Models of Bayesian Learning and Neural Surprise in Somesthesis

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

vorgelegt von

Dipl.-Psych. Kathrin Tertel

Berlin, August 2018

Erstgutachter:	Prof. Dr. Felix Blankenburg
Zweitgutachter:	Prof. Dr. Dirk Ostwald
Tag der Disputation:	14.09.2018

Acknowledgements

First, I want to thank Felix Blankenburg for his ongoing support and supervision, as well as Dirk Ostwald for his mathematical input, endless patience and advice. The Berlin School of Mind and Brain funded most of my work and created a supportive and stimulating structure as well as guidance for this Ph.D., which I am also grateful for.

Thanks also to all of my wonderful colleagues, they created a fun and inspiring work environment and made all the difference: Lisa Velenosi, Pia Schröder, Evgeniya Kirilina, Isil Uluc, Yuan-hao Wu, Timo Schmidt, Till Nierhaus, Alex von Lautz, Jakub Limanowski, Jan Herding, and Simon Ludwig.

Workshops by María Machón and her ongoing support during the whole process of planning and writing have been immensely helpful, I don't know how I could have finished without her. She will be missed.

I thank my family, especially my parents Elisabeth and Jürgen and my siblings Jakob, Anne, and Philipp for being there, being family, listening and proof reading, as well as my nieces Lotta and Thea and nephew Oskar for always brightening my day. Philipp's support, day and night, his pragmatic problem-solving skills paired with his genius mind have been invaluable for me throughout my Ph.D. years.

Contents

Acknowledgements V

Abbreviations IX

Notation X

Zusammenfassung XI

Abstract XIII

1 Introduction **1**

1.1 Computational Theories of Neural Information Processing 3

1.1.1 Theoretical Foundations 3

1.1.2 The Bayesian Brain Hypothesis 6

1.1.3 Predictive Coding 8

1.1.4 Bayesian Predictive Coding and its Competition 10

1.2 Surprise in Sequences 13

1.2.1 Surprise as a Mathematical Concept 14

1.2.2 Surprise as a Model for Brain Signals 16

1.3 The Somatosensory System 21

1.3.1 Architecture of Somatosensory Cortex 21

1.3.2 Perception of Somatosensory Stimuli 22

1.4 General Aim of this Thesis 24

2 Theoretical Modeling **25**

2.1 Introduction 26

2.2 Experimental Stimulation Paradigms 27

2.2.1 Stimulus Sequence Generation 29

2.2.2 Example Paradigm . . . 30
2.3 Computational Models and Surprise Functions . . . 32
2.3.1 Beta-Bernoulli Models . . . 35
2.3.2 Gaussian Random Walk Models . . . 44
2.4 Results . . . 50
2.4.1 Beta-Bernoulli Stimulus Probability Surprise . . . 51
2.4.2 Beta-Bernoulli Alternation Probability Surprise . . . 53
2.4.3 Beta-Bernoulli Transition Probability Surprise . . . 55
2.4.4 Gaussian Random Walk Stimulus Probability Surprise . . . 56
2.4.5 Gaussian Random Walk Alternation Probability Surprise . . . 58
2.4.6 Correlation Between Models . . . 59
2.5 Discussion . . . 60

3 Empirical Application **63**
3.1 Introduction . . . 63
3.2 Material and Methods . . . 66
3.2.1 Participants . . . 66
3.2.2 Stimuli . . . 66
3.2.3 Experimental Procedure . . . 66
3.2.4 EEG Recording and Preprocessing . . . 69
3.2.5 ERP Analysis . . . 72
3.2.6 Single-Trial Analysis . . . 73
3.3 Results . . . 78
3.3.1 Event Related Potentials . . . 79
3.3.2 Single-Trial Analysis . . . 85
3.4 Discussion . . . 89

4 Discussion **94**
4.1 Furthering Computational Modeling of Bayesian Learning . . . 94
4.2 Bayesian Learning in Electrophysiological Potentials . . . 98
4.2.1 Factors Influencing ERP Results . . . 100
4.2.2 Factors Influencing Single-Trial Analysis Results . . . 102

4.3 From Computation to Implementation 105
4.4 Conclusion and Outlook 109

References **111**

Supplement **127**

Participant Instructions and Consent Form 128
Curriculum Vitae 131
Eidesstattliche Erklärung 132

Abbreviations

AP	Alternation probability
BA	Brodman area
BB	Beta-Bernoulli
BL	Bayesian learner
BOLD	Blood-oxygen-level dependent
BPC	Bayesian predictive coding
BS	Bayesian surprise
CS	Confidence-corrected surprise
EEG	Electroencephalography
EM	Expectation maximization
ERP	Event-related potential
FEP	Free-energy principle
(f)MRI	(functional) Magnetic resonance imaging
GLM	General linear model
GRW	Gaussian random walk
ISI	Interstimulus-interval
KLD	Kullback-Leibler divergence
LME	Log-model evidence
MC	Markov-chain
(s)MMN	(Somatosensory) Mismatch-negativity
PC	Predictive coding
PCA	Principal component analysis
PEB	Parametric empirical Bayes
PS	Predictive surprise
ROI	Region of interest
SBL	Sequential Bayesian learner
SEP	Somatosensory evoked potential
SP	Stimulus probability
TP	Transition probability

Notation

o_t	observation of a stimulus at trial t
s_t	(hidden) state at trial t causing observation o_t in a generative model
$[S]$	Iverson brackets, equals 1 if statement S is true, 0 if false
$KL(p(x)\|\|q(x))$	Kullback-Leibler divergence from q(x) to p(x) $\int_{-\infty}^{\infty} p(x)\frac{p(x)}{q(x)}dx$
$x_{a:b}$	$x_a, x_{a+1}, x_{a+2}, \cdots, x_b$
$\delta_a(b)$	Dirac-delta function $\delta(a-b)$
$\text{Bern}(o; s)$	Bernoulli distribution
$\text{Beta}(s; \alpha, \beta)$	Beta distribution
$\Gamma(z)$	Gamma function $\int_0^{\infty} x^{z-1}e^{-x}dx$
$\Psi(z)$	Digamma function (logarithmic derivative of Gamma function)

