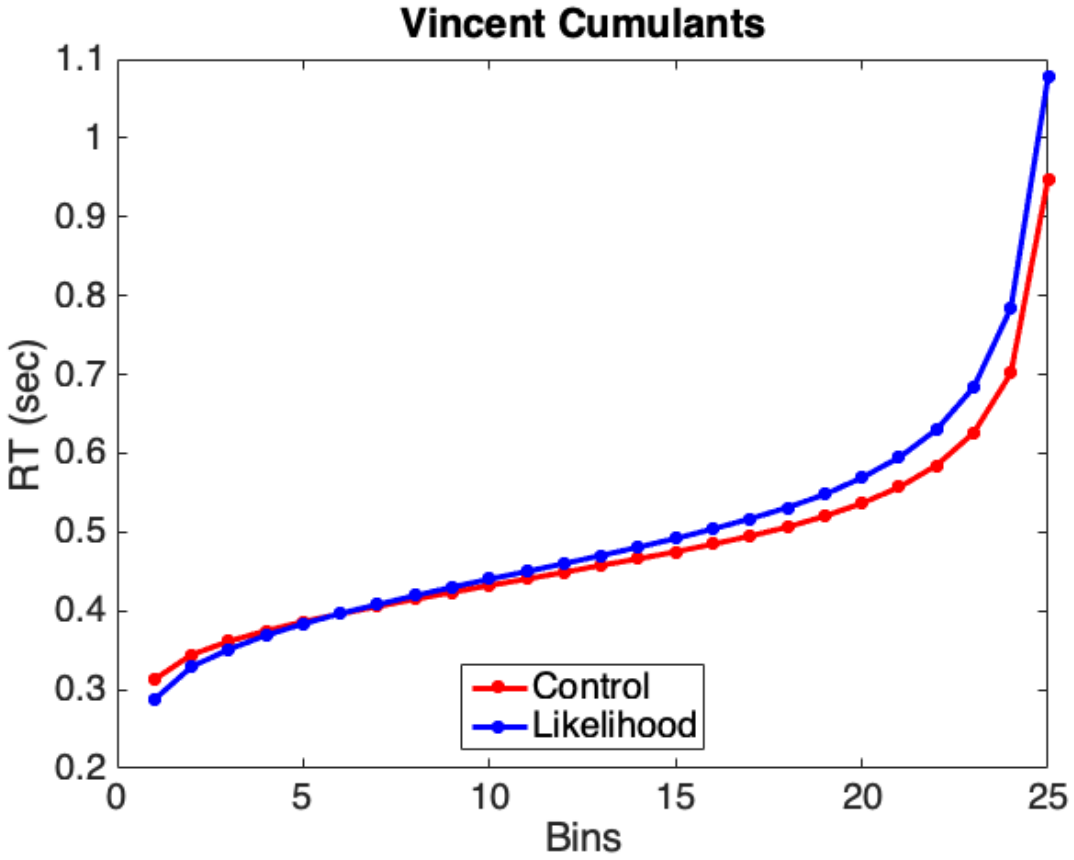

Figure 2.6: Vincent cumulants for control (red) and likelihood (blue) data. Each point represents the midpoint of the RT *Vincent* estimates computed over 25 bins.

Smoothed RT with respect to relevant x-dimension

For RT smoothing with respect to the relevant categorization dimension, RT samples were collected across all subjects for control and likelihood conditions, resulting in vectors of lengths $v_C = 30501$ and $v_L = 25001$, respectively. The RT samples were sorted

with respect to increasing x-coordinates and convolved with a Gaussian kernel to get a smoothed picture of RT behavior with respect to the relevant categorization dimension (or distance to the hypothetical criterion). Since the underlying data vectors differed in size, Gaussian kernels were scaled by 1/6 of the data vector lengths resulting in kernels with $N_C = 5083.5, \sigma_C = 635.31$, and $N_L = 4166.8, \sigma_L = 520.73$ for control and likelihood conditions, respectively. Finally, the smoothed RT estimates were truncated on both ends at 2% of the length of the original RT vector to approximately remove estimates that were denatured by the kernel. The smoothed RT estimates are depicted in Figure 2.7.

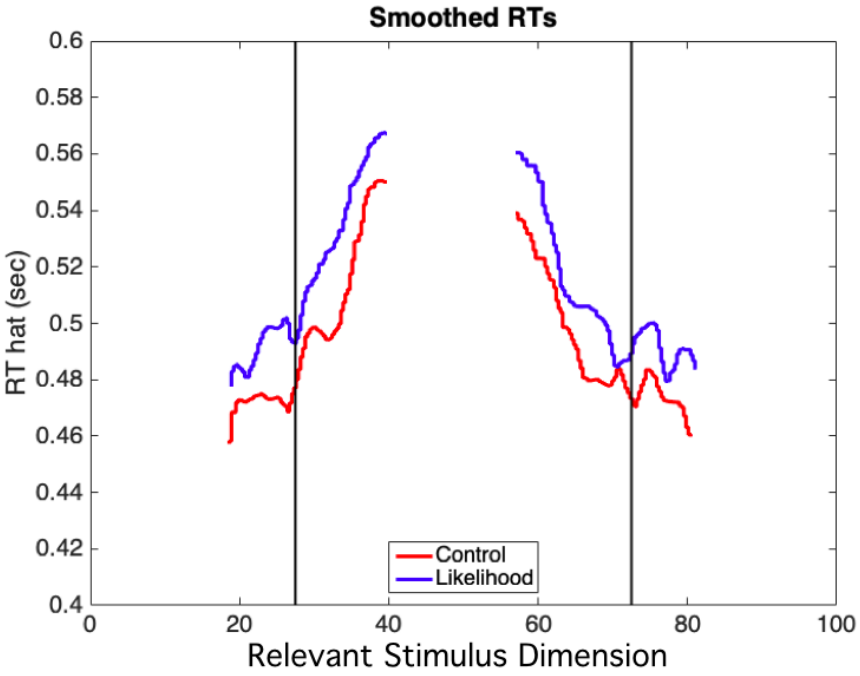

Figure 2.7: Smoothed RT estimates across all subjects for control and likelihood conditions. The vertical solid lines mark the means of the marginal distributions on the stimulus axis relevant to the categorization task.

2.4 Discussion

The data presented above provide evidence that falsify the criterion model as process model in explicit categorization. Furthermore, the data highlight contrasting utility for accuracy and RT measures. Accuracy is sufficient to differentiate whether participants were using procedural or explicit strategies. The steep transition from the first 5-trial block to the next shown in Figure 2.4 is characteristic of rule-based learning [52], and suggested that participants were testing hypothesis during the first few trials before rapidly converging onto the correct categorization rule in ensuing trials. This confirms that participants were, in fact, using explicit strategies in both task conditions. As seen in Figure 2.5, accuracy was monotonically increasing with distance to the equal likelihood-contour dividing stimulus space. This result was consistent with the prediction by the criterion model. However, accuracy is subject to a stark ceiling effect and performance had clearly reached asymptote for the last 500 trials that were analyzed. Therefore, it was not possible to conclusively determine whether accuracy was higher for mean centered stimuli that were associated with a response more frequently than the remaining stimuli. These results imply that the performance parameter *accuracy* may not suffice to reliably differentiate between process models. In contrast, RT offered a far richer perspective.

For the RT analyses, Vincent cumulants depicted in Figure 2.6 did not show a clear ordering which supports general equivalence between the tasks. In other words, alternative explanations for the effects observed above based on a speed-accuracy trade off can be ruled out. Nonetheless, a two-sample Kolmogorov-Smirnof test indicated that the RT cumulants for the control condition were significantly smaller than the RT cumulants from the likelihood condition. It is noteworthy that a comparison at the distributional level is more powerful than comparisons of first moments [49] and that an ordering at the level of the cumulative density function implies the same ordering at the level of the

probability density, and the mean [3, 53]. To gather a more complete picture, smoothed RT functions with respect to the relevant stimulus axis, i.e. as a function of distance to the equal likelihood contour, were depicted in Figure 2.7. Consistent with the significance test above, the functions clearly show that participants in the likelihood condition took longer to respond.

The results presented here falsify the criterion model. First, the criterion model predicts similar RTs in both conditions because category judgements are solely based on comparisons between the sample and a criterion and should ignore the disrupting effects of region inconsistent feedback in the likelihood condition. In contrast, participants in the likelihood condition produced significantly longer RTs than participants in the control condition. Second, the criterion model predicts monotonically decreasing RTs with increasing distance to the criterion. The Figure 2.7 clearly shows that the prediction of monotonicity for RTs is violated. Furthermore, this violation persists for all participants individually. In contrast to the criterion model, these results support a *direct-mapping* model which highlights the regions separated by the criterion. Interestingly, the RTs for all stimuli in the likelihood condition were longer compared to the control condition, even though the disparate feedback was spatially constrained. There are two possible explanations for this observation. One possibility is that perceptual noise associated with the stimuli that received disparate feedback spatially smeared across the entire perceptual space. Another possibility is that a cognitive system which is sensitive to probabilistic feedback such as the procedural system bootstraps the representations in the explicit system. A final possibility is that the disparate feedback increased interference from a cognitive system that monitors prediction errors. The slowing in RTs may have resulted from either of these possibilities or a combination.

2.4.1 Perceptual categorization models

Perceptual processing strategies are associated with different decision models. Two major decision models are independent decision models and information integration models [17]. Independent decision models assume a two-stage process in which decisions are made on each perceptual dimension separately according to some criterion before a response is selected. While information integration models assume a one-stage process that combines all available perceptual information in order to select a response. Evidence from patients with neurologic disorders, e.g. amnesia [54], as well as behavioral [40] and neuroimaging [55] experiments support the notion that implicit [56] and explicit memory are mediated by distinct memory systems [57, 58], which are subject to different constraints [59, 43], and that implement disparate reinforcement strategies [28].

In the field of category learning a neurobiological hybrid model of these decision processes named COVIS has been developed [37]. COVIS assumes two interacting memory systems [36]: a procedural system that gradually learns to associate the information integration decision process with a motor response, and a rule-based system which utilizes a hypothesis testing mechanism in the form of a discriminant function that is constructed around a decision criterion. The procedural system, centered about the basal ganglia [60], is governed by a dopaminergic reinforcement learning algorithm [28] that maximizes accuracy with an implicit similarity-based strategy. The prefrontal connected rule-based system depends on declarative memory and recruits executive attention and working memory processes [36, 42, 43]. One hallmark of explicit decision processes is that the applied rule is verbalizable [36], and individual participants begin rule-based tasks with large discrete jumps of the hypothetical rule boundaries that become incremental at later stages of learning [61]. To date, no complete theories exist of how the explicit system quickly learns and flexibly switches between behaviors in rule-based tasks.

Rule-based tasks involve at least two cognitive operations: rule selection and criterion learning [37]. Rule selection concludes the hypothesis testing phase during which different hypotheses that include one or more perceptual dimensions are tested. Thus, rule selection is synonymous with selecting the relevant stimulus dimension. Criterion learning is assumed to ensue from the rule selection phase. Criterion learning describes convergence onto the partition of perceptual space that maximizes accuracy, however, no neurobiological account of this process exists. The COVIS explicit system generates a trial-by-trial discriminant value that globally quantifies the output. COVIS pre-selects the rule (discriminant function on the relevant dimension) and absorbs the criterion learning phase into the selection process. In 2015, Helie and colleges developed a biologically inspired criterion learning circuit, e.g. HICL [62]. HICL proposes a heterosynaptic circuit that modulates input of the lateral PFC rule-cells in a global fashion. Interestingly, their model implies that the criterion is an epiphenomenon, i.e. it is not explicitly represented at the neurobiological level. Our results presented above are consistent with this prediction. While the HICL can account for extra- and intra-dimensional shifts of relevance during categorization, it remains an open question of how rules are actually represented and selected [63].

The results presented here violate the criterion assumption and discredit models in which participants parametrically estimate a decision criterion to make category judgements in explicit rule-based tasks.

2.4.2 Direct-Mapping Model

The direct mapping model builds on prototype theory or category *typicality* (e.g. category representations are continuous structures in which category membership is graded by degree of similarity of the entire object or object features [64] and relies on trial-by-

trial confusion-based judgments that depend on response-region-specific mappings. These mappings and their strength are contingent on the similarity between the current stimulus and previously seen stimuli and whether that judgment receives positive feedback. For stimuli from bivariate normal distributions, where samples cluster around the mean, the mean of the bivariate stimulus distribution becomes the category-prototype and the mean of the marginal distributions on each stimulus dimension becomes the category-feature-prototype. Rosch et al. [64] showed that category members with an increasing degree of typicality are classified with monotonically increasing accuracy and monotonically decreasing RT. Thus, category members that are more similar to the category-prototype or category-feature-prototype are classified more accurately and faster than category members that are less similar. Additionally, Rosch et al. [64] predicted that frequency (i.e., stimulus repetition) should have an effect equivalent to typicality, and that frequency effects may not always be separable from category structure such that the frequency of features could become a structural variable. The direct-mapping model predicts that accuracy should be highest and RT lowest for highly practiced stimuli that are consistently associated with the same response. Specifically, response-region inconsistent feedback should interfere with mapping strength and result in increased RTs. The results presented here are consistent with the direct-mapping model, which favors a more region based account. However, the nature of the representation remains elusive. In the next chapter, electroencephalography correlates of explicit category learning in humans were explored.

Chapter 3

EEG Correlates of Rule Based Decision Making

3.1 Introduction

It is well known that visual information of object perception underlies a hierarchical processing structure, which supports the idea that tuning of individual neurons increases in complexity when moving upward through the processing hierarchy [65]. Visual categorization is a complex process that involves: perception, attention, working memory, a decision about which category label to associate with the stimulus, and a response. During decision making, sensory representations serve as inputs to a computational machine that converges onto a decision output on a sub-second scale [66]. Electroencephalography (EEG) provides high temporal resolution metrics of information processing that may be sensitive enough to explore feature selective modulations underlying decision making processes during categorization tasks, and provide insight beyond population level inferences that can be drawn from voxel-level functional magnetic resonance imaging analyses [67]. However, this temporal sensitivity comes at a cost: EEG has a hefty limitation in

the spatial domain due to non-negligible volume conduction of the head [68].

Recent evidence from EEG experiments in the field of spatial visual working memory and attention have provoked the following experiment. Several independent studies have discovered that spatial mental representations can be reconstructed from the topography of EEG oscillations in the alpha band (8-12Hz) [69, 70]. More specifically, induced alpha was found to track spatial location in visual working memory representations during a delay working memory task [69]. Traditionally, the working memory literature has focused on direct measures, such as quantifying capacity [41, 71, 72, 73], rather than characterizing the representational medium itself. More recently, however, the quest to discover some fundamental capacity limit has shifted to characterize working memory as a dynamic process of limited storage [44] that is subject to internal noise [74, 75]. One mechanistic explanation suggests that activity in the alpha frequency band reflects a visual working memory coordination mechanism that depends on the synchrony of neural populations within that frequency band [76]. However, detailed descriptions of working memory representations remain a topic for debate [73, 77].

The field of category learning has very recently expanded to include EEG experiments. In 2015, an EEG experiment was conducted to establish EEG correlates in the form of event related potential that dissociate procedural from rule-based categorization tasks [78]. Even more recent, event related potentials were used to compare simple unidimensional categorization rules to conjunctive rules [79]. They found differential mean amplitudes in ERP components when comparing a simple unidimensional to a more complex conjunction rule in frontal and fronto-central electrodes. These experiments are only beginning to scratch the surface of the intersection between EEG and category learning. To my knowledge, the experiment presented here is the first EEG experiment to investigate the nature of rule representation during a unidimensional explicit categorization task using an exploratory approach including spectral analyses of alpha and theta frequency

bands.

The experiment was a within subjects design consisting of three sessions. In session 1, participants performed a two alternative forced choice unidimensional explicit categorization task [26]. During rule learning each participant developed an unbiased categorization strategy, and rule application over the course of 600 trials established robust rule representations. The purpose of the other sessions was to provide intra-individual EEG correlates that would mimic region (session 2) and criterion (session 3) based decision processes. In session 2, participants performed a five alternative forced choice absolute identification task on stimuli drawn from the same stimulus space used in session 1. In session 3, participants performed a two interval forced choice rotation discrimination task. This task is a spin off from the Sternberg memory scanning task [47] with the memory set size reduced to one. All tasks included a delay period of 1 second that was inserted between the stimulus presentation and the categorization response.

3.2 Method

3.2.1 Participants

In this experiment participants were 16 healthy UC Santa Barbara undergraduate and graduate students (8 females; age range: 19-31). All participants self-reported normal or corrected-to-normal visual acuity and provided written informed consent for the three session experiment in accordance with the human participants Institutional Review Board at UCSB and were monetarily compensated for their participation. The sample size is in line with previous studies that investigated the role of induced alpha in visual working memory [69]. All participants had normal or corrected to normal vision.

3.2.2 Categorization Stimuli

The stimuli were gray-scale circular sine-wave gratings that varied across trials on two dimensions: spatial frequency (cycles per degree of visual angle; CPD) and orientation (radians of counterclockwise rotation from horizontal; rad). All stimuli subtended 7° of visual angle and were displayed against a gray background.

Session 1: Categorization Experiment Stimuli

In session 1, the categorization experiment (e.g. two alternative forced choice task), the two categories were defined by three orientation locked distributions each. The orientation parameters are specified in Table 2.1. The orientation distributions for each category were separated by 0.14 rad (8°) and both categories were separated by 0.28 rad (16°) centered about the criterion located at 0.4555 rad (26.1 deg). The exact orientation bins were chosen arbitrarily with the only constraint that no generic orientation rule (e.g. 45°) could be used as criterion to partition the two categories.

Session 2: Identification Experiment Stimuli

In session 2, the identification experiment (e.g. five alternative forced choice task), five orientation locked distributions were separated by 0.2121 rad (12°) measured from the horizontal. The center distribution is set at 0.4555 rad (criterion location of the categorization experiment in session 1) and the the two distributions at either end (e.g. 0.0314 rad and 0.8796 rad) had identical orientation coordinates as in the categorization experiment session 1.

Session 3: Memory Delay Rotation Discrimination Experiment Stimuli

In session 3, the memory comparison experiment (e.g. two interval forced choice rotation discrimination task), the orientation distributions from session 1 were used, including the distribution located at the criterion. This experiment included memory probe stimuli that shared the same CPD coordinate (2.6334). Five memory probes were selected corresponding to the five orientation distributions used as stimuli minus the two distributions at either end.