Zusammenfassung

Laut der *Bayesian Brain Hypothesis* (BBH) verarbeitet das menschliche Gehirn Informationen in Form von Wahrscheinlichkeitsverteilungen und aktualisiert Annahmen über seine Umwelt durch Bayes'sche Inferenz. Da die BBH ein grundlegendes rechnerisches Prinzip für eine Vielzahl von Hirnfunktionen bietet, erlangte sie zunehmend Aufmerksamkeit in den mathematischen Neurowissenschaften. Insbesondere für die Informationsverarbeitung von sequentiellen Reizen bietet die BBH einen geeigneten theoretischen Rahmen, um rechnerische Modelle zur schrittweisen Anpassung von Annahmen über die Umwelt zu erstellen.

Da kein direktes Maß für die unter der BBH postulierten und im Gehirn enkodierten Wahrscheinlichkeitsverteilungen existiert, greift empirische Forschung auf neuronale Überraschungssignale wie zum Beispiel die *mismatch negativity* (MMN) zurück, die mit dem Elektroenzephalogramm (EEG) gemessen wird. Die MMN zeigt sich als ein negativeres Potential nach einem seltenen abweichenden Stimulus in einer Reihe von gleichen Standard-Stimuli. Weil innerhalb eines Beobachtermodells das Ausmaß jeglicher überraschungsbezogenen Aktivierung von der aktuellen Annahme des Beobachters abhängen sollte, kann man Überraschung modellgemäß in Relation zu dieser Annahme quantifizieren. Es bestehen jedoch viele Möglichkeiten, ein Bayes'sches Modell von sequentieller Annahmenaktualisierung und Überraschungsquantifizierung zu erstellen. Die vorliegende Arbeit befasst sich mit der Analyse von mehreren rechnerischen Bayes'schen Modellen für neuronale Überraschungsanworten sowie deren Anwendung in der somatosensorischen MMN als Repräsentation von Überraschung.

Nach einer Einführung in rechnerische Modelle allgemeiner Hirnfunktionen in Kapitel 1 untersucht Kapitel 2 diverse Bayes'sche Modelle von Annahmensaktualisierung in Sequenzen bestehend aus zwei Stimuli, charakterisiert deren respektive Eigenschaften und leitet spezifische Funktionen für jede Modellkategorie her, die ein Ausmaß von Überraschung in Abhängigkeit von einer beobachteten Sequenz quantifizieren. Als Illustration der Modelleigenschaften in einem Testfall werden weiterhin Überraschungsregressoren in einem umfassenden Modellraum bestimmt und gegeneinander kontrastiert.

Kapitel 3 ergänzt diesen theoretischen Ansatz mit einer empirischen Untersuchung von neuronaler Überraschung in der MMN. Überdies geht es auf die Fragestellung zu einer Wahrnehmungsmodalitätenunabhängigkeit von mismatch-bezogener Überraschung ein, indem es sich auf das wenig untersuchte somatosensorische System konzentriert. Versuchspersonen erhielten konsekutive Stimu-

lation des nervus medianus in zwei klar wahrnehmbaren und unterscheidbaren Intensitäten, welche einem Markov-Ketten Paradigma folgte. Um Überraschung im somatosensorischen System, soweit durch EEG messbar, zu untersuchen, nutze ich sowohl gut etablierte Mittelungstechniken, als auch einen feinjustierbareren einzel-trial Analyseansatz, welcher die rechnerischen Modelle und Überraschungsfunktionen aus Kapitel 2 anwendet.

Im Ergebnis zeigt die theoretische Analyse der Überraschungsregressoren, dass scheinbar kleine Unterschiede in Modellspezifikationen zu einander widersprechenden und teilweise kontraintuitiven Überraschungsschätzern führen können. Während die EEG-Studie einen kleinen MMN-Effekt nach Stimuluswechseln aufwies, wurden keine Belege für ein zugrundeliegendes rechnerisches Modell in Kombination mit einer Überraschungsfunktion gefunden. Jedoch zeigte sich auf deskriptiver Ebene konfidenzkorrigierte Überraschung von Stimulus- und Übergangswahrscheinlichkeit unter allen getesteten Modellen als beste Erklärung für die Daten.

Schlussfolgernd lässt sich festhalten, dass der theoretische Ansatz die Wichtigkeit der genauen Konstruktion Bayes'scher Überraschungsmodelle betont. Die alleinige Definition eines rechnerischen Modells für Überraschungsantworten als "Bayesianisch" ist ungenügend, da multiple Bayes'sche Modelle gegensätzliche Vorhersagen über Datenmuster ergeben können. Daher sollten Bayes'sche Modelle sorgfältig erstellt und deren zugrundeliegende Annahmen genau spezifiziert werden. Angesichsts der Studie in Kapitel 3 rechtfertigt weiterhin die Beschaffenheit der somatosensorischen MMN nicht die Annahme eines modalitätsunabhängigen Gehirnprozesses als Basis für diese Komponente. Um einen modalitätsunabhängigen Teil in der MMN in Bezug auf Überraschung zu identifizieren, sollte weitere Forschung Aufmerksamkeits- und Vorhersagbarkeitsfaktoren in somatosensorischen MMN-Paradigmen variieren und rechnerische Überraschungsmodelle eingehend in allen Wahrnehmungsmodalitäten überprüfen.

Abstract

The assumption that the human brain employs information processing in terms of probability distributions and uses Bayesian inference to update beliefs about the world has been summed up in the Bayesian brain hypothesis (BBH). Because of its unifying quality of how to understand brain functioning, the BBH has gained increasing attention in computational neuroscience. Particularly in information processing of sequences spread out in time, the BBH provides an apt framework for computational models of ways in which the human brain updates beliefs about its environment according to new input.

Since there is no direct measure of the belief distributions that are assumed under the BBH, empirical research relies on studying neural surprise responses such as the preattentive mismatch-negativity (MMN), that can be measured using electroencephalography (EEG). Usually, the MMN is found as a more negative EEG potential in response to a rare deviant auditory event embedded in a stream of frequent standard stimuli. Because the magnitude of any surprise-related activity should depend on the current belief held by the observer, surprise can be quantified in relation to this belief. However, there are multiple ways to realize a computational Bayesian model of sequential belief updates as well as of surprise quantification. The scope of this thesis comprises the analysis of several computational Bayesian models for surprise responses as well as their application to the somatosensory MMN as a proxy for surprise.

After an introduction into computational models of brain functioning in Chapter 1, Chapter 2 examines various Bayesian models for belief updates of two-item sequences, categorizes their respective features and derives specific surprise functions for each model category. Further, as an illustration of the models' properties on a test case, I determine surprise regressors from an extensive model space and contrast them against each other.

Chapter 3 complements this theoretical approach by empirically investigating neural surprise in the MMN. Moreover, it addresses the question of perceptual modality-independence of mismatch-related surprise by concentrating on the little-studied somatosensory system. Here, participants received consecutive median nerve stimulation of two clearly perceivable but differentiable intensities according to a Markov-chain roving-like paradigm. To investigate surprise in the somatosensory system as measured from EEG, I use not only established averaging-techniques for event-related potentials, but also a much more fine-grained single-trial analysis approach that employs the computational models and surprise functions from Chapter 2.

As a result, the analysis of surprise regressors from Bayesian models shows that seemingly small differences in model specifications can lead to vastly disparaging, and sometimes counter-intuitive surprise-estimates. While the EEG-study revealed a small MMN-effect after stimulus-alternations, no definite evidence was found regarding an underlying computational model and surprise function. However, on a descriptive level, confidence corrected surprise for stimulus as well as transition probability had the best explanatory value of all computational models tested.

In conclusion, my theoretical approach emphasizes the importance of scrutiny in constructing Bayesian models for surprise. The mere definition of a computational model for surprise responses as "Bayesian" is insufficient, since multiple Bayesian models can make contradicting predictions about data patterns. Hence, Bayesian models have to be carefully built and their assumptions made specific. Furthermore, in light of the study from Chapter 3, the nature of the somatosensory MMN does not warrant the assumption of a modality-independent process in the brain to be the basis of the component. Further research should vary factors of attention and predictability in somatosensory MMN paradigms as well as scrutinize computational models of surprise for the MMN in all perceptual modalities in order to discern a modality-independent part inherent in the MMN relating to surprise.

Chapter 1

Introduction

Imagine walking through grass while looking up into the sky for birds. Not on every step but on most of your steps, grass touches your ankles. You keep walking, not paying attention to the sensation of grass at your ankles, still looking out for birds. At some point, you notice something must have changed on the ground, since there is no more grass touching your ankles. But how did you you notice the absence of something you did not even pay attention to in the first place? And why do you notice it missing now, even though it was not there at every step before? At what point in the sequence of steps, and on what grounds of statistical properties of that sequence of steps did your brain decide that there was a noteworthy change of somatosensory input?

By intuition, we can assume that the more reliable the feeling of grass was in the beginning, and the more reliable and sudden the absence of it, the sooner one would notice a change in the structure of the ground. We can also suspect that the salience of a change depends on the difference in sensory inputs: Suddenly stepping into a puddle of mud might attract ones attention more strongly than the mere absence of grass on the ground.

Having such an internal alert-system that weights a stream of sensory input and detects changes in the surroundings is of enormous importance for highly adaptive organisms such as humans. However, how exactly we perform these functions remains the subject of ongoing debate and extensive research. The present thesis aims at contributing to this work on a theoretical as well as empirical level.

In general terms, the outlined research questions can be framed in the following way: In acquiring continuous sequential sensory information about our surroundings that are not completely

reliable, how do we determine a change in the (hidden) causes for this sensory input? For the integration of new data into an existing hypothesis, Bayes' theorem (1763) provides the optimal answer from a computational viewpoint. In a sequence of time points t, the probability of a hidden state s_t given the current observation o_t is written as $p(s_t|o_t)$ and calculated by

$$p(s_t|o_t) = \frac{p(s_t)p(o_t|s_t)}{p(o_t)}, \tag{1.1}$$

where $p(s_t)$ is the probability of a hidden state s_t prior to having made observation o_t (subsequently called *prior*), $p(o_t|s_t)$ is the probability of making observation o_t given s_t (sometimes also referred to as *likelihood*), and $p(o_t)$ is the unconditioned probability of the observation. Another way of applying this relationship of conditional probabilities is by making use of the proportionality

$$p(s_t|o_t) \propto p(s_t)p(o_t|s_t) \tag{1.2}$$

without any representation of $p(o_t)$.

It is plausible to wonder if, and in what way, the (human) brain processes sequential information and infers possible changes in environmental states according to Bayes' theorem. In terms of the initial example, this would mean that we encode a probability distribution over the sensation of grass on our feet that is updated at each step. When encountering events with a sufficiently low probability according to the encoded distribution, such as a number of steps without the sensation of grass, a surprise response might be triggered and attentional resources allocated to the surroundings on the ground. However, before investigating any variant of this question regarding Bayesian information integration and surprise responses empirically, a range of theoretical computational questions need to be addressed.

In this thesis, I will tackle these theoretical questions by introducing sequential Bayesian learner models. In a second step, I will apply those models to EEG data from an experiment employing sequential stimulation. I study sequential stimulus processing in the *somatosensory* domain, because contrary to auditory and visual perception, few studies have concentrated on somatosensory stimulus processing, and because only evidence across sensory modalities warrants drawing general conclusions regarding theories of information processing in the brain.