Stimulus values for the orientation dimension are shown in table 3.1. For all session stimuli, the values along the CPD dimension were transformed using a non-linear transform (e.g. $x_t = \frac{x}{100*3} - 1$; $x_{new} = 2^{x_t}$; [80]) with minima and maxima set at 3.2490 and 0.6156, respectively. The minima and maxima were chosen to reduce the variability in the CPD dimension and control for unintended stimulus driven activity for extremely large or extremely small CPD values. During the experiment, each (CPD, rad) coordinate pair was used to create a sine-wave grating using Brainard's [48] Psychophysics Toolbox. All participants in each session were presented with the same stimuli. However, presentation order was randomized.

Table 3.1: Orientation Parameters of Experimental Stimulus Distributions in radians (rad)

Session 1		Session 2	Session 3	& *Probe*
A_3	0.0314	[1] 0.0314	0.0314	
A_2	0.1728		0.1728	0.1728
A_1	0.3142	[2] 0.2435	0.3142	0.3142
$Crit$		[3] 0.4555	0.4555	0.4555
B_1	0.5969	[4] 0.6676	0.5969	0.5969
B_2	0.7383		0.7383	0.7383
B_3	0.8796	[5] 0.8796	0.8796	

3.2.3 Procedure

Each participant took part in three experimental sessions that took place in a dimly lit room. Participants sat in front of a computer screen with a keyboard or mouse for responding. The stimulus distributions and task flow for all sessions are shown in figure 3.1. The procedure for each session is described in detail in the following.

Session 1: Categorization Experiment Procedure

Prior to session 1, participants were informed that they would be categorizing circular sine-wave gratings belonging either to category A or B, and that category membership would become apparent through feedback. Each trial began with a blue fixation dot at a visual angle of approximately 0.2° centered on the screen, together with a green circle at a visual angle of approximately 0.4° which signified the participant′s gaze. In order to initiate the trial the participants had to align their gaze (green circle) with the blue fixation dot and press the space bar. Once the trial was initiated, a circular sine-wave grating was displayed at fixation at a visual angle of approximately 7° on the same gray background for 500ms. Next, a 1000ms retention interval accompanied by a gray screen followed. After the retention interval a display with the response options “A” and “B” appeared centered vertically on right and left sides on the screen corresponded to marked keyboard letters “d” and “k” at a visual angle of approximately 10°. The response options appeared in randomized order to prevent participants from anticipating a response location and generating a preparatory motor signal. If a wrong key was pressed, “Wrong key pressed!” was displayed in red for 1000ms, the trial was aborted and added to the end of the block. This session included 600 stimuli that were partitioned into 24 blocks á 25 trials.

Session 2: Identification Experiment Procedure

Prior to session 2, participants were informed that they would identify circular sine-wave gratings numbered 1 through 5 based on their orientation. The distributions were numbered in increasing order with increasing angle from the horizontal. Each trial began with the same trial initiation screen, followed by a 500ms display of the stimulus and a 1000ms retention interval as before. After the retention interval a display with the response options, numbers 1 through 5 at a visual angle of approximately 10° were shown. Again, the response options appeared in randomized order to prevent participants from anticipating a response location and generating a preparatory motor signal. Participants had to select the response by moving a cursor over the desired response and eliciting a mouse-click. If the cursor was not clearly placed on a number, "Click out of bound!" was displayed in red for 1000ms and that trial was added to the end of the block. This session included 600 stimuli that were partitioned into 24 blocks á 25 trials.

Session 3: Memory Delay Rotation Discrimination Experiment Procedure

Prior to session 3, participants were informed that they would compare a second sine-wave gratings to a previously shown one. Participants had to initiate the display of the first stimulus, the memory probe, which was shown for 500ms. Subsequently the trial initiation screen appeared again and participants initiated the trail as in the sessions before. The stimulus was shown for 500ms, followed by a 1000ms retention interval, and then the response options "Up" and "Down" appeared centered vertically on right and left sides on the screen corresponded to marked keyboard letters "d" and "k" at a visual angle of approximately 10°. Participants had to judge whether the second stimulus shown was rotated up or down relative to the memory probe that was shown first. The response options appeared in randomized order. If a wrong key was pressed, "Wrong key

pressed!" was displayed in red for 1000ms and that trial was added to the end of the block. This session included 595 trials that were partitioned into 17 blocks á 35 trials.

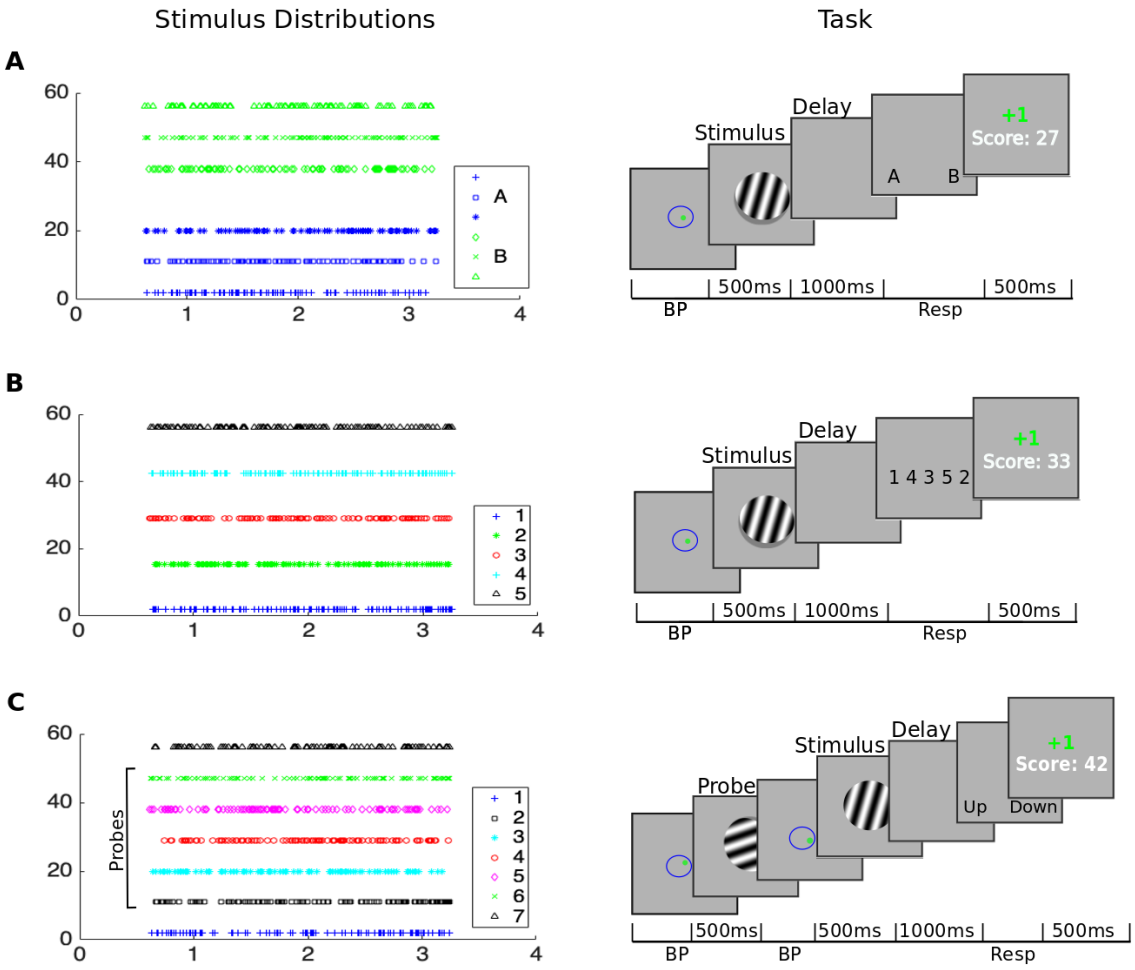

Figure 3.1: Stimulus distributions and task flow for all sessions. Categorization session 1 (row A), identification session 2 (row B), and memory delay session 3 (row C).

In all sessions, participants were told that the maximum time to respond was 2 seconds. Feedback was given together with a point total and displayed for 500ms. All incorrect categorizations were followed by a red -1, whereas all correct categorizations were followed by a green +1. The feedback points were shown in the center of the screen and a total score was shown below in white (*Score: #*). Participants were told that all trials would be gaze contingent and that blinking before initiating the trial

is recommended. The scripts that oversaw the experiments were created on MATLAB (version 2015b, Massachusetts, TheMathWorks Inc.) using Brainard's [48] Psychophysics Toolbox.

3.2.4 EEG Neurophysiology

EEG measures neuro-electric waveforms produced by the brain's dynamic behavior at multiple spatial and temporal scales. The first human measurements were conducted by Hans Berger in the 1920s [81]. Although the neurobiological underpinnings of EEG including event related potentials are not entirely known, it continues to be accepted that EEG as well as intra-cranially recorded local field potentials arise from postsynaptic activity of neural ensembles that manifest as extracellular ion fluxes between extracellular fluid and the pyramidal neurons (and other participating cells such as glial cells) in the vicinity [68, 82].

In order to be able to measure the electric potentials generated by postsynaptic potentials, a phalanx of cells have to be active simultaneously and their geometry relative to the recording electrode is highly constrained, e.g. dipoles with perpendicular orientation in close proximity to the electrode generate voltage fluctuations that are detectable [68]. Since electro-magnetic fields are instantaneous with respect to the sampling frequency of the electrode, the EEG signal has remarkable temporal resolution. However, it is virtually impossible to estimate the number and location(s) of the underlying dipole(s) (or neural generators) that contribute to the measures voltage fluctuations [83]. For instance, the electrical artifacts caused by saccadic eye movements produce voltage fluctuations of 16 Volts per degree of saccade that can be picked up by electrodes on the back of the head. Furthermore, volume conduction of the anisotropic layers of the skull (eg. bone - cerebrospinal fluid - bone) cause current shunting due to the favorable conductive

properties of the cerebrospinal fluid. For instance, one scalp electrode measures space averaged potentials from 10^9 neurons [68].

Neocortex represents the largest contributor to scalp EEG recordings: excitatory connections occur in superficial layers, while inhibitory connections occur in deeper layers of cortex. Cognitive functions are associated with cell assemblies that spontaneously form, disconnect, and re-connect on timescales ranging from $10-100ms$. These cell assemblies are embedded in and interact with global fields to form global networks. EEG recordings are biased to global field activity due to volume conduction and the organization of cortical layers resulting in large amplitude low frequency and small amplitude high frequency recordings. Therefore, oscillations in lower frequency bands, such as alpha and theta, are more robust to noise [68].

Relevant Brain Oscillations

In the following relevant brain oscillation are briefly introduced (for review see Klimesch [84]). A more comprehensive review of oscillatory phenomena can be found in Nunez and Srinivasan [68].

Alpha rhythms $(8-12Hz)$ occur ubiquitously and dominate scalp EEG recordings of healthy adults. These oscillations are spontaneously enhanced during relaxation with closed eyes and are more pronounced in posterior electrode locations. Alpha amplitude typically ranges from 15-45 V at posterior recording sites while amplitudes at frontal recording sites are considerably lower. Interestingly, adults show larger anterior alpha amplitudes and anterior-posterior coherence compared to children. While moderate alpha amplitudes are reported in frontal electrodes of healthy relaxed adults, large amplitudes may co-occur with pathologies such as trauma, disease, and anesthetics. Some alpha activity is blocked by the opening of the eyes, drowsiness, and when a task is experienced as difficult. For instance, Klimesch [85] reports a story about the EEG recordings of

Albert Einstein: "Einstein produced continuous alpha waves during the solution of complex mathematical tasks which he was skilled and trained to solve. When his alpha-band activity suddenly disappeared, he was asked what had happened, he replied that he had just become aware of a mistake in the calculations he had recently done."

In general, alpha activity is thought to reflect a superposition of global activity from large dipole layers that are measured at most electrode sites, and smaller dipole layers at more local electrodes [68]. People tend to display an individual alpha peak frequency that is modulated by age and neurological disease [84]. Importantly, scalp EEG recordings of alpha rhythms are ubiquitous and result from spatial averages over an unknown number of underlying components. It is therefore impossible to identify which component(s) are active in synchrony over which superficial cortical area at a particular instance in time. What further complicates inference about an underlying physiological process is the existence of large individual differences.

Alpha activity has been associated with perception [86], working memory [87], and attention [85]. Furthermore, the finding that posterior alpha power is proportional to working memory load [87] and the inverse relationship between alpha power and fMRI blood-oxygen-level dependent (BOLD) activity [88] have inspired the conclusion that alpha activity is ubiquitously related to inhibition and information selection processes [85].

The alpha rhythm has been described as a quasi-stable and bi-modal phenomenon as some frequencies are differentially modulated by activity and have approximately disparate scalp topographies. Low alpha ($8 - 9Hz$) was found to decrease in amplitude with mental activity, while high alpha ($10 - 12Hz$) and theta activity (discussed below) showed increased amplitudes at frontal electrodes. Again, large variance between individuals has been observed in the alpha band such that inter-individual differences can be as large as age-related differences [84]. For instance, some people show distinct bi-modal

alpha peaks in frontal electrodes while other do not [68]. These large inter-individual differences further complicate robust inferences about an underlying mechanism of these oscillations.

The result that inspired the experiment presented here describes a correlation between working memory related memoranda and induced brain activity in the alpha frequency band $(8 - 12Hz)$ [69, 89]. Induced activity is defined as a correlate of the experimental task that is not strictly phase-locked to any task variable, in contrast to evoked activity which is related to specific task variables [90]. To characterize induced alpha activity presented here, the total activity (e.g. evoked plus induced) in the alpha frequency band was extracted in two disparate approaches presented below. The analyses were computed separately for individual participants because non-negligible individual differences exist, particularly in the alpha frequency band [84].

Theta rhythms $(44-7Hz)$ are much less prominent in the human scalp EEG, however, these rhythms are strongly represented in the hippocampus of animals and increases in animals as well as humans with memory demands [84]. Theta is enhanced during deep relaxation and sleep. These oscillations are important during brain development in early childhood and decrease throughout puberty, while they increase in late adulthood [84]. Neurological disease such as dementia has been associated with increased theta power during wakefulness, while theta power increases during deep sleep and with the administration of Melantonin [84].

Decreased theta and increased upper alpha power together with increased theta and decreased upper alpha coherence have been observed in anterior electrodes during mental activity [68] and coherence during encoding predicted successful recall [84]. Intriguingly, task performance could be inferred from resting state because large pre-stimulus alpha power predicted a large decrease in coherence during the task, while small pre-stimulus theta correlated with a large increase in coherence [91]. Further, coupling between theta

and gamma ($40 > Hz$) bands correlate with short-term memory processing [68]. Theta frequency covaries with alpha frequency [84] and together frequencies in the alpha ($10Hz$) and theta ($7Hz$) range have been associated with attentional sampling to support a cyclic perceptual parsing mechanism in vision [86]. Both theta and alpha band activity has been associated with memory and attention processes, however, the exact purpose of these oscillations remain unknown. It is likely that cognitive activity associates with multiple alpha and theta frequency networks, whose measurements are biased to the global synaptic fields that dominate the scalp EEG [68].

3.2.5 EEG Acquisition

EEG data were recorded for each participant using a Brain Products ActiCHamp system (Brain Vision LLC, Morrisville, NC) with 64 Ag-AgCl sintered active electrodes that were uniformly distributed in an actiCAP (Electro-Cap, USA) elastic cap and placed in accordance with the 10/20 System [92]. The TP9 and TP10 reference electrodes were adhered directly to the right and left mastoids. Data were sampled at 1000 Hz. At the beginning of each session, all impedances were ¡15 k.

3.2.6 Eye Tracking

Gaze contingent eye tracking accompanied the trial from initiation until the response options appeared. Trials on which the participants broke their gaze by blinking or moving their gaze farther than 3° from fixation were aborted and added to the end of the block to ensure that every participant saw the same stimuli. Aborted trials were followed by the trial initiation screen with the blue fixation dot. During the task, participants were positioned in a chin rest at approximately 120cm from the monitor (19-in. ViewSonic E90f CRT). The eye-tracker (Eyelink 1000 plus, SR Research Ltd. Mississauga, Ontario,

Canada) was positioned approximately 60 cm from the right eye (monocular tracking @ 1000Hz, mean error ¡1°).

3.2.7 EEG Data Pre-processing

EEG data analyses were performed offline in MATLAB (version 2015b, Massachusetts, TheMathWorks Inc.) with the EEGLAB v. 14.1.1 toolbox [93]. First, the HEOG and EKG channels were removed and all electrodes were referenced to the average of the left and right mastoid electrodes. Next, noisy channels were identified via visual inspection of the raw data in temporal space as well as their representation in frequency space. The noisy channels were then removed and replaced using spherical interpolation.

Then the pre-processing pipeline to select the relevant frequencies and task sections diverged depending on the analysis that followed. For ERP analyses the continuous EEG recording was bandpass filtered with cutoffs of 0.1 and 30 Hz using Eeglab 'firws' filter function, which implements a zero-phase FIR filter of order 200 and tapers the data with a hamming window prior to convolution with the filter kernel. After filtering the continuous signal was dissected into 700ms epochs ranging from 200ms before stimulus onset throughout the 500ms of stimulus presentation. For session 3, 'stimulus onset' describes the test stimulus shown after the memory probe.

For the spectral and autocorrelation analyses the continuous EEG recording was fist separated into condition specific 1 second memory delay intervals (or epochs) and tapered using a Hamming window to alleviate edge artifacts in the small data sample. For the spectral analysis, the epochs were Fourier transformed and power values were retrieved via the periodogram (see details in section *Fourier Analysis*). For the autocorrelation analysis, the epochs were filtered using an FIR filter that was designed using the Matlab 'designfilt' function. To isolate alpha frequencies the following parameters were selected:

a passband of $8-12Hz$, stopband frequencies set at 7 and $13Hz$ with 60dB equal ripple stopband attenuation. For theta the 60dB equal ripple stop-band cutoff was set at $3Hz$ and $8Hz$, while the passband was $4-7Hz$. These parameter setting showed the best results for the trade-off between passband ripple and transition band roll-off.

In general, baseline corrected epochs with voltage fluctuations beyond -75 and 75 Volts at any electrode were rejected to exclude prominent myogenic and occulo-muscular artifacts. Since brain activity of interest is typically contained in scalp EEG amplitudes between 0 and 45 Volts [90, 68] no cerebral EEG activity of interest was removed. Furthermore, small muscular artifacts that bypassed this thresholding procedure and remaining artifacts with sharp transitions were identified via visual inspection of the raw as well as filtered signals and removed. Any potentially remaining myogenic and occulo-muscular artifacts are of little concern because these affect higher frequencies (e.g. 20 - 300 Hz) [90, 94].