The thesis is structured as follows: In Chapter 1, I review theories and empirical evidence related to Bayesian learning and surprise signals in the brain. While most of the evidence for the

reviewed theories comes from visual and auditory perception, I will also regard the much-neglected somatosensory system (which is concerned in the experiment of Chapter 3).

Chapter 2 entails the formulation and comparison of several versions of sequential Bayesian learner models and ways to quantify surprise for time series of two different stimuli. Here, the degrees of freedom and the ways in which computational models derived from identical theories can make disparate predictions will be of particular interest.

These models are put to the test in Chapter 3, where I will study trial-by-trial variations of the somatosensory mismatch negativity as a proxy for neural surprise. Finally, all findings and their connections to existing theories as well as open research questions are discussed in Chapter 4.

1.1 Computational Theories of Neural Information Processing

During the last 150 years, two fundamental theoretical concepts have emerged in neuroscience describing perception and information processing in the brain in unified computational frameworks: the Bayesian Brain Hypothesis (BBH) and Predictive Coding (PC). In this Section, I will review their historic-scientific foundations in a larger context (1.1.1) and examine each one more closely on its own (1.1.2 and 1.1.3) before considering their joint value and criticism (1.1.4).

1.1.1 Theoretical Foundations

Many current computational theories of information processing in the brain (e.g., Friston, 2005; Dayan et al., 1995; Heeger, 2017; Daw, 2013) refer back to the work of Helmholtz (1891). In his study of perception using psychophysical experiments of visual illusions, he treated the brain as a system trying to solve an inverse problem, namely inferring the causes of sensory input. He went on to conclude that perception is not a passive process, a mere reception of input, but to the contrary, an unconscious probabilistic inferential process in which the nervous system infers the *most likely* causes of any given sensory input.

In the same line of reasoning, Mach (1902) assumed in his writings that all sensory modalities create experiences which are continuously updated according to the latest observations in the most economic way, leading to a present *temporary state of collective science*. Mach draws an analogy between the scientific system of evidence collection and theory support or refutation, and the way

that we perceive and infer what is "true" in our current surroundings. Hence, around the turn of the 19$^{\text{th}}$ century, the idea that perception is an inferential process much like science that is constantly updating, emerged. However, a clearer concept of this updating process as a common principle of the brain was lacking at the time.

Half a century later, the cyberneticist Ross Ashby added to these ideas with his view of the brain as a self-organizing system, striving for homeostasis (Ashby, 1947). More generally, in an undated aphorism[1], Ashby summed up the whole function of the brain as error-correction. Along his lines, perception is rather seen as a consequence of this error-correction process and the achievement of homeostasis, and not so much defined by the scientific building of an inner model of the world.

Attneave (1954) connects to this view, emphasizing how perception itself is guided by expectations and predictions, and inextricably linked with these. On the other hand, he notes that the information received by any higher organism is highly redundant, and from an information-theoretical standpoint, a major function of the nervous system is to discard this redundancy and to encode a more economic form of information than what initially hits the receptors (Attneave, 1954, p. 189). For an informational analysis of these functions, he deems events that are ordered in time particularly suitable (which we will return to later).

Shortly after, Barlow (1961) formulated this *redundancy-reduction hypothesis* mathematically (p. 225) and derived testable predictions regarding neuronal spikes along a sensory hierarchy. In his view, this hypothesis is especially useful as it does not regard the senses blindly as some machinery to be figured out, but as tools that an organism employs for certain goals. This, he assumes, is why these goals (e.g., redundancy reduction) can be expected to be reflected in the nervous system itself. Using the case of the retina, Srinivasan et al. (1982) showed possibly one of the first instances of experimental data from the nervous system supporting *predictive coding* (PC). This code is predictive in the sense that correlations about features of natural visual scenes known to the organism are employed to reduce redundancy (according to Attneave, 1954; Barlow, 1961) in passing on the information to the next neuronal relay.

Notably, error-correction is one way for sparse coding and redundancy reduction that has since been studied extensively in nervous systems. Especially Mumford (1992) laid out a detailed theory of how the neocortex can achieve such a computational goal. Because the brain's architecture and in particular that of the cortex are thought to be highly hierarchically structured (with sensory input first reaching lower-level primary cortices and being processed with increasing abstraction

[1] from the Ross Ashby Digital Archive at `http://ashby.de/rossashby.info/`

in higher levels), Mumford's model accommodates two levels of cortical hierarchy, one closer to the sensory input and one of higher abstraction. While the lower level feeds its input forward to the higher level, the higher one tries to create an abstraction fitting to the lower level and sends back that model to the lower level via deep pyramidal cells. The lower level in turn will pursue a reconciliation of the model with its own input, passing up in the hierarchy only that information which has not been entailed by the previous feedback from the abstract level. Thus, Mumford's model depicts a scheme for error correction, strive for homeostasis, and greater efficiency with less redundancy all at once. In addition, it reflects a model of the *causes* of sensory input (feedback connections from higher, more abstract processing levels) previously assumed by Helmholtz.

The most prominent study of predictive coding in cortical neurons arguably comes from Rao and Ballard (1999). Connecting to the works of both Srinivasan et al. and Mumford, they propose a computational model in which predictions are fed back to lower levels of a perceptual hierarchy, and prediction error signals are sent forward up the hierarchy to improve predictions. Their suggested algorithm for prediction optimization is equivalent to Bayesian model selection as formulated in Equation (1.1), whereby the model or state s with the highest posterior probability $P(s|o)$ is assumed to be true. With their algorithm, Rao and Ballard could simulate certain extra-classical receptive field effects observed in visual cortex neurons, thus providing strong evidence in favour of PC in the visual system.

Around the same time, Knill and Richards (1996) publish their concept of a *Bayesian Brain Hypothesis* (BBH), stating "that the brain represents sensory information probabilistically, in the form of probability distributions" (Knill and Pouget, 2004, p. 712). Invoking the numerous ways in which humans act as optimal Bayesian observers (i.e., according to Bayes' theorem), they review psychophysical evidence in support of their hypothesis. In accordance with the BBH, for example, perceptual and motor-signals are integrated in a Bayes-optimal fashion, taking into account their respective uncertainties. After a long time of disregard for Bayes-optimality in human behavior and brain processing (Friston, 2012), the idea regained pupularity.

Under the larger framework of the free-energy principle (FEP; Friston, 2005; Friston and Stephan, 2007), Friston described a way in which redundancy-minimization and Bayesian updating complement each other in the brain (2010). The FEP rests on the notions that adaptive systems (i.e., biological agents) update their beliefs about the world in a Bayesian way and the minimize prediction errors throughout the cortical hierarchy. In addition to these concepts, his

model contains an *enactive* component. Free energy quantifies the amount of data not explained under the current model or internal states, which can be minimized either by adjusting the internal states (model-updating according to Bayes) or by acting upon the world in order to create a better fit between inner and external states. Both will lead to internal model distributions that have a smaller variance (or a higher precision, which equals inverse variance), and hence, decrease uncertainty about the world. Thus, according to Friston, organisms not only build internal models of the world to reduce free energy, but also act, explore and experiment to make the world fit the inner model. In a recent interview with Friston (2018), he relates the FEP to PC conforming to a principle-to-process-theory relationship. As a principle, the FEP is admittedly unfalsifiable but it employs PC mechanisms as a process theory, from which falsifiable hypotheses can readily be derived.

Finally, to categorize this plethora of frameworks or models, Marr's *levels of analysis* for machines or organisms carrying out an information-processing task are a very useful tool (1982, p. 25). For Marr, a *computational* theory specifies the goal of a computation that is carried out, as well as the logic behind the strategy to do so. On the level of *representation and algorithm*, representations for in- and output along with a possible algorithm reaching the computational goal are defined. Lastly, the level of *hardware implementation* assumes a physically possible implementation of the algorithm.

In the next two subsections, the concepts of BBH and PC are discussed in more detail as they relate to the human brain and bear implications for my work in Chapters 2 and 3. Furthermore, an attempt to place them in Marr's levels of analysis according to their varying levels of specificity is made.

1.1.2 The Bayesian Brain Hypothesis

According to the Bayesian Brain Hypothesis (BBH), new observations o_t will change an internal model regarding the causes of the observation s_t in agreement with the Bayes' theorem (see Equation (1.1)). The BBH stresses the notion of the (human) perceptual system as a statistical inference tool, with the function to infer probable causes of sensory input (Dayan et al., 1995). Because such a form of inference does not require labeled data, it is regarded as unsupervised learning. Moreover, since it is generally assumed to function over probability distributions and not only point estimates, it carries a strong emphasis on the processing of the uncertainty of an

input along with the input itself. Using Bayes' theorem to update an hypothesis about a state of the world after a new piece of information (e.g., perceptual input) is the optimal strategy from a probability-theoretical point of view. Seeing as humans do not always perceive and behave optimally, BBH is subject to strong criticism (c.f., Bowers and Davis, 2012a,b). However, aside from assuming optimality for the brain, the BBH can contribute to neuroscience by making *normative* predictions about how an ideal perceptual system combines prior knowledge with perceptual input. Furthermore, the BBH can providing algorithms that can function as mechanistic interpretations of neural circuits in the brain (Doya, 2007, p. xi). Since the human brain faces problems of information integration, and humans solve them, under many circumstances, close to optimally (Knill and Pouget, 2004), a good model of how the brain works must be able to solve those problems as well. Framing problems in a Bayesian way can then provide better understanding of the relevant variables and their interplay (Kording, 2014).

Aitchison and Lengyel (2017) call Bayesian inference one of the brains fundamental computational goals, and thus, place it on the *computational* of Marr's levels. Friston (2018) sees it as not committing to a particular process theory, in that it only asks of implicit beliefs to conform to Bayes' rule. Unlike PC, the BBH makes explicit how generative models should be adjusted according to new input, and while it can utilize PC, it can also be realized by other representations in the brain (Aitchison and Lengyel, 2017).

Initially, the most pieces of supporting evidence for the BBH stemmed from psychophysical studies on integration of cues and of sensory and motor signals (Knill and Pouget, 2004). Knill and Pouget show that complicated patterns of perceptual biases can be explained by simple Bayesian models without explicit codes for those biases in them (2004, p. 718).

After Bayesian theories of motor coordination (Todorov and Jordan, 2002) at different timescales (Kording et al., 2007), and of inductive learning and reasoning (Tenenbaum et al., 2006), the BBH has also become a powerful model in reinforcement-learning. Here, Bayesian models can improve classical theories by describing how a learner should actively probe the environment to learn optimally (Kruschke, 2008). This property of the BBH makes it an apt tool for Friston's FEP, which stresses the ability of organisms to act upon the world to understand it (2018). In the seminal work by Behrens et al. (2007), participants tracked the volatility of the experimental environment to improve their decision making, wherein volatility was found to be represented by an fMRI signal in the anterior cingulate cortex.

As for a possible neural implementation of Bayesian inference in the brain, many models with varying degrees of specificity have been proposed. In accordance with his PC implementation, Rao (2004) suggested a way for recurrent networks of noisy integrate-and-fire neurons to perform approximate Bayesian inference, with belief propagation in the log domain represented by membrane-potential dynamics of neurons. The formulation of probabilistic population codes for Bayesian inference by Ma et al. (2006), where larger populations of neurons automatically represent probability distributions, has also gained considerable attention and was later adapted to model decision making (Beck et al., 2008). Beck et al. assume that neurons in the lateral intraparietal cortex involved in evidence accumulation encode a probability distribution over the perceptual input in a trial-wise manner that should be able to predict performance in a perceptual decision-making task, which they support by presenting consistent evidence. To make up for unjustified assumptions in most implementations using probabilistic population codes (e.g., a uniform distribution of sensory variables), Ganguli and Simoncelli (2014) propose a model with *heterogeneous* neural population codes that contains less strict assumptions about sensory input.