3.3 Analysis Methods

3.3.1 Analysis of Evoked Event Related Potentials

When event related potentials (ERPs) were introduced, EEG correlates of cognitive processes became popular around the 1970s. ERPs quantify neuro-electric responses that are evoked by a sensory stimulus. However, events in the external environment that provoke characteristic evoked potentials are confounded with the spontaneous background noise of the brain. Therefore, several trials need to be averaged to generate ERPs with robust components. An ERP was defined as [82]: *scalp-recorded neural activity that is generated in a given neuroanatomical module when a specific computational operation is performed*. The overall ERP waveform shape consists of consecutive peaks and throughs

that reflect sums over several independent components. In this sense, amplitude and timing of observed maxima and minima are relatively meaningless. The amount of trials necessary to generate reliable representations depends on the ERP components of interest and their associated sensory modality.

For instance, the visual N1 component describes a large negative Voltage deflection that typically peaks in anterior electrode sites around $100 - 150ms$ post stimulus, while two posterior components observed in parietal and occipital electrode sites follow the anterior peak with a delay of approximately $50ms$. The N1 component is thought to be associated with visual discrimination, perceptual processing, and expert recognition [95]. Since the number of trials necessary to estimate robust N1 components takes approximately 300-1000 trials [95], any analysis with less than that should be interpreted with great caution. The largest problem in ERP analysis is the confident association between ERP component and experimental manipulation since the EEG recording at any point in time represents a sum over several processes and components that overlap in space and time. For instance, the two hemispheres may contribute differentially to an ERP component of interest [82]. Fortunately, intra-subject variability of ERP waveforms is small compared to inter-subject variability. Nonetheless, ERP component analysis is further complicated by an intra-trial additive effect due to component overlap and that ERP activity does not follow the same time course across trials and subjects. These unavoidable sources of variance result in grand group averages with smaller amplitudes in which waveforms are smeared in time [95]. Therefore, to confidently separate ERP components that arise from feed-forward or feed-back processing is virtually impossible. However, some ERP components occur early enough that it is likely that they result from feed-forward processes, while the later components are likely a mixture of feed-forward and feed-back processing.

Following the recent results from Rabi et al. [79], the N1 over frontal and fronto-

central scalp sites were investigated. In their exploratory investigation comparing ERPs of unidimensional and conjunctive rules during categorization, their results show mean amplitude differences of the N1 component and late positive complex. The analysis presented here will focus on the mean amplitude of the N1 component over anterior (frontal and fronto-central) electrode sites. The time-window for the N1 for each orientation bin was empirically defined by the component zero-crossings in a grand average across all subjects and conditions. The window size for the zero-crossings satisfy the recommended minimum length of 40ms [82] and defining the window latency via grand averages is recommended as alternative approaches based on individual subjects has been shown to bias the significance tests [96]. Lastly, one important rule of ERP analysis is to never assume a linear or even monotonic relationship between an ERP components' amplitude or latency and the quality or timing of an underlying cognitive process [82].

3.3.2 Frequency Domain Analyses of Induced Activity

Induced activity describes a neuroelectric correlate of the experimental task that is not strictly phase-locked to any task variable. Spectral analysis decomposes temporal waveforms into frequency components. Frequency representation has several advantages over standard time representation in which the signal was recorded. The most obvious benefit is the visualization of periodicities in the recorded signal that may correlate to some underlying physical phenomenon. Spectral analysis of EEG data became popular in the 1970s and replaced the traditional analysis method of counting zero crossings [68].

Theoretical Background and Assumptions

From the perspective of analytical mathematics, it is convenient to regard EEG measures as stochastic processes and sample EEG recordings as single realization of such a

process [97, 68]. Assuming the underlying process is a bandlimited stationary random process then it can be sampled without aliasing and filtered out [98]. Note the stochastic nature of EEG shall not imply that the process has no statistical structure, rather it implies that this structure has not been characterized. The following describes a mathematical theory that provides the necessary framework for spectral analyses and is not to be considered as a biophysical model of EEG.

A stochastic (or random) process is the ensemble of all the possible realizations (e.g. sample functions) that the underlying stochastic phenomenon of study could have generated. Consider the stochastic process $X(t)$ which represents a series of realizations, such that they construct an ensemble. $X(t)$ follows a density that depends on time, and when t is fixed, i.e. $t = t_o$ then $X(t_o)$ becomes a random variable of the density $f(x_{t_o})$. Then for some integer N, the random variables $X_1, ..., X_N$ are defined by the stochastic process $X(t)$ at times $t_1, ..., t_N$. If the joined density function of $(X_1, ..., X_N)$ does not depend on time, then the stochastic process $X(t)$ is stationary. It was pointed out that to gather an ensemble of neuro-electric waveforms it would require the simultaneous acquisition of an approximately infinite number of brains with identical statistical properties [99]. In application this is clearly not possible and one realization is not sufficient to characterize the stochastic process. However, assuming the phenomenon under investigation is a stochastic process and assuming that EEG recording samples are realizations of that stochastic process provides the theoretical framework for investigating the phenomenon of interest.

A further necessary assumption for the spectral analysis of the EEG sample data is time-invariance (or signal stationarity). That is that the signal mean and variance (e.g. power spectrum) are assumed to be independent of time. The assumption of stationarity is likely violated due to the neocortical source dynamics described above as well as overall state changes in vigilance, caffeine effects, etc. However, the assumption

of stationarity for induced or spontaneous activity becomes more viable when the time window of interest is shrunk to 1 second [92]. In contrast, the assumption of stationarity is clearly violated in evoked event related potentials that follow some sensory input (see *Section Event Related Potentials above*), whereby one exception are steady state evoked potentials [68]. The stationarity assumption precludes robust detection of transient information (for example synchronous inhibitory volleys that do not occur on every oscillation cycle [100]). Finally, brain activity measured using EEG recordings can be regarded as non-stationary stochastic processes in space and time whereby short sample of EEG data, for instance an EEG epoch, may be thought of as one realization of some underlying stochastic phenomenon that satisfies the stationarity assumption.

Fourier Analysis

Fourier analysis was developed by Jean Baptiste Fourier (1768-1830) and is essential to the analytical armamentarium of mathematical, physical, and natural sciences. Since the development of the efficient Fast Fourier Transform (FFT; [101]) this analysis has gained computational allure. The Fourier transform decomposes a real signal into a linear superposition of sines and cosines at particular frequencies. Since most recorded signals, including the EEG signal analyzed here, are discrete I will only present the *discrete Fourier Transform*. For a signal represented by N discrete samples $x(n) = x_o, x_1, ..., x_{N-1}$ the discrete Fourier Transform is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-i2\pi kn}{N}}, k = 0, ..., N-1 \tag{3.1}$$

and its inverse is defined as:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{\frac{i2\pi kn}{N}}. \tag{3.2}$$

When the decomposed time series is real, the discrete frequencies are defined as: $f_k = k/N$. Thus, the frequency resolution depends on the frequency resolution of the original signal and the length of the segment that is analyzed. There are two errors that can occur when applying the discrete Fourier Transform: aliasing and leakage [102]. The highest detectable frequency within a period of N samples is called the *Nyquist frequency* and corresponds to $f_{Nyquist} = N/2$. If the signal contains frequencies above the $f_{Nyquist}$, these will be forced into the frequency range $f_k < f_{Nyquist}$ causing spurious effects known as *aliasing*. Leakage refers to a broad-band artifact at the window border. In general, leakage artifacts are prevented by applying a windowing function, commonly a *Hanning* window, to the signal before the Fourier Transform is computed [97].

The periodogram computes the power spectrum of the signal. The power spectrum represents the signal variance at particular frequencies:

$$I(k) = |X(k)|^2 . \tag{3.3}$$

Since the periodogram is not a consistent estimator [97], in order to gather a robust statistical estimate of the power spectral density the periodograms of several EEG sample realizations were averaged [68]. The power spectrum has been referred to as the Fourier transform of the autocorrelation function [103].

For the spectral decomposition of the 1 second delay interval EEG epochs, the epochs were first shortened to remove any evoked potentials. Specifically, the first 200ms were removed since the stimulus disappearing from the screen usually results in evoked activity that continues until approximately 150ms after stimulus offset [90]. Next, a Hanning windowing function was applied to each epoch before the Fourier Transform was computed using the FFT algorithm. Since the EEG recordings commenced in an unshielded environment and one wall was shared with an animal research laboratory, the sampling

rate of the signal was kept at $Fs = 1000ms$ to omit potential aliasing due to unknown high frequency sources. After computing the power spectra per epoch, all epochs for each orientation bin were averaged per participant per condition and the relevant frequencies in alpha $(8 - 12Hz)$ and theta $(4 - 7Hz)$ bands were extracted resulting in a 4-dimensional matrix [participants x condition x EEG channel x frequency bins] for each frequency band.

Autocorrelation Analysis

Assuming that the EEG epoch of interest resembles a zero mean stationary random signal, the autocovariance C_{xx} of that signal sampled at equally spaced intervals δt is defined by [97]:

$$C_{xx}(k\delta t) = E\left[x(n\delta t)x((n+k)\delta t)\right]. \tag{3.4}$$

When the signal is stationary, the right side depends on k only. And when $k = 0$, we get the variance:

$$C_{xx}(0) = E\left[x(n\delta t)^2\right]. \tag{3.5}$$

The autocorrelation function is the standardized autocovariance [97]:

$$R_{xx}(k\delta t) = \frac{C_{xx}(k\delta t)}{C_{xx}(0)}. \tag{3.6}$$

In contrast to the autocovariance the autocorrelation is scale independent. The caveats associated with using the autocorrelation function on the known alpha and theta frequency bands are: 1) alpha and theta activity itself is periodic, and 2) the onset of induced alpha is likely not time-invariant across trials.

In oder to compute the autocorrelation functions, the 1 second delay EEG epochs were first filtered to select the frequency band of interest. For each trial, the autocorrelation

was computed over the last 800ms of the delay interval to reduce the presence of evoked activity due to stimulus-offset. For alpha, autocorrelations were computed at 4 time-lags, comparing $200ms$ intervals, which produced a 4-dimensional matrix [participants x condition x EEG channel x time lag]. It is noteworthy, that 4 time lags is less than the recommended number of 20 lags [102]. However, 4 time lags provided a time window that should capture approximately two alpha-cycles and seemed appropriate given that the frequency of the signal was known.

3.4 Results

The two interval forced choice rotation discrimination task and the similarity based $(A, notA)$ forced choice identification task provide the data for a presumed cognitive process underlying orientation judgements relative to a previously shown criterion orientation or relative to a previously established rule region, respectively. The similarity between EEG correlates of these different tasks is quantified in the following. For all analyses, the number of trials between conditions are subject to great variability. For instance, in session 2, the five alternative forced choice identification task, motor noise has contributed to trial loss. While in session 3, corrupted data files have largely contributed to the stark difference in data loss between conditions. All analyses presented in the following only include correct categorization trials.

3.4.1 ERP Analysis

After artifact rejection, the artifact free number of trials for each participant varied between conditions and orientation bins. For session 1, the average number of trials for orientation bin 1 were $M = 54.0909$ with $\sigma = 15.5528$, and for orientation bin 7 they were $M = 56.0909$ with $\sigma = 17.184$. For session 2, the average number of trials for

orientation bin 1 were $M = 73$ with $\sigma = 27.0592$, and for orientation bin 7 they were $M = 52.6364$ with $\sigma = 20.9918$. And for session 3, the average number of trials for orientation bin 1 were $M = 23.0909$ with $\sigma = 9.2136$, and for orientation bin 7 they were $M = 16.2727$ with $\sigma = 4.7136$. Since the ERP results presented here were generated from differential numbers of trials between conditions, the interpretation of results ought to be approached with caution [82].

In accordance with Rabi et al. [79], the ERP analyses were restricted to frontal and fronto-central electrodes (23 electrodes in total). The ERPs for grand averages across participants and electrodes are depicted in figure 3.2, for orientation bins 1 and 7. The figure clearly shows the presence of pre-stimulus alpha in session 3 in contrast to session 1 and 2. With respect to ERPs, alpha activity not related to the stimulus event represents the largest source of noise [95].

The N1 component mean amplitude was computed for each orientation separately. The window latencies based on grand averages across subjects and conditions were $94 - 147ms$ and $87 - 151ms$ for orientation bin 1 and 7, respectively. The resulting window size for orientation bin 1, $w1 = 54ms$, and for orientation bin 7, $w7 = 65ms$, satisfied the length recommendation by Luck [82]. Figure 3.3 shows topographic images for the grand average maxima of orientation 1 (maximum at 119ms) and orientation 7 (maximum at 120ms). The 23 frontal and fronto-central electrodes were averaged prior to an overall analysis of variance. The analysis showed no significant difference in mean amplitude of the N1 component between conditions for orientation bin 1 ($F(2, 10) = 0.057, p = 0.945, \eta^2 = 0.003$), as well as orientation bin 7 ($F(2, 10) = 1.415, p = 0.266, \eta^2 = 0.066$).

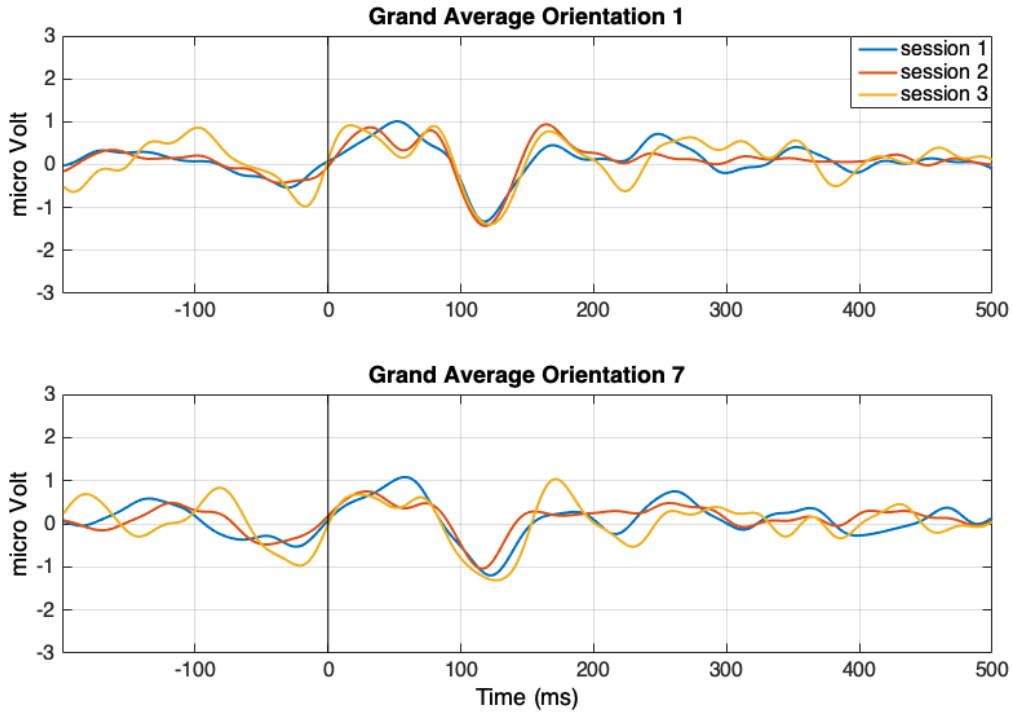

Figure 3.2: Grand averages of frontal and fronto-central electrodes for orientation bin 1 (top) and orientation bin 7 (bottom). Abscissa shows the time in ms relative to stimulus onset and ordinate references the amplitude in μV. Session 1 (blue trace), session 2 (red trace), and session 3 (yellow trace).

3.4.2 Spectral Analysis

After artifact rejection, the trial numbers for the delay interval varied between conditions and orientation bins. For session 1, the average number of trials for orientation bin 1 were $M = 86.18$ with $\sigma = 28.5$, and for orientation bin 7 they were $M = 89.64$ with $\sigma = 29.65$. For session 2, the average number of trials for orientation bin 1 were $M = 73.64$ with $\sigma = 24.02$, and for orientation bin 7 they were $M = 55.46$ with $\sigma = 23.94$. And for session 3, the average number of trials for orientation bin 1 were $M = 22.82$ with $\sigma = 5.58$, and for orientation bin 7 they were $M = 16$ with $\sigma = 5.75$.

The power values are depicted in a topographic manner referenced to the electrode position on the scalp. The results for alpha power are seen in figure 3.4, and the results for

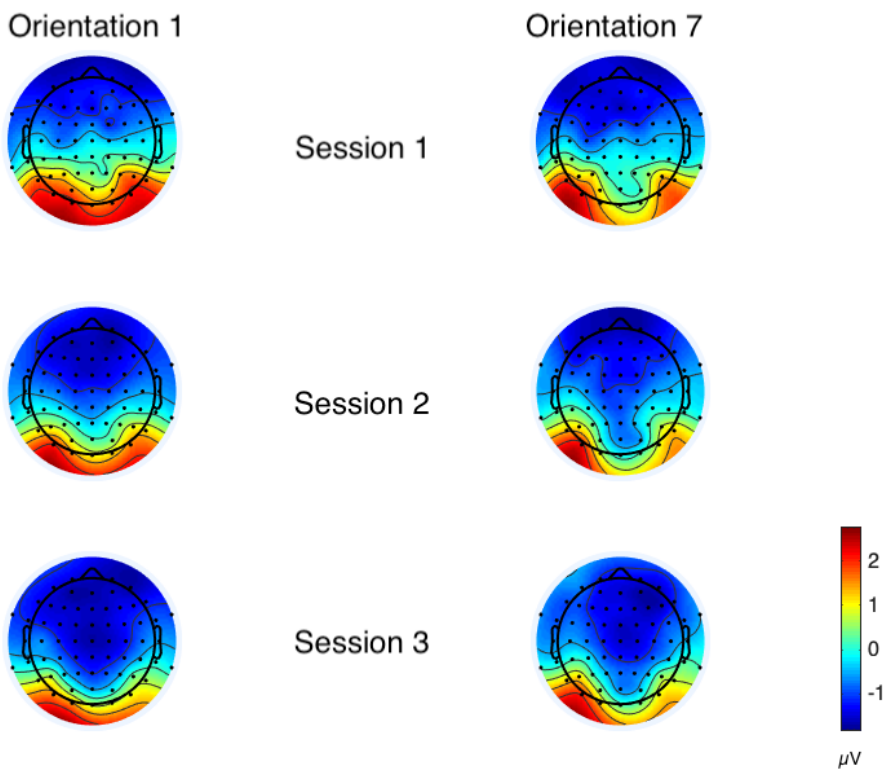

Figure 3.3: Topographic plots of ERPs at N1 maxima defined by grand average across subjects and conditions. Orientation bin 1 (left column) orientation bin 7 (right column), session 1 (top row), session 2 (middle row), and session 3 (bottom row). The colorbar references μV.

theta power are depicted in figure 3.5. Power was differentially distributed between the two frequency bands, with alpha power being highest in posterior electrodes, and theta power was highest in anterior electrodes. Although within frequency bands differences between tasks and orientation bins were evident they were far less pronounced. Power values across electrodes were averaged prior to all repeated measures analyses of variance.