One objection often raised against the BBH is the apparent inability of humans to efficiently use conditional probability statements in reasoning and the fact that they often resort to heuristics and biases (as famously described by Tversky and Kahneman, 1974). A convincing model to bridge this supposed gap comes from Sanborn and Chater (2016), who assume that the brain does not directly represent probability distributions but that it samples from them (see also Griffiths et al., 2012), and that many cognitive biases and reasoning errors actually result from such a sampling process.

Finally, while many BBH models rely on PC (such as Rao, 2004), there are also numerous ideas of a Bayesian brain not relying on predictive codes: The aforementioned probabilistic population codes (Ma et al., 2006), probability and log-probability codes representing parameters of the posterior (Fiser et al., 2010), or direct-variable coding (Olshausen and Field, 1996), just to name a few.

1.1.3 Predictive Coding

In Predictive coding (PC), sensory input or bottom-up information x is not necessarily passed on as is. Rather, only that part which is not explained by prediction μ from a higher level, i.e. the prediction error ϵ, gets relayed. In classical predictive coding algorithms, this is achieved by

subtraction:

$$\epsilon = x - \mu \tag{1.3}$$

The prediction error ϵ can then be used to make adjustments to the prediction, whereas its absence can understood as a perfect prediction. *Hierarchical* PC uses such a feedforward-feedback loop on several hierarchical layers of abstraction, such as in the seminal work by Rao and Ballard (1999). In such a hierarchical setup, PC emphasizes a very economic information transmission through the cortical hierarchy: The better the predictions, the less capacity or "band-width" for information-processing is needed on average, because the sensory input will be "explained away" (Friston and Stephan, 2007), and thus, corresponds to the redundancy reducing hypothesis by Attneave (1954); Barlow (1961). In addition, the assumption of PC in the brain makes sense of the numerous feedback-connection to be found in the cortex (Friston, 2005).

While many different algorithms for PC in the nervous system exist in the literature (see Spratling, 2017, for a comprehensive overview), the theory itself can be placed on the computational of Marr's three levels, in detailing just the computational goal (error reduction through prediction of incoming signal) and not necessarily the way in which an updating of predictions through prediction errors takes place. However, Aitchison and Lengyel (2017) argue for PC to be a *common algorithmic motif* emerging in different computations a neural system has to make, such as maximizing information transmission, canceling effects of self-generated actions, representing continuous quantities using spikes, or reinforcement learning. Hence, in their view, PC describes the representation and algorithm of an information-processing task.

For Friston (2018), the predictive part of PC is not about predictions in time, but more about what is happening in the present, under an organism's current beliefs or expectations about how its sensations are caused. In this way, PC renders the system to be self-adjusting and self-organizing.

Many different ideas about the implementation of PC in the brain have been formulated. A common theme among these representations is a duplex architecture of two functionally distinct subpopulations of neurons, one for representations of the input signal and one for prediction error (Friston, 2005). The leading idea for those is that superficial pyramidal cells (prominent in forward connections) act as error units, sending prediction error signals up the hierarchy, while deep pyramidal cells pass predictions downward (Friston, 2005; Mumford, 1992). Oscillatory brain activity associated with either message-flow might also be distinct: feedforward prediction errors are likely passed in the high gamma ($40 - 90$ Hz), feedback predictions in the lower alpha/beta

bands ($5-15$Hz van Kerkoerle et al., 2014). These features are embedded in the influential work by Bastos et al. (2012), outlining a canonical microcircuits that specifies how the basic computations of PC can be performed under the FEP.

In visual perception, PC can explain firing patterns of single neurons on several stages of the cortical hierarchy (Srinivasan et al., 1982; Rao and Ballard, 1999), as well as fMRI BOLD activiation (Murray et al., 2002). It is also applied to many brain-functions outside of basic perception, for example as a mechanism for analogue digital-conversion (Denève and Machens, 2016) or in reward-learning (Schultz et al., 1997). Nevertheless, Aitchison and Lengyel (2017) conclude that the overall-evidence for predictive coding as a unified brain theory seems inconclusive and cannot rule out other forms of coding because experimental paradigms did not allow for a clear distinction.

1.1.4 Bayesian Predictive Coding and its Competition

In light of the implementations of the BBH and PC discussed above, the theories can easily be integrated but could also be realized without one another in the brain. Here, I briefly discuss their mutual advantages, a common form of Bayesian predictive coding (BPC), and criticism, as well as a competing model for cortical function by Heeger (2017).

PC and BBH both formulate computational goals, and for both, algorithms and ideas for a concrete neuronal implementation are available from the literature. Thus, in my view, they cannot be clearly placed on different layers of Marr's levels of analysis. However, PC is mostly considered to be more on the algorithmic or process-theory side than the BBH (Aitchison and Lengyel, 2017; Clark, 2013), most likely due to its computational simplicity (subtractions are easier to perform by neuronal networks than multiplication and division Aitchison and Lengyel, 2017) and abundant formulations in hierarchical form for cortical microcircuits (e.g., Mumford, 1992; Spratling, 2008; Bastos et al., 2012).

According to Aitchison and Lengyel (2017), PC is an algorithmic motif possibly serving many computational goals, with Bayesian inference being one of those conceivable goals. For Friston, the BBH is a corollary of the FEP and can be realized through a process like PC, while he still deems both to be incomplete as they are missing an "enactive" component of how we infer states of affairs (2018).

To formulate a joint theory from PC and BBH, Aitchison and Lengyel (2017) set the updated

prediction μ_t to

$$\mu_t \coloneqq \int o_t p(o_t|s_t) d_{o_t} \tag{1.4}$$

yielding a PC scheme with Bayesian prediction-updates. Models of BPC mostly use both *direct coding* neurons to ease the computation of predictions as well as *predictive coding* neurons (as for exaple in Rao and Ballard, 1999). Other possible versions of BPC are given by, among others, Mathys et al. (2011, 2014) and Gershman et al. (2015).

It should be noted, that behavior based on Bayesian computations, while being optimal in the use of new information, can, in principle, come with certain disadvantages. This is the case because priors will influence the integration of information and could be misleading if they are set to strong biases. However, this is rarely the case, and most of the time, priors help to make better choices (Vilares and Körding, 2011).

In light of these compelling unified theories, it is very tempting to accept such a common principle of information processing in the brain, in the form of either BBH, PC, or BPC, as a simpler and thus preferred explanation for brain functioning (according to Occam's razor; see Rasmussen and Ghahramani, 2001). While the BBH is admittedly a very loose concept demanding only that the brain somehow integrate information according to Bayes' rule and take uncertainty into account, Clark (2013) sees hierarchical PC to lead to a clearer picture, going as far as the specification of cortical microcircuits (Bastos et al., 2012).

Nonetheless, there remains a large gap between BPC models and their implementation (Clark, 2013). In general terms, these models mostly lack the specification of a cognitive architecture, of how the brain divides its cognitive labors (although here, some progress has been made by Zénon et al., 2017), and of what aspects of the world get sensorially coded at all (Clark, 2013, p. 14). For the Bayesian side, Vilares and Körding (2011) agree, stating that although there are many theoretical proposals for how uncertainty is represented in the brain, there is not much experimental support for any of them. In more detail, Aitchison and Lengyel (2017) note that hybrid models of PC and direct-coding neurons still fall short of determining which phenomena are specific to predictive coding neurons and of clarifying how to map both types of neurons onto specific cell types in the cortex. Further, Kogo and Trengove (2015) have pointed out problems of the inner logic of the highly acclaimed canonical microcircuit model by Bastos et al. (2012), pertaining to the computations and implementations of prediction error, that should not be ignored.

Heilbron and Chait (2017) have recently reviewed evidence for (mostly Bayesian) PC schemes

in the auditory domain, which is much more dependent on information that is extended in time than the ample-studied visual perception. They declare that evidence for the commonly proposed functions of oscillatory frequency bands for predictions (beta/alpha) and prediction errors (gamma) is scarce, indirect and limited (but see Sedley et al., 2016, for evidence from electrocorticography). Furthermore, Heilbron and Chait argue that the literature is not clear on how the concept of precision relates to regularity in the auditory signal, as there are studies claiming it enhances (Barascud et al., 2016), others suggesting it suppresses (Sohoglu and Chait, 2016) neural activity.

Other model types have been proposed that account for the same effects as the BPC schemes. Carandini and Heeger (2012), for example, propose a model of divisive normalization, in which cells compute a ratio between bottom-up inputs and the summed activity of a pool of neurons, as a canonical computation in the brain. Their model can account for neuronal effects such as saturation, cross-orientation suppression, and surround suppression, and can also establish a modulation of neural activity by attention.

Later, Heeger (2017) introduced an even more global theory of cortical function that does not rely on BPC. Rather, it aims at being able to carry out three major *functions* of the neocortex: the abilities to make inferences (with perception being unconscious inference), to explore and to predict (while not necessarily using a predictive code in the neural system). In summary, Heeger's hierarchical model describes neural activity in each brain area and at each level of the processing-hierarchy to consist of *feedforward* (i.e., bottom up from a lower hierarchical level), *feedback* (top-down context from a higher level) and *prior* drive, which he calls expectation. State parameters control their relative contributions, that could be implemented through neuromodulators and oscillations. Accordingly, the model can perform inference through a combination of input and prior states (and in such a way approximate Bayesian inference), encourage exploring behavior because inference relies on neural response variability (i.e., noise), and make predictions in time by utilizing the inference results from different layers of the hierarchy and thusly make predictions on various timescales. This model connects well with previous ideas of neural population codes for Bayesian inference (Ma et al., 2006; Beck et al., 2008; Ganguli and Simoncelli, 2014), because it as well entails such an implicit representation of the posterior distribution. However, it differs greatly from PC in the way that prediction errors are implemented: Heeger models them as feedback connections, and as such they have the opposite directionality of conventional BPC models (as, e.g., Rao and Ballard, 1999). In the case of neural adaptation phenomena, where repeated repre-

sentation of a stimulus leads to a swift decrease of neuronal activity, PC theories assume a total "explaining away" of receptor input and thus zero (or baseline) neuronal activity, which typically does not happen, while Heeger's model can represent this remaining activity accurately.

In summary, many open questions regarding the brain's use of BPC remain. Empirical studies are necessary that do not only enable the support of one model, but also the distinction between several models that claim to express a unified way of how the brain works. Lastly, there seems to be a consensus among researchers that the brain is able to, in at least some instances, perform Bayesian inference as well as entertain a predictive code in the sense that unpredictable input leads to a stronger neuronal response than a predictable one in at least some neuronal populations. The main questions that computational neuroscience has to answer then relate to the circumstances under which Bayesian inference and PC are performed, how exactly they are implemented by the brain on a micro- and macro-scale, and how a unified computational model could encompass both predictive and direct coding, as well as the degree to which Bayesian inference plays a role in brain activity (as, for example, done with the state parameters in the theory by Heeger, 2017). The present thesis aims at contributing to this ongoing debate by investigating simple sequential updating processes within a Bayesian scheme, thusly in accordance with the BBH. While PC could be one mechanism with which surprise is processed during such Bayesian updating, other means of coding are possible as well.