The analysis of variance to investigate significance of the factor condition showed no significant difference for alpha power averaged across all electrodes between conditions for orientation bin 1 ($F(2,10) = 0.741, p = 0.489, \eta^2 = 0.005$) as well as orientation bin 7 ($F(2,10) = 1.606, p = 0.226, \eta^2 = 0.020$). To quantify significance separately for

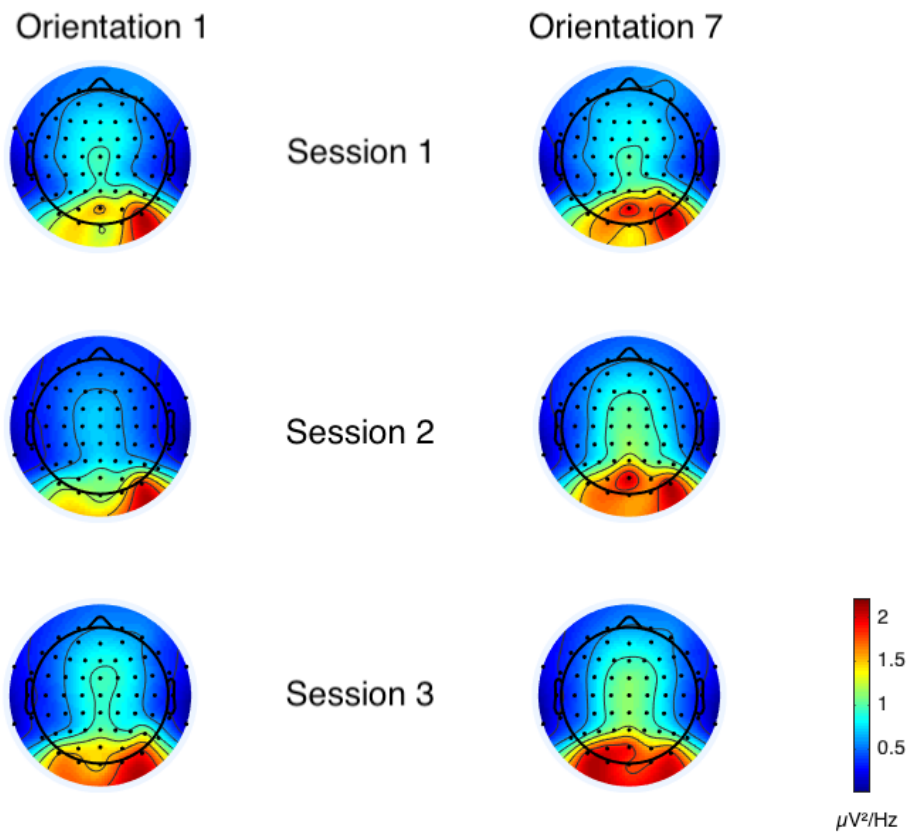

Figure 3.4: Topographic plots of alpha power averaged across participants and 800ms delay periods. The columns depict alpha power for orientation bin 1 (left) and orientation bin 7 (right), while the rows reference the tasks: session 1 (top row), session 2 (middle row), and session 3 (bottom row). The colorbar on the bottom right references the power values in $\mu V^2/Hz$.

selected topographic regions, electrode clusters in frontal, central, parietal, and occipital regions were investigated. The analyses of variance did not reveal any significant effect for condition for either orientation bin. Specifically, no significant difference for power averaged across a subset of frontal electrodes (AFz, FCz, F1, F2) between conditions was found for orientation bin 1 ($F(2, 10) = 0.772, p = 0.475, \eta^2 = 0.008$), as well as orientation bin 7 ($F(2, 10) = 1.987, p = 0.163, \eta^2 = 0.028$). No significant difference for power averaged across a subset of central electrodes (CPz, FCz, C1, C2) between conditions was found for orientation bin 1 ($F(2, 10) = 0.645, p = 0.536, \eta^2 = 0.005$), as well as orientation

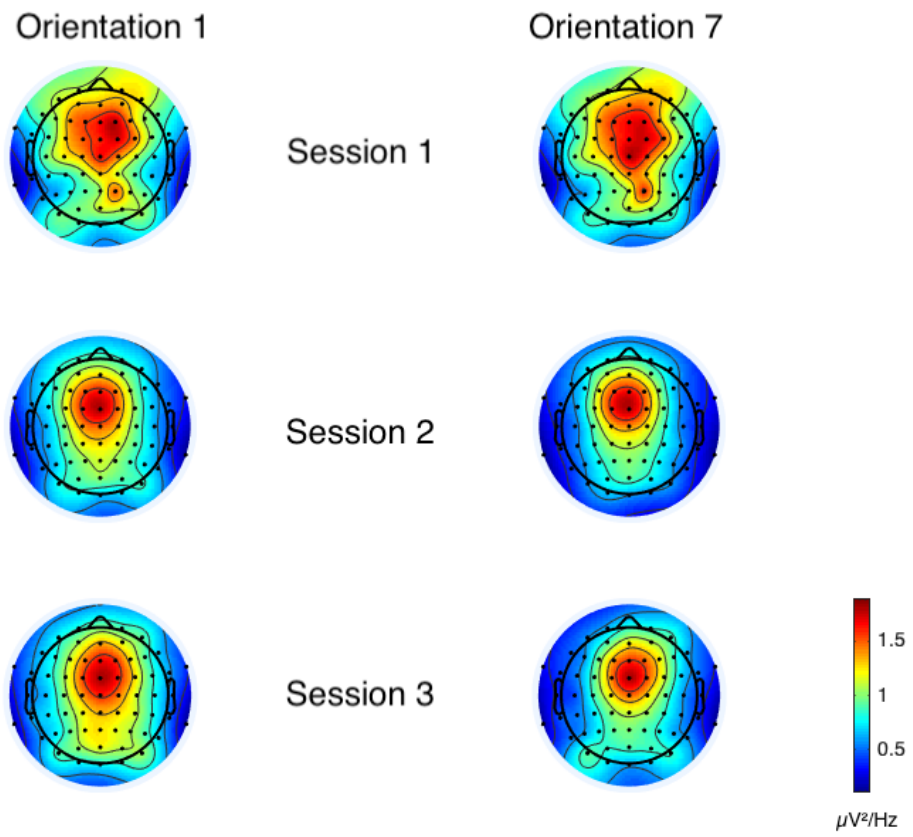

Figure 3.5: Topographic plots of theta power averaged across participants and 800ms delay periods. The columns depict theta power for orientation bin 1 (left) and orientation bin 7 (right), while the rows reference the tasks: session 1 (top row), session 2 (middle row), and session 3 (bottom row). The colorbar on the bottom right references the power values in $\mu V^2/Hz$.

bin 7 ($F(2,10) = 3.089, p = 0.068, \eta^2 = 0.028$). No significant difference for power averaged across a subset of parietal electrodes (CPz, POz, P1, P2) between conditions was found for orientation bin 1 ($F(2,10) = 0.379, p = 0.689, \eta^2 = 0.003$), as well as orientation bin 7 ($F(2,10) = 2.11, p = 0.147, \eta^2 = 0.015$). No significant difference for power averaged across a subset of occipital electrodes (POz, O1, O2) between conditions was found for orientation bin 1 ($F(2,10) = 0.926, p = 0.412, \eta^2 = 0.007$), as well as orientation bin 7 ($F(2,10) = 1.185, p = 0.326, \eta^2 = 0.014$). Since alpha frequencies appear to be differentially modulated (see Section the analyses above were repeated to

investigate high alpha ($10 - 12Hz$). Non significance of all tests above was maintained for high alpha.

The analysis of variance for theta power averaged across all electrodes showed a marginal significant difference after sphericity corrections between sessions for orientation bin 1 ($F(2, 10) = 4.251, p < 0.05, \eta^2 = 0.051$). While a greater significant effect was found for orientation bin 7 ($F(2, 10) = 5.391, p < 0.05, \eta^2 = 0.043$). For orientation bin 1, post-hoc Bonferroni corrected paired comparisons showed a marginal non-significant difference between theta power for session 1 ($M = 290.78$) and session 3 ($M = 402.88; p = 0.052$), while a marginal significant difference between session 1 and session 2 ($M = 359.48; p < 0.05, d = 0.36$) was found. For orientation bin 7, post-hoc Bonferroni corrected paired comparisons showed theta power differed significantly between session 1 ($M = 293.26$) and session 2 ($M = 383.11; p < 0.05, d = 0.41$), as well as session 1 and session 3 ($M = 412.81; p < 0.05, d = 0.34$).

To quantify which topographic locations drove the significant effect, several subsets of electrodes in frontal, central, and parietal regions were compared. The analysis of theta power averaged across a subset of frontal electrodes (AFz, FCz, F1, F2) showed a significant difference for power between conditions for orientation bin 1 ($F(2, 10) = 6.233, p < 0.01, \eta^2 = 0.06$), as well as orientation bin 7 ($F(2, 10) = 3.714, p < 0.05, \eta^2 = 0.043$). For orientation bin 1, post-hoc Bonferroni corrected paired comparisons revealed a significant difference between session 1 ($M = 435.57$) and session 2 ($M = 612.46; p < 0.05, d = 0.5$), as well as session 1 and session 3 ($M = 644.98; p < 0.05, d = 0.44$), while no significant difference was found between session 2 and session 3. For orientation bin 7 post-hoc Bonferroni corrected paired comparisons showed a significant difference existed only between session 1 ($M = 445.14$) and session 2 ($M = 729.22; p < 0.05, d = 0.54$).

The analysis of theta power averaged across a subset of central electrodes (CPz, FCz, C1, C2) showed a significant difference for power between conditions for orientation bin

1 ($F(2,10) = 6.445, p < 0.01, \eta^2 = 0.073$), as well as orientation bin 7 ($F(2,10) = 5.245, p < 0.05, \eta^2 = .047$). Post-hoc Bonferroni corrected paired comparisons revealed a significant difference between mean power across electrodes for orientation bin 1, between session 1 ($M = 402.95$) and session 2 ($M = 577.31; p < 0.05, d = 0.52$), as well as session 1 and session 3 ($M = 622.53; p < 0.05, d = 0.48$). For orientation bin 7, post-hoc Bonferroni corrected paired comparisons showed a significant difference between session 1 ($M = 432.62$) and session 2 ($M = 646.23; p < 0.05, d = 0.48$), and a marginally non-significant difference between session 1 and session 3 ($M = 687.48; p = 0.066, d = 0.37$).

The analysis of theta power averaged across a subset of parietal electrodes (CPz, POz, P1, P2) showed a marginal significant difference for power between conditions for orientation bin 1 ($F(2,10) = 4.202, p < 0.05, \eta^2 = 0.072$), and a significant effect for orientation bin 7 ($F(2,10) = 4.981, p < 0.05, \eta^2 = 0.045$). Post-hoc Bonferroni corrected paired comparisons revealed a significant difference between mean power across electrodes for orientation bin 1, between session 1 ($M = 329.87$) and session 2 ($M = 440.91; p < 0.05, d = 0.47$), and a marginally non-significant difference between session 1 and session 3 ($M = 509.09; p = 0.055, d = 0.39$). For orientation bin 7, post-hoc Bonferroni corrected paired comparisons showed a marginally significant difference between session 1 ($M = 349.18$) and session 2 ($M = 464.24; p < 0.05, d = 0.43$), and a significant effect between session 1 and session 3 ($M = 482.79; p < 0.05, d = 0.41$).

3.4.3 Autocorrelation Analysis

The number of trials per session are the same as for the spectral analysis above. To investigate the evolution of the alpha frequency, autocorrelations for the last 800ms of the delay interval were computed at 4 time lags. The condition specific autocorrelation statistics were then averaged across trials. By its nature, the autocorrelation function

decreases monotonically with increased time lag and averaging across participants and electrodes did not produce a particularly interesting result between sessions. However, autocorrelation statistics for each electrode evolved differentially, and thus, cross-sections of the autocorrelation functions were investigated. The figure 3.6 shows the correlation statistics on the ordinate plotted for each electrode on the abscissa averaged across participants for the last two time lags. The figure clearly shows larger variance across electrodes for session 2 and session 3 in comparison to session 1. In addition, two troughs corresponding to occipital (Oz, O1, O2) and occipital-parietal (PO3, PO4, PO7, PO8, POz) electrodes are evident. Both effects become more pronounced at the last time lag.

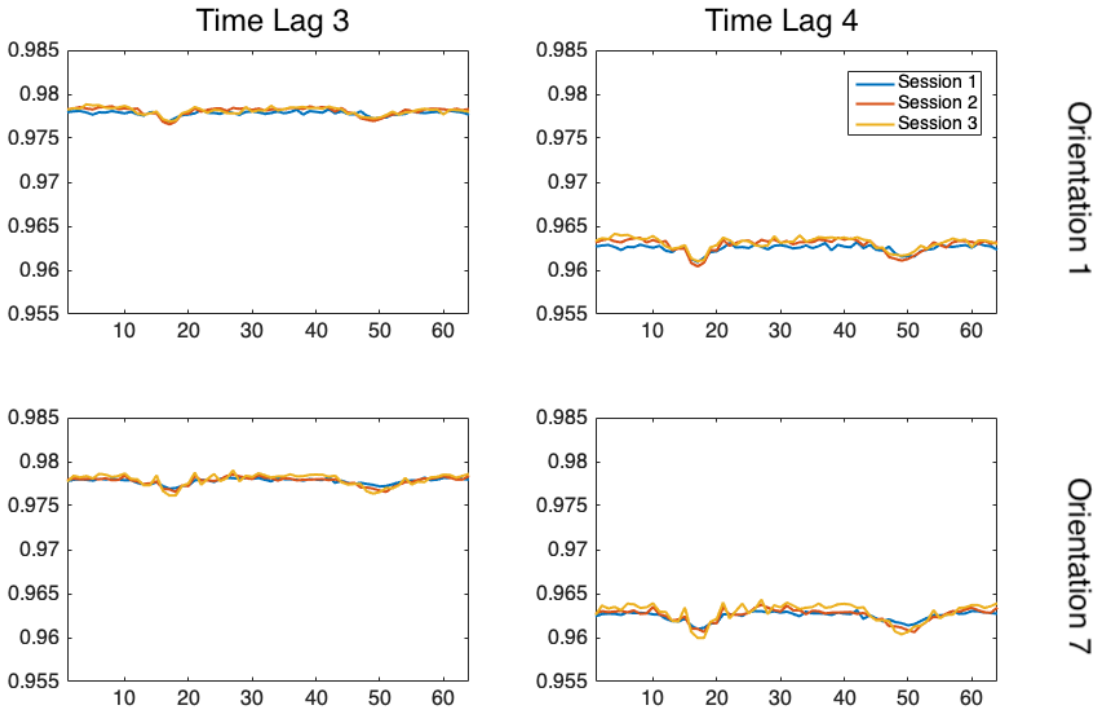

Figure 3.6: Alpha autocorrelation cross-sections. Cross-sections at lag 3 (left) and 4 (right) for orientation bin 1 (top row) and orientation bin 7 (bottom row). The ordinate references the autocorrelation statistic and the abscissa indexes electrodes.

Two-sample T-tests comparing the autocorrelation values of all electrodes between sessions revealed a significant difference between session 1 ($M = 0.9625$) and session 2

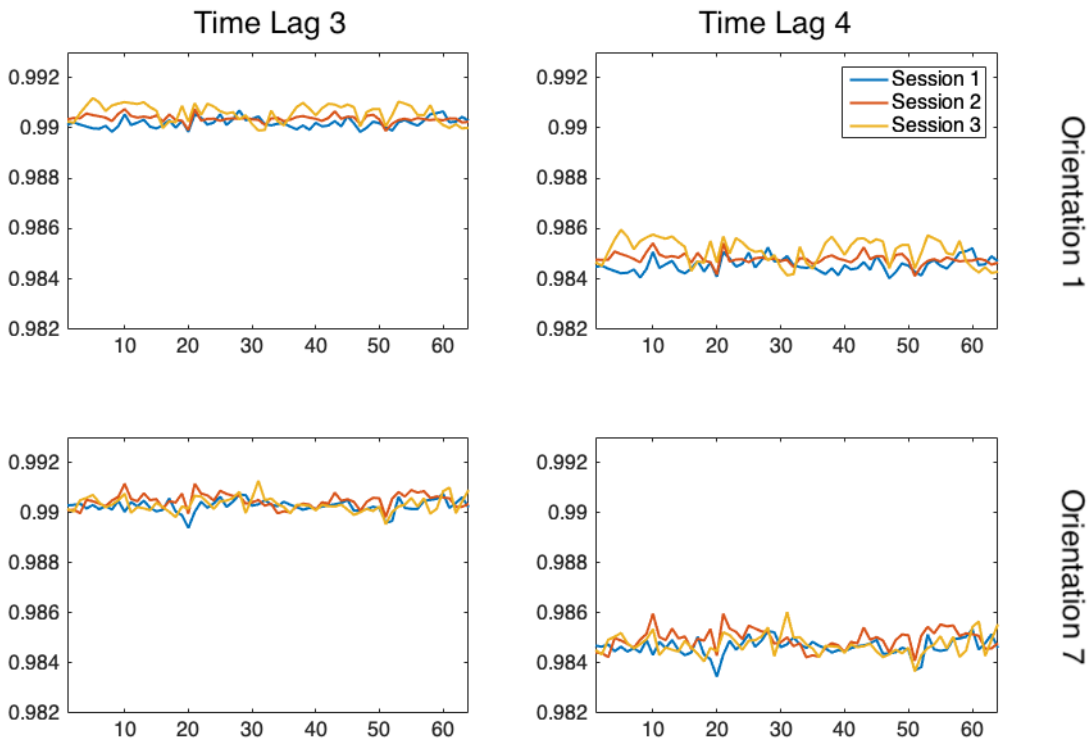

Figure 3.7: Theta autocorrelation cross-sections. Cross-sections at lag 3 (left) and 4 (right) for orientation bin 1 (top row) and orientation bin 7 (bottom row). The ordinate references the autocorrelation statistic and the abscissa indexes electrodes.

($M = 0.9629; p < 0.01, d = 0.005$), as well as session session 1 and session 3 ($M = 0.9631; p < 0.001, d = 0.0048$), while no significant difference was found between session 2 and session 3 ($p = 0.121$) for orientation bin 1. For orientation bin 7, a significant difference was found only between session 1 ($M = 0.9625$) and session 3 ($M = 0.9628; p < 0.05, d = 0.0067$), while no significant differences were found between session 1 and session 2 ($M = 0.9626; p = 0.419$), as well as session 2 and session 3 ($p = 0.076$). It appears that these differences are driven by the increased variance in autocorrelation statistics across all electrodes in session 2 and session 3 compared to session 1. The significance tests suggest a consistently greater difference between session 1 and session 3 in contrast to the difference between session 1 and session 2.

To investigate the evolution of the theta rhythm, autocorrelations for the last 800ms of the delay interval were computed at 4 time lags. The condition specific autocorrelation statistics were then averaged across trials. Again, cross-sections of the autocorrelation functions were investigated. The figure 3.7 shows no clear trend to dissociate electrodes by variance. Specifically, only two right posterior electrodes (P8 indexed by x=20, and PO8 indexed by x=51), show a decreased correlation statistic that is consistent across sessions and orientation bins. Two-sample T-tests comparing the autocorrelation values of all electrodes between sessions revealed a significant difference between all sessions for orientation 1. Specifically, session 1 ($M = 0.9846$) and session 2 ($M = 0.9848; p < 0.001, d = 0.002$), session 2 and session 3 ($M = 0.9851; p < 0.0001, d = 0.003$), and session 1 and session 3 ($p < 0.001, d = 0.0032$). For orientation bin 7, a significant difference was found between session 1 ($M = 0.9846$) and session 2 ($M = 0.9849, p < 0.0001, d = 0.003$), as well as session 2 and session 3 ($M = 0.9847, p < 0.01, d = 0.0032$). While no significant difference was found between session 1 and session 3 ($p = 0.1884$).

3.5 Discussion

Since no clearly defined EEG correlates of rule representation during explicit categorization exist, this analysis was an exploratory endeavor. In the ERP analysis, no significant differences were found comparing the mean amplitude of the N1 ERP component between conditions. The N1 component has been associated with perceptual processing, expert recognition, and visual discrimination [95]. For instance, Vogel and Luck [104] found evidence that supports the hypothesis that the posterior N1 (which peaked around 160ms) reflects some general visual discrimination process, that is not modulated by temporal expectation [105]. In contrast, Vogel and Luck concluded that the origin of the anterior N1 (which peaked around 100ms) is less clear although in their

study the component appeared to be modulated by response-related activity. It is possible that the anterior N1 reflects an executive attention signal for a specific location in space, however, this has not been established.

It is important to note that waveforms elicited by visual stimuli as well as motor outputs can last up to 1 second [95, 82]. In the paradigm presented here, the inter-trial-interval was short and each trial began with a button press that was immediately followed by the visual stimulus. Therefore, waveforms elicited by previous activity are included in the pre-stimulus baseline and are likely to overlap with the stimulus. Furthermore, if the underlying phenomenon of interest is independent of the alpha rhythm, then alpha represents the largest source of noise corrupting the ERP analysis [95]. In particular, the strong presence of pre-stimulus alpha in session 3 may have resulted from the two interval nature of the task where successful performance was contingent on the memory of an immediately preceding stimulus. In this sense, the null result comparing the anterior N1 is favorable because it may provide some control for task equivalence between sessions 1, 2, and 3.

The finding that induced alpha activity in posterior channels was sufficient to track disparate spatial locations [69] could not be extended to characterize the categorization rule. This result may in part be due to the fact that spatial locations are associated with specific retinotopic mappings [68] and spatial attentional priority maps [106] that are supported by a differential processing stream [107]. Differential topographic dominance of alpha power (strongly represented in posterior electrodes) and theta power (strongly represented in frontal electrodes) implicate theta as candidate for top-down control, while alpha may reflect a more localized information selection mechanism. Although alpha activity appears to be inversely related to the amount of available cortical resources [108], it has been assigned a prominent role in attention processes [85].

Strong inferences are cautioned since scalp recordings of alpha oscillations represent

space averages of multiple processes that support cognitive operations including dynamic long-range coherence between regions that carry relevant information and desynchronization of regions that carry irrelevant information [68]. Therefore, it is likely that the tasks in the three sessions were too complex to generate neuroelectric correlates of the ubiquitous alpha rhythm to support robust dissociations. In addition, EEG spectra result from a superposition of white (non-interesting activity due to artifacts), pink (underlying unstructured component), and colored noise (task relevant activity in form of characteristic spectral peaks) [97] that produce topographic measurements subject to variance. For instance, the absence of theta measurements in posterior electrodes does not imply an actual absence, rather it implies a strong alpha dominance at posterior electrode sites [68]. Similarly, alpha rhythms may contribute to the cognitive task at hand without producing outstanding alpha power measured at the scalp. Indeed, frontal alpha activity may produce low amplitude measurements due to highly specialized and localized activity. This is consistent with the nonsignificant result for upper alpha in frontal electrodes. It is possible that more localized activity at frontal electrodes did not have sufficient power, or the localized activity was too similar to generate dissociations.

One intriguing finding is that alpha power did not differ significantly between sessions while the autocorrelation comparisons at the last time lag did. Significance tests suggested a greater difference between session 1 and session 3 for both orientations, while the difference between session 1 and session 2 only existed for orientation 1. However, these results ought to be interpreted with great caution since session 2 and session 3 were constructed from fewer trials than session 1 and will naturally be subject to larger variance as a result. In addition, session 3 may have been subject to larger *alpha-noise* as was seen in the presence of pre-stimulus alpha in the ERP analyses above. The different results for the spectral and autocorrelation analyses for alpha could be explained by the finding of Herrmann et al [109] that total alpha power in a delay match to sample paradigm

was differentially modulated early compared to late in the retention interval, such that 500ms after stimulus offset task specific differences were detectable. Their result is in line with the finding that activity in the alpha band during the retention period of a working memory task showed significant task dependent changes over posterior and central regions only after 300ms from the onset of a 2.8 second retention interval [87]. Thus, total alpha power computed over the last 800ms of a 1 second delay interval was not sufficient to detect task dependent differences while the autocorrelation analysis comparing $200ms$ intervals was sensitive enough. The results in the spectral and autocorrelation analyses were not entirely different however, since the increased variance in occipital and occipital-parietal electrodes corresponding to lower autocorrelation statistics matched the results of the spectral analyses showing increased alpha power for the same electrode sites. Autocorrelation statistics with respect to a particular frequency band may relate to phase analyses, in that higher correlation values imply phase stability while lower values imply phase shifts.

In that light, the finding that information processing is tied to the phase of oscillations is in line with the idea of *temporal attention* where temporal regularities in the environment contribute to select relevant information [110], possibly via a cortical excitability bias [105]. Specifically, a relationship between alpha oscillations and a temporal selection mechanism have been established. At posterior recording sites alpha oscillations have been associated with a perceptual inhibition timing mechanism [85, 111, 105]. Posterior upper alpha activity was found to prioritize perceptual information in tandem with activity in the gamma frequency band [112]. Further, a thalamo-cortical pathway via the pulvinar [113] was suggested to implement a top-down spatial attention mechanism [114]. And local cortico-cortical or corico-thalamic networks are thought to generate resonant frequencies due to local or specific mono-synaptic neuronal delays [68]. Even though the executive attentional pathway proposed by Jensen et al. [114] is thought to mediate

spatial attention, the pulvinar has been shown to relay information traveling in dorsal as well as ventral pathways [115]. Finally, the frequency of alpha waves is faster at posterior and slower at anterior recording sites [84], which may relate to differential information selection strategies (supported by the finding of Bonnefond [112] mentioned above) or an anterior desynchronization mechanism to ablate interfering information/networks. It is tempting to assume that spatial attention recruits a mechanism similar to feature based attention, however, the latter is apparently more complex.

The analyses of theta power revealed a significant effect between sessions. Specifically, comparisons between frontal, central, and parietal electrode sub-sets showed a consistent difference between session 1 and session 2 for both orientations. While session 1 and session 3 were significantly different in frontal and central electrodes for orientation 1 only, this difference vanished in the posterior subset. These results are consistent with the finding that increased theta power in frontal electrodes and decreased theta power in posterior electrodes tracked mental activity [68]. The mental activity in the anterior part of the brain ought to be more similar in tasks that require visual working memory, such as the categorization task (session 1) and the two interval forced choice task (session 3). In contrast, the object identification task (session 2) that is more similar to an object recognition task would be less likely to tax working memory to the same degree. One clear difference between the tasks is that in the categorization and identification sessions the stimulus was singular, while in the memory comparison session the stimulus represented a pair of objects. Taken together, these results support the notion that category learning and presumably rule representation require working memory.

The presence of theta power implies theta stationarity over the 800ms delay. The idea of frequency dependent perception is thought to include theta frequencies. A meta-analysis over several EEG and magneto-encephalography studies found a temporal parsing mechanism was associated with alpha (peak around 10Hz) and theta (peak around

7Hz) rhythms in the context of visual attentional sampling [86]. Specifically, alpha was thought to relate to sensory information, while theta was thought to reflect attention processes. The implication of theta in the context of executive attention (i.e. a top-down mediated information selection process) matches the finding of increased power at fronto-central electrode sites seen in all sessions. Furthermore, VanRullen [86] also suggested that sensorimotor synchronizations are likely to produce theta band oscillations. Indeed, theta oscillations may provide a temporal gate essential to the formation of multisensory associations in episodic memory. Specifically, small phase modulations (approximately 125ms) between information feeding visual and auditory sensory modalities strongly affected the associative multisensory memory [116]. Taken together, cyclic periodic information sampling mechanisms may pose a mechanistic solution to the perceptual binding problem, and may inspire solutions to rule representation in explicit categorization.

3.5.1 Alternative Accounts

It was noted that alpha frequency correlated with familiar, while theta frequency correlated with novel stimuli [85]. The alternative explanation that stimulus novelty drove the significant effect comparing theta frequencies can not be ruled out entirely, since all participants first participated in session 1 to establish stable rule representations that could be compared with sessions 2 3. In general, differences between orientation bin 1 and orientation bin 7 could be explained by the similarity to a horizontal line of orientation bin 1, which was rotated counter-clockwise $0.0314rad$ from a horizontal line. In contrast, orientation bin 7 was more similar to an oblique line, that was shown to be less accurately remembered than a horizontal line, e.g. *oblique effect* [117].

Alternative explanations for differences in the activity of the alpha frequency include

global changes that could result from caffeine, fatigue, and stress, that could confound the analyses of phasic changes that relate to a particular task when individual alpha frequencies are considered [84]. Furthermore, it was shown that alpha activity was approximately stationary in a resting state, while peak frequency decreased and power increased throughout the course of a mental task [118]. The existence of individual differences contribute a non-negligible confound that complicates straight forward inferences from EEG data that translate into simple models of decision making. In this regard, it is difficult to tease apart the contribution of sensory driven representations that travel up the processing hierarchy in a potentially bayesian way [119] and top-down modifications from executive attention and working memory processes that update internal mappings between sensory and motor regions that may ultimately result in efficient optimized automatic behavior [120]. The isolation of processes and processing components is further complicated by hemispheric lateralization, where the left hemisphere is thought to achieve uncertainty reduction, while the right hemisphere integrates new information [121].

The results presented here are consistent with the categorization literature in that categorization depends on attention and working memory processes. Finally, the small sample size and trial loss paired with the individual difference confound and possible stochasticity during delay make strong inferences about rule representation unfortunately impossible, although these constraints inspire further experimental investigation.

Chapter 4

General Discussion

In the recent past, predictions at the processing level were linked to behavioral measurements in the human psychophysics literature [6]. The successful interface between behavior and mathematical relationships in stimulus space have provoked extensive formalizations of perceptual processes with regard to categorization [9, 12]. The convenient mathematical construct of a decision boundary has precipitated in several cognitive process models, here collectively referred to as *criterion model*. The validity of categorization models can be evaluated either by goodness-of-fit testing or by testing the axioms that are used to build the models (e.g. [30]). Although the categorization literature has been dominated by the former, this thesis focused on the latter and explored the validity of the criterion as a processing component in models of decision making.

In chapter 2, predictions of the criterion model were challenged and falsified the criterion as a processing component in rule-baed categorization. This result represents an important first step away from a classic relic. An alternative direct-mapping model was introduced towards a more parsimonious and intuitive approach of rule representation. Chapter 3 put forth an attempt to characterize the nature of rule representation further using EEG correlates. Even though, it was likely that the first EEG experiment con-

ducted to characterize rule representation would not offer a panacea for abstract rule representation in the human brain, a venture to the realm of local field potentials and neuroelectric waveforms certainly marked an important step. Taken together, the findings presented previously motivate a new neutrally inspired theory of rule representation. In this chapter, the nature of rule representation in the form of direct maps with regard to past models and recent findings across various neuroscience fields is explored.

4.1 Existing Models

In the past, scientists have asked questions that were limited to the algorithmic level [1] and collected data to support or refute assumptions inherent to these models. For example, the serial vs. parallel, exhaustive vs. self-terminating [3], and discrete stage vs. continuous flow [122] type models are difficult to falsify when limiting assumptions are made about the supporting architecture. For instance, Sternberg's [47] memory scanning experiment showed a linear increase in response time with the number of items to be remembered. This evidence motivated Sternberg to propose a serial exhaustive search model that assumes equal processing times for all items. Although this model produced astonishing fits to the memory scanning (or visual search) data, it was not the only model that could account for the linear increase in mean RTs proportional to added memory items. A parallel capacity reallocation model that assumes limited capacity which is equally (re)allocated between uncompleted items mimicked the serial exhaustive model exactly. This example illustrates that goodness-of-fit testing by itself is necessary but limited, such that evaluating the assumptions underpinning the axioms built into both process models becomes a natural pursuit.

Categorization models that utilize a decision bound or criterion are parametric models and require assumptions about the underlying category distributions. For instance, linear

and quadratic decision bounds require multivariate normal category distributions [30]. In contrast, models that highlight associations between perceptual regions and a specific response are equivalent to non-parametric classifiers. It has been suggested that only linear bounds perpendicular to the category relevant dimension are permitted in explicit models. However, in the absence of response bias, the GRT optimal classifier assumes that accuracy is maximized when the participant computes the likelihood ratio of the PDFs of the perceptual effects associated with each category (e.g. $l(x) = g_A(x)\ g_B(x)$). The equivocality contour $h(x)$ is then defined by [19]:

$$h(x) = -log[l(x)] = 0. \tag{4.1}$$

In most cases the equivocality contour will be nonlinear. Therefore, likelihood seems to be an ill-chosen candidate for models of explicit categorization. However, assuming the nonlinear bound can be approximated by the sum of smaller linear bounds that satisfy the requirement of orthogonality to the relevant stimulus/perceptual dimension, the perceptual space can be partitioned into rectangular regions that are deterministically (within the limits of perceptual noise and under a particular set of external task requirements and internal goals) mapped to those response regions that maximize accuracy. Note, even though the likelihood equivocality contour is valid in both, it is not synonymous in stimulus and perceptual space. The RT results presented in Chapter 2 are consistent with a classification strategy that relies on the likelihood ratio. GRT assumes sub-optimal performance is contingent on: perceptual noise, suboptimal decision bounds, variability in the memory of the decision bound, and response bias [19]. Out of these, perceptual and memory noise are explained by neurobiology, while the others are related to individual differences [39, 15].

Explicit rule-learning is thought to be incompatible with pure exemplar theory [123,

124, 125]. Pure exemplar models can be thought of as non-parametric classifiers in which category distributions are estimated using a Parzen kernel [126]. In contrast, explicit representations have been associated with prototype models in which the participant makes strong assumptions about the underlying category distributions to parametrically estimate the decision criterion. Equivalence relations between parametric exemplar models and GRT could only be established at the level of the decision boundary [19]. And a rule-exemplar hybrid model was developed where discrete stimuli are independently represented in addition to the rule [123]. Their model assumes that stimulus presentation frequency affects generalization (i.e. extrapolation in stimulus space). In an experiment where participants had to learn four exceptions embedded in a rule-based category structure the data showed a trend to partition stimulus space according to regions that favored divisions parallel to the categorization bound [127]. This is important because it highlights object or feature representation in a complete model of explicit categorization. Although exemplar theory might be able to account for presentation frequency effects [128], and adding a rule-module that resembles the verbalizable rules in COVIS by dividing the perceptual space according to some single dimensional value might allow to fit accuracy data.

In the current model of category learning, the categorization rule is represented by a criterion based discriminant function [36, 37]. In general, human categorization behavior begins with a hypothesis testing mechanism supported by the explicit system, although control over the response is maintained by this system only if decisional separability is satisfied (i.e. the equal likelihood contour that separates perceptual space is orthogonal to the relevant perceptual axis). The explicit categorization system relies heavily on working memory and executive attention processes that have been formalized in a model of working memory maintenance (FROST; [39]). FROST assumes that working memory maintenance relies on top-down attentional control, which is implemented via a PFC

mediated subcortical pathway that selects what information is maintained in parallel, frontal cortical-thalamic reverberating loops between PFC and association cortex. Object information is maintained via two types of reverberating loops: one between lateral PFC and ventral infero-temporal cortex (ITC; high level visual representations) and the other between lateral PFC and dorsal post-parietal cortex (PPC; spatial representation). Although, FROST highlights the distributed nature of representations during working memory maintenance by postulating a macro-circuit including PPC, thalamus, and the basal ganglia that drives a micro-circuit within lateral PFC, these circuits maintain static contents.

Finally, in a neurobiological model of automaticity, automatic cortico-cortical associations between sensory input and motor output regions are learned slowly via a two factor Hebbian learning rule (SPEED; [120]). The Hebbian connections are formed passively, while either the implicit procedural or explicit rule-based system actively controls the behavior. For the procedural system where behavior is controlled via direct cortico-striatal maps between visual input and motor output Hebbian learning is mediated by the basal ganglia [129]. In contrast, the maps that train the Hebbian connections remain elusive for the explicit system. Importantly, the process supporting explicit control of behavior ought to be flexible enough to account for changes until a definite automatic behavior is optimal. It appears that the uncharted holy grail of cognitive flexibility has prevented a successful merger of COVIS, FROST, and SPEED. In an attempt to uncover the mysterious computations underlying cognitive flexibility, rule-based categorization in particular the integral processes of executive attention and working memory are revisited.

4.2 Executive Attention and Working Memory

Rule-based categorization is supported by an explicit declarative memory system [130, 55]. Declarative representational memories are flexible, accessible to conscious awareness, and supported by medial temporal lobe structures [58, 131]. These memories become available through recollection or retrieval, which is associated with working memory. Working memory has been defined as limited capacity process that temporarily maintains information to support various cognitive functions by interfacing perception, long-term memory, and actions [41, 42, 43, 44, 132]. Even though the exact biochemical properties that support working memory and distinguish short-term memory memory from other encoding stages are not clearly defined [133], pre-automatic retrieval may be thought of as a decision process in itself where a stimulus is mapped to an action goal. Specifically, the process of retrieval and re-consolidation represents a destabilization and re-stabilization of the memory trace and influences long-term modifications [133]. Attention processes selectively modulate retrieval and encoding of representations. In this way working memory and attention work in tandem to select relevant information and mitigate information loss [134].

Attentional templates held in working memory operate on object representations and motor plans to resolve resource competition and guide behavior [107]. In primates two visual processing streams that maintain an attentional bias in working memory have been identified. The dorsal stream maintains a visuospatial bias via a loop between the dlPFC and PPC, and the ventral stream maintains object or feature selective bias via a loop between the vlPFC and ITC [107]. Both streams involve the pulvinar nucleus as thalamocortical relay during top-down control [115]. Clearly, top-down signals are essential during the learning phase to bias recognition toward features of interest, however. Riesenhuber & Poggio [135] assert that object discrimination results from a bottom-up process in

which activation patterns of population codes that represent prototypes are compared. Indeed, there are two opposing attentional forces that are mediated by segregated fronto-parietal networks [136]. The internally driven dorsal fronto-parietal network controls goal directed visuospatial attention, while the right lateralized ventral fronto-parietal network involuntarily responds to unexpected salient stimuli in the external environment [136]. This is consistent with data that found the ventro-lateral PFC associated with implementing previously learned associations, while the dorso-lateral PFC was associated with internally guided rule selection and working memory [137]. Both networks interact via the medial frontal gyrus to select and filter task relevant information [138].

The fronto-parietal network represents complex visual stimuli and task manipulations [139] and supports a perceptual updating mechanism for working memory representations [140]. A meta analysis of differential working memory manipulations revealed consistent activations of brain regions involved in the fronto-parietal network [141]. Finally, there are two sources for attentional top-down control signals that update working memory representations. One is the dorsal fronto-parietal network (intra-parietal sulcus and frontal eye field), and the other is the PFC (ACC and dlPFC) [138]. Taken together, the maps that connect visual object representations to behavioral goals are a mixture between bottom-up stimulus driven representations and top-down feature and response selection mechanisms. How do these interact to support flexible goal directed behaviors?

General representations in PFC result from an interplay between top-down filtering of information and recursive cortico-ganglia networks to generate predictions and action plans [142, 143] that are guided by reward and uncertainty reduction [144]. The PFC receives considerable input from the basal ganglia which supports the idea that these structures collaborate to support behavior via cortico-striatal-thalamic loops [145] associated with category learning [36], working memory [39], and selective attention [146]. Seger [147] proposed four primary interacting [148] cortico-striatal-thalamic loops (visual,

motor, motivational, and executive) that support category learning.

Visual and motor networks show sensitivities to reward prediction and support category learning via loops through posterior caudate and putamen, respectively [120, 37]. The visual loop relays input information to the executive and motor loops for response selection, in return these influence the visual loop to refine the selection of the relevant stimulus component. Recent evidence supports a bi-directional influence between striatum and PFC, however, striatal LFPs were found to exert a greater influence on PFC LFPs than vice versa [149]. While this result is consistent with the hypothesis that the ganglia train representations in PFC [142, 143], no changes in the directional influence between striatum and PFC was observed as a result of learning [149]. Instead these results are consistent with the idea that stimulus driven representations are relayed to PFC, who in turn refines the representations to emphasize task relevant information. The reverberating loop actively relays object representations such that new information about changing environmental demands or internal goals can be integrated. Lastly, the motor loop was found to provide response information to the executive loop such that the success of the response can be integrated with future action goals, and both motor and executive loops were found to interact with the motivational loop, possibly providing information regarding reward history [150, 148].

The motivational and executive loops support feedback processing and intersect at the anterior caudate, which is largely innervated by Dopamine [147]. Feedback from the environment is integrated to update internal goals via ventro-medial PFC (anterior-cingulate and orbito-frontal cortices) and the anterior caudate. The executive loop integrates reward information with action goals via dlPFC and PPC through subcortical relays that meet in the anterior caudate, and updates working memory representations throughout category learning. Interestingly, the motivational loop was found to correlate with prediction errors in general, while the executive loops was specific to reward prediction error

[150]. In addition to the basal ganglia, PFC also forms a loop with the medial temporal lobe, which is thought to store and provide stimulus specific category representations [27], which can be uploaded into working memory [133]. Taken together, category learning is supported by a complex network of subsequently evolved components that process information via dynamically interacting cortico-ganglia loops. But where exactly are categorization rules represented?

The PFC is a powerful integrator of information and plays a central role in the coordination of goal directed behaviors [151, 152]. The PFC is heavily interconnected with higher associations areas rather than with primary sensory or motor cortices and acts as global attention controller by relaying task information to posterior brain systems [142]. The PFC is thought to maintain information about the overall structure of the task and track performance [153, 143] via a wide variety of cellular activity profiles [154]. The PFC is a top candidate for reward processing since it receives the highest density of dopaminergic innervation of with respect to the rest of frontal cortex [155].

Lateral PFC appears to be a critical actor in rule learning, patients with damage to the lateral PFC are incapable to switch behavioral control to more appropriate associations and have to rely on automatic associations [156]. Specifically, medial frontal gyrus is thought to store maps connecting visual stimuli and motor goals while the lateral PFC oversees and mediates behavioral performance. The medial frontal gyrus is associated with a hierarchical processing gradient where posterior regions associate motor function, central regions correlate with cognitive control and pain, while the anterior region integrates reward and episodic memory [157].

An adaptive coding model was introduced where PFC neurons are suggested to multitask (e.g. involved in working memory, selective attention, and cognitive control; [158]). Given that PFC neurons appear to multitask, deterministic neurons by themselves are most likely insufficient to generate flexible rule representation. After reviewing some

evidence that executive attention and working memory operate via complex networks centered around the PFC, the prefrontal representations are explored in more detail to characterize the *rule* maps that support behavior [151].

4.2.1 Dynamic Population Codes

The evidence presented in the following shows that rule representation at the neurobiological level is a conglomeration of electro-chemical signals that is integrated over several neurons in PFC. Although the emphasis of category relevant visual features of complex objects occurs in ITC [159, 160], the process of pooling stimulus features into an explicit category associated with a particular behavioral response relies on PFC [161]. Neurons in PFC increase category selective activity throughout a memory delay while the number of neurons in ITC is pruned [162]. It is noteworthy that half of the PFC neurons encoded category specific response selection, while a smaller number of PFC neurons encoded actions independent of the visual category. Furthermore, representations in PFC are strongly modulated by behavioral goals and mediate a temporary enhancement of task related feature representations in ITC [160], suggesting that increased category selectivity in ITC was at least in part contingent on PFC mediated top-down information biasing [163, 164, 162], that is consistent with the aforementioned cottico-ganglia feedback selection mechanism [150, 148]. In addition, parietal cortex was found to be sensitive to categorical representations [165, 166], that represent another source of behaviorally relevant input to complete our world view with location specific information [167, 168]. Consistent with the aforementioned fronto-parietal control of visuospatial attention and cortico-striatal integration mechanisms, stimulus specific representations fed forward from PPC in tandem with executive feed-back biasing signals from PFC prepared the cortical environment to process expected incoming stimuli in a robust goal-directed

manner [167]. Interestingly, stimulus identity was encoded in PPC although to a lesser degree than location specificity, likely due to large interconnectivity at every stage along the processing hierarchy between ventral and dorsal streams to support robust representations. These findings support the idea that several maps between PFC and association cortices support robust sensory integration with behavior.

How do these maps support flexible cognition? Recordings from the frontal eye field (an area of PFC involved in coordinating saccadic eye movements, visuo-spatial attention, and visuo-motor decisions) of primates showed that population representations were sufficient to separate dynamically evolving sensory representations while single cell activity was not [169]. The task required flexible selection and integration of superimposed color and motion information towards a saccade. Specifically, population trajectories (where each point in the trajectory represented a unique pattern of activity across the population) in neural state space (constructed from principal component analysis and linear regression to identify four task relevant orthogonal axes) followed a dynamic path towards a choice that maintained inter- and intra-dimensional separability of motion or color information regardless which context was relevant. Importantly, entirely different parts of state space were occupied depending whether motion or color was relevant.

In another experiment where two orthogonal category schemes were mapped onto the same stimuli, most neurons (30%) in lateral PFC were selective for one category scheme, while some neurons (7%) were selective for both category schemes [170]. This finding is expected when the same perceptual regions are mapped to both categories. Furthermore, selectivity for stimuli close to the boundary was enhanced regardless whether the preferred scheme was relevant, and category selective populations followed a dynamic generalization from stimulus to category representation during the delay when the preferred category scheme was relevant; presumably to resolve competition between maps and support robust behavioral performance. Importantly, Roy et al. [170] observed a

reduction in selectivity when the encoded category scheme was irrelevant that did not result from a suppression of overall activity, rather the decreased selectivity may have been contingent on pre-synaptic inputs from other parts of PFC (e.g. frontopolar cortex [156]).

Finally, Meyers et al. [171] used a decoding approach to reveal dynamic population codes of category relevant information in PFC neurons of primates performing a delayed match-to-category task. While a small subset of eight neurons sufficiently encoded category information at the time of the response, a non-negligible amount of redundant information was contained in a sparse code of approximately 64 PFC neurons at any moment throughout the task. These studies support that PFC actively selects and integrates relevant information using dynamic populations codes, rather than passively biasing sensory information via pure a priori top-down filtering approach that is implemented by the same few neurons. Consistent with the complex network of reverberating loops, these findings provoke the hypothesis that selection and integration processes exist on a dynamic continuum that begins with recruiting a sufficient sample of neurons to reliably represents the sensory evidence that is subsequently pruned to select a robust subsample which is subsequently connected with a response. The idea that PFC constantly integrates sensory information via recurrent loops such that representations evolve to converge onto a reliable subset of sensory input and response neurons is consistent with a selection process that supports the evolution of cortico-cortical Hebbian synapses underlying automatic behavior [120, 172].

4.3 Time Consoles

PFC serves two fundamental purposes: the selection and maintenance of behaviorally relevant information, and the integration of information across remote brain regions that

encode relevant information. In continuance with the idea that selection and integration processes exist on a continuum, these complex functions are supported by a dynamically evolving heterogeneous network of neurons. How these complex population codes could be implemented in a parsimonious model may become more apparent when they are viewed with regard to the temporal structure that supports them.

Different remote brain areas are though to communicate through coherence where effective connectivity occurs during windows of rhythmic synchronization that temporally focuses neuronal output and sensitivity to inputs [173]. This communication is constrained by non-negligible temporal signaling delays that differentially affect feedforward and feedback transmission depending on the frequency band [174]. Attention processes have been associated with activity in several frequency bands that support a controlled top-down temporal selection mechanism to selectively gate relevant information [175, 85, 111, 105, 86, 112]. For instance, top-down prioritization and bottom-up stimulus salience were found to correlate with beta/alpha and gamma band coherence, respectively [112, 174]. Cortical theta phase was found to modulate the strength of gamma coherence, thereby offering a frequency mediated attentional sampling mechanism [86]. This mechanism is related to capacity, since it has been shown that an attentional sampling process around 7-8 Hz is divided over one to three objects [174]. Similarly, working memory processes that maintain attention selected information appear to utilize oscillations to disentangle information [176, 177, 178, 179]. The continuous information selection supported by the ongoing exchange between brain areas provides a mechanism to counterbalance noise [75].

Category learning and performance were associated with increased phase locking between LFPs in striatum and PFC during a delay prior to the response during correct trials and a decrease in synchrony during error trials, while no changes in power were observed [149]. In another experiment, coherence between PPC spikes and PFC LFPs

during a delay period revealed that both stimulus selective and non-selective parietal neurons were synchronized with pre-frontal LFPs, and that several PPC-PFC pairs showed selectivity for multiple locations and/or objects [167]. These findings are consistent with a pre-frontal temporal attention gating mechanism. In addition, it has been suggested that a functional role between LFPs and spike timing of individual neurons exists [100]. For example, in visual cortex perceptual grouping is achieved via synchronized initial spikes of cells that code for related features and that this temporal precision increases with viewing frequency [180]. The idea that spike timing adjustment is a fundamental cortical mechanism that supports a coherent world view has been shown across a variety of oscillation frequencies [105, 112, 86, 116].

Oscillations allow for activity dependent coupling of sparsely connected ensembles that support rapid stimulus-response associations of novel stimuli with goal driven responses. Thereby offering a mechanism to support perceptual binding as well as cognitive flexibility. The empirically supported hypothesis that PFC dynamically integrates sensory information with behavioral context via synchronous cell assemblies could be implemented using oscillations.

4.4 Direct Mapping Model Revisited

At the coarsest level the human brain can be characterized by a gradient from depictive (new sensory input) to propositional (internal model) representations that are weighted differentially across hemispheres [121]. The decision making process underlying categorization is highly complex and involves many sub-processes that include selecting relevant while suppressing irrelevant sensory information, making a category judgement according to the current rule, and eliciting a response.

Goal directed stimulus representations result from a combination of activity in pre-

existing neural circuits that respond to stimulus salience, and biasing activity from the dorsal fronto-parietal network that integrates visual information with behavioral goals (possibly by predicting the location of the target), as well as other parts of the PFC that oversee successful performance and update response goals [136, 138, 150, 148]. Perceptual representations are temporarily modulated by attention and can bootstrap permanent changes. For instance, perceptual separability of novel category dimensions can be trained [181]. The attentional networks that select stimulus driven and goal directed representations dynamically co-evolve task relevant perceptual representations.

Multiple potential maps between stimulus specific (identity and location) information and category labels or action goals are modulated by context [160], and are separable only at the population level in PFC [170, 169]. The observed population dynamics in PFC are clearly contingent on the task: competitive associations between the overlapping regions in perceptual space are resolved by dynamic recruitment of the entire category map [170], superimposed sensory inputs are represented by unique neural codes [169], and deterministic stimulus-category associations converge onto a sufficient sample of neurons [171]. The emergence of one map as dominant controller of behavior depends on dynamic processes that recursively integrate stimulus inputs with reward and action goals [147, 150, 148] until robust representations that support behavior and inform future scenarios are generated. It is therefore not the resulting map, but the previous complex integration process via subcortical loops and the precise orchestration of feedforward and feedback loops [167, 149, 174] that marks the holy grail of cognitive flexibility. As a result, the nature of rule representation is not static but rather a dynamic process that is mediated by a complex cortico-ganglia network of reverberating feedforward and feedback connections. Under this framework, the suggestion that differences between ITC and PFC during visual categorization in primates are *a matter of degree rather than a strict division of labor* [162], likely refers to one snapshot of the dynamic process that is taken at

the culmination point just before the response. In this way, dynamically refined category relevant (stimulus-response) associations eventually converge onto a deterministic map, which supports behavior and trains automatic associations.

In the direct mapping model explicit rules are mediated by prefrontal cortical population codes that select relevant information via executive attention that operates on higher visual representations and connects these to context dependent behavioral goals in a transient electrical working memory space. Accordingly, the current model of working memory maintenance, FROST, is revisited with regard to a rule representation module. FROST highlights the distributed nature of representations during working memory maintenance by postulating a macro-circuit including PPC, thalamus, and the basal ganglia that drives a micro-circuit within lateral PFC. FROST relies on cortical interneurons to connect the cortical layers in lateral PFC. These interneurons interface the cortico-thalamic reverberating loop with the information carrying loops to association cortices (e.g. PPC and ITC), and allow for additional regulating inputs from other frontal regions.

Laminar differences provide the framework for microcircuits that integrate feedforward and feedback information within PFC and between PFC and subcortical as well as other cortical areas. Microcircuits in PFC are organized into a laminar structure of five layers, where layer 2/3 support horizontal communication between neocortical areas as well as local processing in layers 4 to 6 within PFC microcircuits [182]. Furthermore, cortico-thalamic macrocircuits are driven by deeper layers that connect parallel cortico-thalamic loops in addition to the connections in layer 2/3 [183]. This laminar organization offers a doorway to integrate afferent pre-synaptic bias from many sources. The electrical stimulus-response maps formed in lateral PFC may be flexibly modulated by pre-synaptic biasing signals from medial PFC [156, 157]. In particular, orbitofrontal cortex mediates stimulus-reward associations (current stimulus value) while anterior cingulate cortex is a likely source for a pre-synaptic bias of response selection [147, 137].

These areas are heavily interconnected with lateral PFC [184].

The microcircuits in PFC that drive the cortico-thalamic macrocircuits result from a complex network between excitatory pyramidal cells and inhibitory GABAergic interneurons. Furthermore, neocortical interneurons show large morphological diversity that occur in differential proportions throughout the cortical laminar layers and show divergent spiking behaviors [185], which might complicate micro-to-macro-circuit maps. However, temporal gating offers a signaling mechanisms via local field potentials that may provide implementational solutions beyond the electrochemical level of individual neurons. In particular, superficial cortical layers that receive anatomical forward projections preferentially show gamma band activity, while deeper cortical layers that are the primary source of feedback projections show strong coherence in the alpha/beta band [186]. These findings inspire categorization models at the level of LFPs that could potentially generate predictions that are testable at the level of human EEG. The challenge to implement these models is beyond the scope of this thesis and offers a largely uncharted territory for neurocomputational models of rule-based category learning.

4.5 To Consider

Creating models of stochastic neurobiological processes is at the very least challenging. In the following a final note on the topic of *noise* is presented.

4.5.1 Is all Noise Equal?

An important distinction between biological phenomena is whether they were generated by a stochastic process of a deterministic process with chaotic dynamics. Assumptions about the mechanics of the underlying biological process determine how to analyze data describing the process and how a model ought to be constrained, and should be

made prior to either endeavor. For instance, stochastic and deterministic systems are indistinguishable at the level of the Fourier generated power spectrum [187]. In the case of neurobiological data, one group concluded that EEG time series, 1 second long, were not generated by deterministic processes [188]. Often, the stochastic nature of processes in the brain is modeled by adding Gaussian white noise [189, 190, 33].

In general, noise has had a reputation for corrupting the relationship between input and output of a system or for corrupting measurement accuracy. In the field of control engineering, models with superimposed noise assume that noise and input to a system are independent and that this corrupting noise is defined to be *white*, i.e. Gaussian with $\mu = 0$ and some variance σ^2. Similarly, measurement noise in the filed of EEG has been characterized by white noise, given that the brain filtered signals are defined as colored noise [97]. Assuming the brain is a closed electro-chemical system that is not affected by external electric, magnetic, electro-magnetic, or chemical disturbances other than the sensory interpretations of these, the sensory inout signals may be viewed as pure inputs. And the task is to extract the wanted noise from the unwanted background noise, it then appears over-simplified that a noise variable in an implementational model shall be independent with a Gaussian functional form. Furthermore, the brain follows computational rules that show highly optimized tolerance and robustness to expected environments by allocating processing resources according to specific design principles [191], , for instance in primary visual cortex area V1 approximately 25% of cortex is devoted to processing the central 2.5° of visual angle [192]. It may be reasonable to assume that most of these design principles are either undiscovered or have not yet become widely known. Since the relationship between sensory input and robust behavioral performance is not entirely clear, it is even less clear what functional role is associated with white noise. So what then is the purpose of noise in the brain and how should we model it?

Traditionally deterministic models of decision processes have been awarded stochastic

character by adding Gaussian white noise to linear systems equations [189]. The brain, however, is a dynamic system where processes are inherently stochastic because they are unpredictable with our current understanding. Therefore, what we view as *white noise* from the perspective of determinism is likely related to the complex feedforward and feedback networks that exist on several coupled levels of granularity, which are fundamental to the phenomenon of cognitive flexibility. The solution of adding gaussian noise to make deterministic systems look more like natural stochastic systems may at best serve some cosmetic purpose but seems entirely too simplistic to capture the structural nuances that are contained in the marginal or noisy contributions of information (e.g. [171, 170, 169]). In general, noise changes signal attributes such as peak frequency or tuning curves [193], and it has been shown that correlated noise was sufficient to decode uncertainty [194]. The brain is a noisy apparatus that has time and space to constrain and separate particular operations. It could be a worthwhile venture to explore the brain's filter functions by comparing system outputs between deterministic inputs and those where structured noise has been added.

4.6 In Closing

One essential element of human categorization behavior is generalization, or the ability to apply knowledge from past experience to novel situations. For instance, multiplication can be learned through serial addition. On the most superficial level, categorization behavior can be divided into three components: perception, decision making, and response generation. The nature of rule representation was explored with regard to attention and working memory processes and how these modulate task relevant stimulus response associations. This thesis contributes evidence against the popular criterion as a processing component in rule-based decision making and proposed a neutrally in-

spired direct-mapping model. In this model, the rules that support explicit decisions are described by dynamic pre-frontal micro-to-macro circuits that flexibly integrate current reward goals into this complex circuitry via pre-synaptic modulations.

Future work is required to implement these circuits. In the light that rule based categorization depends on a complex cortico-ganglia network that dynamically integrates sensory information with reward and behavioral goals to support robust responses that maximize accuracy, it might useful to begin with models based on functional connectivity. Furthermore, dynamical system approaches might offer promising solutions for population LFP models that may subsequently find relevance to new brain computer interface applications.

Bibliography

[1] D. Marr, *Vision: A computational approach*. Freeman and Co San Francisco, 1982.

[2] R. Ratcliff, *A theory of memory retrieval.*, *Psychological review* **85** (1978), no. 2 59.

[3] J. T. Townsend and F. G. Ashby, *Stochastic modeling of elementary psychological processes*. CUP Archive, 1983.

[4] D. R. King and M. B. Miller, *Influence of response bias and internal/external source on lateral posterior parietal successful retrieval activity*, *cortex* **91** (2017) 126–141.

[5] R. M. Roe, J. R. Busemeyer, and J. T. Townsend, *Multialternative decision field theory: A dynamic connectionst model of decision making.*, *Psychological review* **108** (2001), no. 2 370.

[6] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*, .

[7] H. Stanislaw and N. Todorov, *Calculation of signal detection theory measures*, *Behavior research methods, instruments, & computers* **31** (1999), no. 1 137–149.

[8] Y. Jang, J. T. Wixted, and D. E. Huber, *Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory.*, *Journal of Experimental Psychology: General* **138** (2009), no. 2 291.

[9] F. G. Ashby and J. T. Townsend, *Varieties of perceptual independence.*, *Psychological review* **93** (1986), no. 2 154.

[10] F. G. Ashby and W. W. Lee, *Perceptual variability as a fundamental axiom of perceptual science*, *Advances in Psychology* **99** (1993) 369–399.

[11] H. Kadlec and J. T. Townsend, *Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions*, *Journal of Mathematical Psychology* **36** (1992), no. 3 325–374.

[12] F. G. Ashby and F. A. Soto, *Multidimensional signal detection theory*, *Oxford handbook of computational and mathematical psychology* (2015) 13–34.

[13] F. G. Ashby and N. A. Perrin, *Toward a unified theory of similarity and recognition.*, *Psychological review* **95** (1988), no. 1 124.

[14] M. J. Wenger and E. M. Ingvalson, *Preserving informational separability and violating decisional separability in facial perception and recognition.*, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29** (2003), no. 6 1106.

[15] F. A. Soto, L. Vucovich, R. Musgrave, and F. G. Ashby, *General recognition theory with individual differences: a new method for examining perceptual and decisional interactions with an application to face perception*, *Psychonomic bulletin & review* **22** (2015), no. 1 88–111.

[16] F. G. Ashby and W. T. Maddox, *A response time theory of separability and integrality in speeded classification*, *Journal of Mathematical Psychology* **38** (1994), no. 4 423–466.

[17] F. G. Ashby and R. E. Gott, *Decision rules in the perception and categorization of multidimensional stimuli*, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14** (1988) 33–53.

[18] F. G. Ashby and W. T. Maddox, *Complex decision rules in categorization: Contrasting novice and experienced performance.*, *Journal of Experimental Psychology: Human Perception and Performance* **18** (1992), no. 1 50.

[19] F. G. Ashby and W. T. Maddox, *Relations between prototype, exemplar, and decision bound models of categorization*, *Journal of Mathematical Psychology* **37** (1993), no. 3 372–400.

[20] W. T. Maddox and C. J. Bohil, *Base-rate and payoff effects in multidimensional perceptual categorization.*, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24** (1998), no. 6 1459.

[21] J. Balakrishnan, *Decision processes in discrimination: fundamental misrepresentations of signal detection theory.*, *Journal of Experimental Psychology: Human Perception and Performance* **25** (1999), no. 5 1189.

[22] S. S. Stevens, *To honor fechner and repeal his law*, *Science* **133** (1961), no. 3446 80–86.

[23] J. D. Wallis, K. C. Anderson, and E. K. Miller, *Single neurons in prefrontal cortex encode abstract rules*, *Nature* **411** (2001), no. 6840 953–956.

[24] L. Veit and A. Nieder, *Abstract rule neurons in the endbrain support intelligent behaviour in corvid songbirds*, *Nature communications* **4** (2013) 2878.

[25] F. G. Ashby, *A stochastic version of general recognition theory*, *Journal of Mathematical Psychology* **44** (2000), no. 2 310–329.

[26] F. G. Ashby and V. V. Valentin, *The categorization experiment: Experimental design and data analysis*, *The Stevens? Handbook of Experimental Psychology and Cognitive Neuroscience* (2017).

[27] C. A. Seger and E. K. Miller, *Category learning in the brain*, *Annual Review of Neuroscience* **33** (2010) 203–219.

[28] F. G. Ashby, V. V. Valentin, and S. S. von Meer, *Differential effects of dopamine-directed treatments on cognition*, *Neuropsychiatric disease and treatment* **11** (2015) 1859.

[29] J. D. Smith, A. C. Zakrzewski, J. J. Johnston, J. L. Roeder, J. Boomer, F. G. Ashby, and B. A. Church, *Generalization of category knowledge and dimensional categorization in humans (homo sapiens) and nonhuman primates (macaca mulatta).*, *Journal of Experimental Psychology: Animal Learning and Cognition* **41** (2015), no. 4 322.

[30] F. G. Ashby and E. M. Waldron, *On the nature of implicit categorization*, *Psychonomic Bulletin & Review* **6** (1999), no. 3 363–378.

[31] F. G. Ashby and M. J. Crossley, *A computational model of how cholinergic interneurons protect striatal-dependent learning*, *Journal of Cognitive Neuroscience* **23** (2011), no. 6 1549–1566.

[32] V. V. Valentin, W. T. Maddox, and F. G. Ashby, *A computational model of the temporal dynamics of plasticity in procedural learning: Sensitivity to feedback timing*, *Frontiers in Psychology* **5** (2014), no. 643.

[33] G. Cantwell, M. J. Crossley, and F. G. Ashby, *Multiple stages of learning in perceptual categorization: evidence and neurocomputational theory*, *Psychonomic bulletin & review* **22** (2015), no. 6 1598–1613.

[34] F. G. Ashby, S. W. Ell, and E. M. Waldron, *Procedural learning in perceptual categorization*, *Memory & Cognition* **31** (2003), no. 7 1114–1125.

[35] M. B. Casale, J. L. Roeder, and F. G. Ashby, *Analogical transfer in perceptual categorization*, *Memory & Cognition* **40** (2012), no. 3 434–449.

[36] F. G. Ashby, L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldron, *A neuropsychological theory of multiple systems in category learning.*, *Psychological Review* **105** (1998), no. 3 442–481.

[37] F. G. Ashby, E. J. Paul, and W. T. Maddox, *COVIS*, in *Formal approaches in categorization* (E. M. Pothos and A. Wills, eds.), pp. 65–87. Cambridge University Press, New York, 2011.

[38] K. R. Ridderinkhof, W. P. Van Den Wildenberg, S. J. Segalowitz, and C. S. Carter, *Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning*, *Brain and cognition* **56** (2004), no. 2 129–140.

[39] F. G. Ashby, S. W. Ell, V. V. Valentin, and M. B. Casale, *Frost: a distributed neurocomputational model of working memory maintenance*, *Journal of cognitive neuroscience* **17** (2005), no. 11 1728–1743.

[40] F. G. Ashby and V. V. Valentin, *Multiple systems of perceptual category learning: Theory and cognitive tests*, in *Handbook of Categorization in Cognitive Science (Second Edition)*, pp. 157–188. Elsevier, 2017.

[41] S. J. Luck and E. K. Vogel, *The capacity of visual working memory for features and conjunctions*, *Nature* **390** (1997), no. 6657 279.

[42] W. T. Maddox, F. G. Ashby, A. D. Ing, and A. D. Pickering, *Disrupting feedback processing interferes with rule-based but not information-integration category learning*, *Memory & Cognition* **32** (2004), no. 4 582–591.

[43] D. Zeithamova and W. T. Maddox, *Dual-task interference in perceptual category learning*, *Memory & Cognition* **34** (2006), no. 2 387–398.

[44] P. M. Bays and M. Husain, *Dynamic shifts of limited working memory resources in human vision*, *Science* **321** (2008), no. 5890 851–854.

[45] S. W. Ell and F. G. Ashby, *The effects of category overlap on information-integration and rule-based category learning*, *Perception & Psychophysics* **68** (2006), no. 6 1013–1026.

[46] F. G. Ashby and W. W. Lee, *Predicting similarity and categorization from identification.*, *Journal of Experimental Psychology: General* **120** (1991), no. 2 150.

[47] S. Sternberg, *High-speed scanning in human memory*, *Science* **153** (1966), no. 3736 652–654.

[48] D. H. Brainard, *The psychophysics toolbox*, *Spatial Vision* **10** (1997) 433–436.

[49] T. Van Zandt, *Analysis of response time distributions*, *Stevens? handbook of experimental psychology* **4** (2002) 461–516.

[50] R. Ratcliff, *Group reaction time distributions and an analysis of distribution statistics.*, *Psychological bulletin* **86** (1979), no. 3 446.

[51] E. Parzen, *On estimation of a probability density function and mode*, *The annals of mathematical statistics* **33** (1962), no. 3 1065–1076.

[52] J. D. Smith and S. W. Ell, *One giant leap for categorizers: One small step for categorization theory*, *PloS one* **10** (2015), no. 9 e0137334.

[53] J. T. Townsend, *Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference.*, *Psychological Bulletin* **108** (1990), no. 3 551.

[54] B. Gordon, *Preserved learning of novel information in amnesia: Evidence for multiple memory systems*, *Brain and Cognition* **7** (1988), no. 3 257–282.

[55] E. Nomura, W. Maddox, J. Filoteo, A. Ing, D. Gitelman, T. Parrish, M. Mesulam, and P. Reber, *Neural correlates of rule-based and information-integration visual category learning*, *Cerebral Cortex* **17** (2007), no. 1 37–43.

[56] D. L. Schacter, *Priming and multiple memory systems: Perceptual mechanisms of implicit memory*, *Journal of Cognitive Neuroscience* **4** (1992), no. 3 244–256.

[57] R. A. Poldrack and M. G. Packard, *Competition among multiple memory systems: converging evidence from animal and human brain studies*, *Neuropsychologia* **41** (2003), no. 3 245–251.

[58] L. R. Squire, *Memory systems of the brain: A brief history and current perspective*, *Neurobiology of Learning and Memory* **82** (2004), no. 3 171–177.

[59] W. T. Maddox and A. D. Ing, *Delayed feedback disrupts the procedural-learning system but not the hypothesis testing system in perceptual category learning*, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31** (2005), no. 1 100–107.

[60] F. G. Ashby and J. M. Ennis, *The role of the basal ganglia in category learning*, *Psychology of Learning and Motivation* **46** (2006) 1–36.

[61] L. A. Alfonso-Reese, *Dynamics of category learning*, Unpublished doctoral dissertation, University of California, Santa Barbara, 1996.

[62] S. Helie, S. W. Ell, J. V. Filoteo, and W. T. Maddox, *Criterion learning in rule-based categorization: Simulation of neural mechanism and new data*, *Brain and cognition* **95** (2015) 19–34.

[63] E. J. Paul and F. G. Ashby, *A neurocomputational theory of how explicit learning bootstraps early procedural learning*, *Frontiers in computational neuroscience* **7** (2013) 177.

[64] E. Rosch, C. Simpson, and R. S. Miller, *Structural bases of typicality effects.*, *Journal of Experimental Psychology: Human perception and performance* **2** (1976), no. 4 491.

[65] M. Riesenhuber and T. Poggio, *Hierarchical models of object recognition in cortex*, *Nature neuroscience* **2** (1999), no. 11 1019.

[66] J. O. Garcia, R. Srinivasan, and J. T. Serences, *Near-real-time feature-selective modulations in human cortex*, *Current Biology* **23** (2013), no. 6 515–522.

[67] T. C. Sprague and J. T. Serences, *Using human neuroimaging to examine top-down modulation of visual perception*, in *An introduction to model-based cognitive neuroscience*, pp. 245–274. Springer, 2015.

[68] P. L. Nunez, R. Srinivasan, *et. al.*, *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

[69] J. J. Foster, D. W. Sutterer, J. T. Serences, E. K. Vogel, and E. Awh, *The topography of alpha-band activity tracks the content of spatial working memory*, *Journal of neurophysiology* **115** (2015), no. 1 168–177.

[70] J. Samaha, T. C. Sprague, and B. R. Postle, *Decoding and reconstructing the focus of spatial attention from the topography of alpha-band oscillations*, *Journal of cognitive neuroscience* **28** (2016), no. 8 1090–1097.

[71] E. K. Vogel, G. F. Woodman, and S. J. Luck, *Storage of features, conjunctions, and objects in visual working memory.*, *Journal of Experimental Psychology: Human Perception and Performance* **27** (2001), no. 1 92.

[72] E. Awh, B. Barton, and E. K. Vogel, *Visual working memory represents a fixed number of items regardless of complexity*, *Psychological science* **18** (2007), no. 7 622–628.

[73] W. Zhang and S. J. Luck, *Discrete fixed-resolution representations in visual working memory*, *Nature* **453** (2008), no. 7192 233.

[74] P. Wilken and W. J. Ma, *A detection theory account of change detection*, *Journal of vision* **4** (2004), no. 12 11–11.

[75] P. M. Bays, *Noise in neural populations accounts for errors in working memory*, *Journal of Neuroscience* **34** (2014), no. 10 3632–3645.

[76] P. Sauseng, W. Klimesch, K. F. Heise, W. R. Gruber, E. Holz, A. A. Karim, M. Glennon, C. Gerloff, N. Birbaumer, and F. C. Hummel, *Brain oscillatory substrates of visual short-term memory capacity*, *Current biology* **19** (2009), no. 21 1846–1852.

[77] P. M. Bays, N. Gorgoraptis, N. Wee, L. Marshall, and M. Husain, *Temporal dynamics of encoding, storage, and reallocation of visual working memory*, *Journal of vision* **11** (2011), no. 10 6–6.

[78] R. G. Morrison, P. J. Reber, K. L. Bharani, and K. A. Paller, *Dissociation of category-learning systems via brain potentials*, *Frontiers in human neuroscience* **9** (2015) 389.

[79] R. Rabi, M. F. Joanisse, T. Zhu, and J. P. Minda, *Cognitive changes in conjunctive rule-based category learning: An erp approach*, *Cognitive, Affective, & Behavioral Neuroscience* **18** (2018), no. 5 1034–1048.

[80] B. Treutwein, I. Rentschler, and T. Caelli, *Perceptual spatial frequency—orientation surface: psychophysics and line element theory*, *Biological Cybernetics* **60** (1989), no. 4 285–295.

[81] H. Berger, *Über das elektrenkephalogramm des menschen*, *European archives of psychiatry and clinical neuroscience* **87** (1929), no. 1 527–570.

[82] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[83] H. v. Helmholtz, *Ueber einige gesetze der vertheilung elektrischer ströme in körperlichen leitern, mit anwendung auf die thierisch-elektrischen versuche (schluss.)*, *Annalen der Physik* **165** (1853), no. 7 353–377.

[84] W. Klimesch, *Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis*, *Brain research reviews* **29** (1999), no. 2-3 169–195.

[85] W. Klimesch, *Alpha-band oscillations, attention, and controlled access to stored information*, *Trends in cognitive sciences* **16** (2012), no. 12 606–617.

[86] R. VanRullen, *Perceptual cycles*, *Trends in cognitive sciences* **20** (2016), no. 10 723–735.

[87] O. Jensen, J. Gelfand, J. Kounios, and J. E. Lisman, *Oscillations in the alpha band (9–12 hz) increase with memory load during retention in a short-term memory task*, *Cerebral cortex* **12** (2002), no. 8 877–882.

[88] R. Scheeringa, P. J. Koopmans, T. van Mourik, O. Jensen, and D. G. Norris, *The relationship between oscillatory eeg activity and the laminar-specific bold signal*, *Proceedings of the National Academy of Sciences* **113** (2016), no. 24 6761–6766.

[89] M. J. Wolff, J. Jochim, E. G. Akyürek, and M. G. Stokes, *Dynamic hidden states underlying working-memory-guided behavior*, *Nature Neuroscience* **20** (2017), no. 6 864.

[90] C. S. Herrmann, M. Grigutsch, and N. A. Busch, *Eeg oscillations and wavelet analysis*, *Event-related potentials: A methods handbook* (2005) 229.

[91] M. M. Doppelmayr, W. Klimesch, T. Pachinger, and B. Ripper, *The functional significance of absolute power with respect to event-related desynchronization*, *Brain topography* **11** (1998), no. 2 133–140.

[92] M. R. Nuwer, M. Aminoff, J. Desmedt, A. A. Eisen, D. Goodin, S. Matsuoka, F. Mauguière, H. Shibasaki, W. Sutherling, and J.-F. Vibert, *Ifcn recommended standards for short latency somatosensory evoked potentials. report of an ifcn committee*, *Electroencephalography and clinical Neurophysiology* **91** (1994), no. 1 6–11.

[93] A. Delorme and S. Makeig, *Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis*, *Journal of neuroscience methods* **134** (2004), no. 1 9–21.

[94] S. Muthukumaraswamy, *High-frequency brain activity and muscle artifacts in meg/eeg: a review and recommendations*, *Frontiers in human neuroscience* **7** (2013) 138.

[95] G. F. Woodman, *A brief introduction to the use of event-related potentials in studies of perception and attention*, *Attention, Perception, & Psychophysics* **72** (2010), no. 8 2031–2046.

[96] S. J. Luck and N. Gaspelin, *How to get statistically significant effects in any erp experiment (and why you shouldn't)*, *Psychophysiology* **54** (2017), no. 1 146–157.

[97] G. Dumermuth and L. Molinari, *Spectral analysis of the eeg*, *Neuropsychobiology* **17** (1987), no. 1-2 85–99.

[98] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time signal processing. Vol. 2.* Upper Saddle River, NJ: Prentice Hall, 2001.

[99] J. Campbell, E. Bower, S. Dwyer, and G. Lago, *On the sufficiency of autocorrelation functions as eeg descriptors*, *IEEE Transactions on Biomedical Engineering* (1967), no. 1 49–52.

[100] K. Benchenane, P. H. Tiesinga, and F. P. Battaglia, *Oscillations in the prefrontal cortex: a gateway to memory and attention*, *Current opinion in neurobiology* **21** (2011), no. 3 475–485.

[101] J. W. Cooley and J. W. Tukey, *An algorithm for the machine calculation of complex fourier series*, *Mathematics of computation* **19** (1965), no. 90 297–301.

[102] G. E. Box and G. M. Jenkins, *Time Series Analisis: Forescasting and Control*. Holden Day, 1970.

[103] J. L. Semmlow, *Biosignal and medical image processing*. CRC press, 2008.

[104] E. K. Vogel and S. J. Luck, *The visual n1 component as an index of a discrimination process*, *Psychophysiology* **37** (2000), no. 2 190–203.

[105] G. Rohenkohl and A. C. Nobre, *Alpha oscillations related to anticipatory attention follow temporal expectations*, *Journal of Neuroscience* **31** (2011), no. 40 14076–14084.

[106] T. C. Sprague and J. T. Serences, *Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices*, *Nature neuroscience* **16** (2013), no. 12 1879.

[107] R. Desimone and J. Duncan, *Neural mechanisms of selective visual attention*, *Annual review of neuroscience* **18** (1995), no. 1 193–222.

[108] A. Gevins, M. E. Smith, L. McEvoy, and D. Yu, *High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice.*, *Cerebral cortex (New York, NY: 1991)* **7** (1997), no. 4 374–385.

[109] C. S. Herrmann, D. Senkowski, and S. Röttger, *Phase-locking and amplitude modulations of eeg alpha: two measures reflect different cognitive processes in a working memory task*, *Experimental psychology* **51** (2004), no. 4 311–318.

[110] A. C. Nobre and F. Van Ede, *Anticipated moments: temporal structure in attention*, *Nature Reviews Neuroscience* **19** (2018), no. 1 34.

[111] H. Van Dijk, J.-M. Schoffelen, R. Oostenveld, and O. Jensen, *Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability*, *Journal of Neuroscience* **28** (2008), no. 8 1816–1823.

[112] M. Bonnefond and O. Jensen, *Gamma activity coupled to alpha phase as a mechanism for top-down controlled gating*, *PloS one* **10** (2015), no. 6 e0128667.

[113] Y. B. Saalmann, M. A. Pinsk, L. Wang, X. Li, and S. Kastner, *The pulvinar regulates information transmission between cortical areas based on attention demands*, *Science* **337** (2012), no. 6095 753–756.

[114] O. Jensen, M. Bonnefond, T. R. Marshall, and P. Tiesinga, *Oscillatory mechanisms of feedforward and feedback visual processing*, *Trends in Neurosciences* **38** (2015), no. 4 192–194.

[115] J. H. Kaas and D. C. Lyon, *Pulvinar contributions to the dorsal and ventral streams of visual processing in primates*, *Brain research reviews* **55** (2007), no. 2 285–296.

[116] A. Clouter, K. L. Shapiro, and S. Hanslmayr, *Theta phase synchronization is the glue that binds human associative memory*, *Current Biology* **27** (2017), no. 20 3143–3148.

[117] C. S. Furmanski and S. A. Engel, *An oblique effect in human primary visual cortex*, *Nature neuroscience* **3** (2000), no. 6 535.

[118] C. S. Benwell, R. E. London, C. F. Tagliabue, D. Veniero, J. Gross, C. Keitel, and G. Thut, *Frequency and power of human alpha oscillations drift systematically with time-on-task*, *NeuroImage* **192** (2019) 101–114.

[119] K. Friston, *The free-energy principle: a unified brain theory?*, *Nature reviews neuroscience* **11** (2010), no. 2 127.

[120] F. G. Ashby, J. M. Ennis, and B. J. Spiering, *A neurobiological theory of automaticity in perceptual categorization.*, *Psychological Review* **114** (2007), no. 3 632–656.

[121] B. O. Turner, N. Marinsek, E. Ryhal, and M. B. Miller, *Hemispheric lateralization in reasoning*, *Annals of the New York Academy of Sciences* **1359** (2015), no. 1 47–64.

[122] F. G. Ashby, J.-Y. Tein, and J. Balakrishnan, *Response time distributions in memory scanning*, *Journal of Mathematical Psychology* **37** (1993), no. 4 526–555.

[123] M. A. Erickson and J. K. Kruschke, *Rules and exemplars in category learning*, *Journal of Experimental Psychology: General* **127** (1998), no. 2 107–140.

[124] J. N. Rouder and R. Ratcliff, *Comparing exemplar-and rule-based theories of categorization*, *Current Directions in Psychological Science* **15** (2006), no. 1 9–13.

[125] F. G. Ashby and L. Rosedahl, *A neural interpretation of exemplar theory.*, *Psychological review* **124** (2017), no. 4 472.

[126] F. G. Ashby and L. A. Alfonso-Reese, *Categorization as probability density estimation*, *Journal of mathematical psychology* **39** (1995), no. 2 216–233.

[127] M. A. Erickson and J. K. Kruschke, *Rule-based extrapolation in perceptual categorization*, *Psychonomic Bulletin & Review* **9** (2002), no. 1 160–168.

[128] R. M. Nosofsky, *Similarity, frequency, and category representations.*, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14** (1988), no. 1 54.

[129] S. Hélie, S. W. Ell, and F. G. Ashby, *Learning robust cortico-cortical associations with the basal ganglia: an integrative review*, *Cortex* **64** (2015) 123–135.

[130] B. J. Knowlton, J. A. Mangels, and L. R. Squire, *A neostriatal habit learning system in humans*, *Science* **273** (1996), no. 5280 1399–1402.

[131] L. R. Squire and A. J. Dede, *Conscious and unconscious memory systems*, *Cold Spring Harbor perspectives in biology* **7** (2015), no. 3 a021667.

[132] A. Baddeley, *Working memory: looking back and looking forward*, *Nature reviews neuroscience* **4** (2003), no. 10 829.

[133] E. R. Kandel, Y. Dudai, and M. R. Mayford, *The molecular and systems biology of memory*, *Cell* **157** (2014), no. 1 163–186.

[134] T. C. Sprague, S. Saproo, and J. T. Serences, *Visual attention mitigates information loss in small-and large-scale neural codes*, *Trends in Cognitive Sciences* **19** (2015), no. 4 215–226.

[135] M. Riesenhuber and T. Poggio, *Models of object recognition*, *Nature neuroscience* **3** (2000), no. 11s 1199.

[136] M. Corbetta and G. L. Shulman, *Control of goal-directed and stimulus-driven attention in the brain*, *Nature reviews neuroscience* **3** (2002), no. 3 201.

[137] M. J. Buckley, F. A. Mansouri, H. Hoda, M. Mahboubi, P. G. Browning, S. C. Kwok, A. Phillips, and K. Tanaka, *Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions*, *Science* **325** (2009), no. 5936 52–58.

[138] M. Corbetta, G. Patel, and G. L. Shulman, *The reorienting system of the human brain: from environment to theory of mind*, *Neuron* **58** (2008), no. 3 306–324.

[139] A. M. Albers, P. Kok, I. Toni, H. C. Dijkerman, and F. P. de Lange, *Shared representations for working memory and mental imagery in early visual cortex*, *Current Biology* **23** (2013), no. 15 1427–1431.

[140] J. Coull, C. Frith, R. S. J. Frackowiak, and P. Grasby, *A fronto-parietal network for rapid visual information processing: a pet study of sustained attention and working memory*, *Neuropsychologia* **34** (1996), no. 11 1085–1095.

[141] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, *N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies*, *Human brain mapping* **25** (2005), no. 1 46–59.

[142] E. K. Miller and T. J. Buschman, *Rules through recursion: how interactions between the frontal cortex and basal ganglia may build abstract, complex rules from concrete, simple ones*, in *Neuroscience of rule-guided behavior*. Oxford University Press, 2007.

[143] T. J. Buschman and E. K. Miller, *Goal-direction and top-down control*, *Philosophical Transactions of the Royal Society B: Biological Sciences* **369** (2014), no. 1655 20130471.

[144] J. Gottlieb, M. Hayhoe, O. Hikosaka, and A. Rangel, *Attention, reward, and information seeking*, *Journal of Neuroscience* **34** (2014), no. 46 15497–15504.

[145] 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* **9** (1986), no. 1 357–381.

[146] M. I. Posner and S. E. Petersen, *The attention system of the human brain*, *Annual review of neuroscience* **13** (1990), no. 1 25–42.

[147] C. A. Seger, *How do the basal ganglia contribute to categorization? their roles in generalization, response selection, and learning via feedback*, *Neuroscience & Biobehavioral Reviews* **32** (2008), no. 2 265–278.

[148] D. Lopez-Paniagua and C. A. Seger, *Interactions within and between corticostriatal loops during component processes of category learning*, *Journal of Cognitive Neuroscience* **23** (2011), no. 10 3068–3083.

[149] E. G. Antzoulatos and E. K. Miller, *Increases in functional connectivity between prefrontal cortex and striatum during category learning*, *Neuron* **83** (2014), no. 1 216–225.

[150] C. A. Seger, E. J. Peterson, C. M. Cincotta, D. Lopez-Paniagua, and C. W. Anderson, *Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and granger causality modeling*, *Neuroimage* **50** (2010), no. 2 644–656.

[151] E. K. Miller and J. D. Cohen, *An integrative theory of prefrontal cortex function*, *Annual review of neuroscience* **24** (2001), no. 1 167–202.

[152] E. Koechlin and C. Summerfield, *An information theoretical approach to prefrontal executive function*, *Trends in cognitive sciences* **11** (2007), no. 6 229–235.

[153] N. D. Daw, Y. Niv, and P. Dayan, *Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control*, *Nature neuroscience* **8** (2005), no. 12 1704–1711.

[154] J. M. Fuster, *The prefrontal cortex: anatomy, physiology, and neuropsychology of the frontal lobe, 2nd ed.* New York: Raven Press, 1989.

[155] P. S. Goldman-Rakic, C. Leranth, S. M. Williams, N. Mons, and M. Geffard, *Dopamine synaptic complex with pyramidal neurons in primate cerebral cortex*, *Proceedings of the National Academy of Sciences* **86** (1989), no. 22 9015–9019.

[156] S. A. Bunge, *How we use rules to select actions: a review of evidence from cognitive neuroscience*, *Cognitive, Affective, & Behavioral Neuroscience* **4** (2004), no. 4 564–579.

[157] A. De La Vega, L. J. Chang, M. T. Banich, T. D. Wager, and T. Yarkoni, *Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization*, *Journal of Neuroscience* **36** (2016), no. 24 6553–6562.

[158] J. Duncan and E. K. Miller, *Cognitive focus through adaptive neural coding in the primate prefrontal cortex*, *Principles of frontal lobe function* (2002) 278–291.

[159] N. Sigala and N. K. Logothetis, *Visual categorization shapes feature selectivity in the primate temporal cortex*, *Nature* **415** (2002), no. 6869 318.

[160] J. L. McKee, M. Riesenhuber, E. K. Miller, and D. J. Freedman, *Task dependence of visual and category representations in prefrontal and inferior temporal cortices*, *Journal of Neuroscience* **34** (2014), no. 48 16065–16075.

[161] D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller, *Categorical representation of visual stimuli in the primate prefrontal cortex*, *Science* **291** (2001), no. 5502 312–316.

[162] D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller, *A comparison of primate prefrontal and inferior temporal cortices during visual categorization*, *Journal of Neuroscience* **23** (2003), no. 12 5235–5246.

[163] H. Tomita, M. Ohbayashi, K. Nakahara, I. Hasegawa, and Y. Miyashita, *Top-down signal from prefrontal cortex in executive control of memory retrieval*, *Nature* **401** (1999), no. 6754 699.

[164] M. Riesenhuber and T. Poggio, *Neural mechanisms of object recognition*, *Current opinion in neurobiology* **12** (2002), no. 2 162–168.

[165] S. K. Swaminathan and D. J. Freedman, *Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex*, *Nature neuroscience* **15** (2012), no. 2 315.

[166] J. K. Fitzgerald, D. J. Freedman, and J. A. Assad, *Generalized associative representations in parietal cortex*, *Nature neuroscience* **14** (2011), no. 8 1075.

[167] R. Salazar, N. Dotson, S. Bressler, and C. Gray, *Content-specific fronto-parietal synchronization during visual working memory*, *Science* **338** (2012), no. 6110 1097–1100.

[168] L. N. Katz, J. L. Yates, J. W. Pillow, and A. C. Huk, *Dissociated functional significance of decision-related activity in the primate dorsal stream*, *Nature* **535** (2016), no. 7611 285.

[169] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, *Context-dependent computation by recurrent dynamics in prefrontal cortex*, *nature* **503** (2013), no. 7474 78.

[170] J. E. Roy, M. Riesenhuber, T. Poggio, and E. K. Miller, *Prefrontal cortex activity during flexible categorization*, *Journal of Neuroscience* **30** (2010), no. 25 8519–8528.

[171] E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, and T. Poggio, *Dynamic population coding of category information in inferior temporal and prefrontal cortex*, *Journal of neurophysiology* **100** (2008), no. 3 1407–1419.

[172] S. Helie and F. G. Ashby, *A neurocomputational model of automaticity and maintenance of abstract rules*, in *2009 International Joint Conference on Neural Networks*, pp. 1192–1198, IEEE, 2009.

[173] P. Fries, *A mechanism for cognitive dynamics: neuronal communication through neuronal coherence*, *Trends in cognitive sciences* **9** (2005), no. 10 474–480.

[174] P. Fries, *Rhythms for cognition: communication through coherence*, *Neuron* **88** (2015), no. 1 220–235.

[175] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone, *High-frequency, long-range coupling between prefrontal and visual cortex during attention*, *science* **324** (2009), no. 5931 1207–1210.

[176] H. Lee, G. V. Simpson, N. K. Logothetis, and G. Rainer, *Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex*, *Neuron* **45** (2005), no. 1 147–156.

[177] M. Siegel, M. R. Warden, and E. K. Miller, *Phase-dependent neuronal coding of objects in short-term memory*, *Proceedings of the National Academy of Sciences* **106** (2009), no. 50 21341–21346.

[178] N. Cashdollar, U. Malecki, F. J. Rugg-Gunn, J. S. Duncan, N. Lavie, and E. Duzel, *Hippocampus-dependent and-independent theta-networks of active maintenance*, *Proceedings of the National Academy of Sciences* **106** (2009), no. 48 20493–20498.

[179] K. Benchenane, A. Peyrache, M. Khamassi, P. L. Tierney, Y. Gioanni, F. P. Battaglia, and S. I. Wiener, *Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning*, *Neuron* **66** (2010), no. 6 921–936.

[180] H. G. Rey, M. Ahmadi, and R. Q. Quiroga, *Single trial analysis of field potentials in perception, learning and memory*, *Current opinion in neurobiology* **31** (2015) 148–155.

[181] F. A. Soto and F. G. Ashby, *Categorization training increases the perceptual separability of novel dimensions*, *Cognition* **139** (2015) 105–129.

[182] M. Kritzer and P. Goldman-Rakic, *Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey*, *Journal of Comparative Neurology* **359** (1995), no. 1 131–143.

[183] S. N. Haber and R. Calzavara, *The cortico-basal ganglia integrative network: the role of the thalamus*, *Brain research bulletin* **78** (2009), no. 2-3 69–74.

[184] J. M. Fuster, *The prefrontal cortex*, 2015.

[185] H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu, *Interneurons of the neocortical inhibitory system*, *Nature reviews neuroscience* **5** (2004), no. 10 793.

[186] A. M. Bastos, R. Loonis, S. Kornblith, M. Lundqvist, and E. K. Miller, *Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory*, *Proceedings of the National Academy of Sciences* **115** (2018), no. 5 1117–1122.

[187] T. Schreiber and A. Schmitz, *Surrogate time series*, *Physica D: Nonlinear Phenomena* **142** (2000), no. 3-4 346–382.

[188] J. Jeong, J. C. Gore, and B. S. Peterson, *A method for determinism in short time series, and its application to stationary eeg*, *IEEE Transactions on biomedical engineering* **49** (2002), no. 11 1374–1379.

[189] J. T. Townsend and M. J. Wenger, *A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series.*, *Psychological review* **111** (2004), no. 4 1003.

[190] S. A. Oprisan and C. V. Buhusi, *Modeling pharmacological clock and memory patterns of interval timing in a striatal beat-frequency model with realistic, noisy neurons*, *Frontiers in Integrative Neuroscience* **5** (2011) 52.

[191] J. M. Carlson and J. Doyle, *Highly optimized tolerance: Robustness and design in complex systems*, *Physical review letters* **84** (2000), no. 11 2529.

[192] K. Anton-Erxleben and M. Carrasco, *Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence*, *Nature Reviews Neuroscience* **14** (2013), no. 3 188.

[193] D. A. Butts and M. S. Goldman, *Tuning curves, neuronal variability, and sensory coding*, *PLoS biology* **4** (2006), no. 4 e92.

[194] R. van Bergen and J. F. Jehee, *Modeling correlated noise is necessary to decode uncertainty*, *Neuroimage* **180** (2018) 78–87.