1.2 Surprise in Sequences

Research on computational models of brain activity often relies on the concept of surprise to empirically test a model. From a theoretical point of view, surprise is a convenient approach to find out what a model already "knows" by establishing what it does not know (i.e., by which input it is "surprised"). On the empirical side of neuroscience, surprise can be a relevant factor in determining costs for cognitive processes (as done by Zénon et al., 2017). Moreover, it has been found to guide attention (Itti and Baldi, 2009) and can have a strong impact on memory formation (Calvillo and Gomes, 2011; Wallenstein et al., 1998; Ranganath and Rainer, 2003). Therefore, surprise constitutes a relevant cognitive process in itself.

So-called neural surprise responses such as the mismatch negativity (MMN; Näätänen et al., 1978) and the P300 (Sutton et al., 1965) have been investigated for decades, and thus, comprise very well-studied responses to serve as a testing bed for computational modeling. In addition,

surprise responses have also been found with fMRI (Iglesias et al., 2013; O'Reilly et al., 2013).

Many researchers define surprise conceptually as an emotion originating from a mismatch between an expectation and the actual experience of an event (e.g., Ekman and Davidson, 1994). Thus, for experiencing surprise, an expectation or prediction are necessary, which typically relate to specific moments in time. This connection to time is a key distinction of surprise from *novelty*, which is determined in relation to a memory recall and, theoretically, should not depend on the moment in time at which a novel item is presented (Barto et al., 2013, although some researchers define surprise as novelty, e.g., Wessel et al., 2012). Next, we will see how concrete, quantified definitions vary on a theoretic-computational (1.2.1) and on an empirical neuro-scientific (1.2.2) level.

1.2.1 Surprise as a Mathematical Concept

The amount of surprise in a model or organism depends on the current state of predictions at the moment of surprise. Different mathematical formulations have been proposed to relate surprise to a predictive distribution. Here, we review surprise in the forms of predictive surprise, Bayesian surprise, and confidence-corrected surprise.

For these general surprise formulations we define a sequential Bayesian learning (SBL) scheme that updates its assumptions about hidden states of the world s_t according to new observations o_t and uses the posterior at instance t as the prior at $t+1$:

$$p(s_{t+1}) \coloneqq p(s_t|o_t) = \frac{p(o_t|s_t)p(s_t)}{p(o_t)} = \frac{p(o_t|s_t)p(s_t)}{\int_{s_t} p(o_t|s_t)p(s_t)ds_t} \tag{1.5}$$

Notably, the concept of a hidden state of the world can be seen on various different levels. While the term itself refers to a state of the environment that is not immediately accessible to the observer, this immediacy is entirely up for debate and to be defined in the realm of the object of research. Usually, in psychophysical experiments, any property of a perceived stimulus is seen as a hidden state, since there is no instance of perception that is not mediated via the senses. However, within the scope of sequential Bayesian learning, I will investigate hidden states relating to stimulus, alternation, and transition probability in Chapter 2.

Apart from hidden states, the exact quantifications of surprise hinge on other concrete model parameters such as the distributions chosen to model prior and posterior, as well as the volatility assumed in the environment (see, for example, Meyniel et al., 2016). Chapter 2 is concerned with

different explicit computational models that involve Bayesian learning during sequential stimulus input of two different items. Here, as well as in Chapter 2, we treat observations and states as random variables with continuous state spaces. It should be noted, however, that some other models also operate on a discrete state space and point estimates instead of distributions (e.g., Seer et al., 2016).

Predictive surprise (PS) is usually defined as the negative logarithm of the predictive distribution that resulted from all previous observations. Under the assumption of our sequential Bayesian learning scheme from Equation (1.5), all previous observations are subsumed in s_t, giving

$$PS(o_t) \coloneqq -\ln(p(o_t|s_t)) \tag{1.6}$$

as a PS function. This formulation has its roots in Shannon's concept of information (1948), and can also be found as *surprisal* in linguistic science (Tribus, 1961).

Bayesian surprise (BS) refers to the Kullback-Leibler divergence (KLD) of the prior and posterior distributions for the hidden state s_t:

$$BS(o_t) \coloneqq KL(p(s_t)||p(s_t|o_t)). \tag{1.7}$$

BS has initially been proposed as a measure of surprise by Itti and Baldi (2009).

Confidence-corrected Surprise (CS) has recently been suggested by Faraji et al. (2018) as a more apt measure of surprise. It is defined as the KLD between the prior $p(s_t)$ and the posterior of a naïve observer $\hat{p}(s_t|o_t)$, i.e. an observer who has a flat prior $\hat{p}(s_t)$ and observed o_t.

$$CS(o_t) \coloneqq KL\left(p(s_t)||\hat{p}(s_t|o_t)\right) \tag{1.8}$$

This formulation is motivated by the reasoning that an event should not elicit surprise if the observer has not yet committed to a model that favors one specific event over others. Faraji et al. show that, taking into account a measure for commitment to model $p(s_t)$, which is defined as the negative entropy of the model

$$C(p(s_t)) = -H(p(s_t)) = \int_{s_t} p(s_t)\ln(p(s_t))ds_t, \tag{1.9}$$

and a data-dependent constant scaling the state space

$$O(t) \coloneqq \int_s p(o_t|s_t)ds_t, \tag{1.10}$$

CS can also be expressed as a linear combination of these measures:

$$CS(o_t) = BS(o_t) + PS(o_t) + C(p(s_t)) + \ln O(t) \tag{1.11}$$

A derivation of the equality is given in Equation (2.17) in Chapter 2.

Both PS and BS are widely used in neuroscientific applications of SBL models, and the very recent CS as a combination of BS and PS can also readily be employed with such a model. In Chapter 2, I will clarify circumstances in two-item sequences and SBL models under which these three definitions yield strikingly different quantifications of surprise.

A completely different perspective on surprise quantification is given by Maguire et al. (2018). In their view, surprise does not merely relate to the experience of an improbable event, but specifically to an event that exhibits *randomness deficiency*, i.e., observing patterns when a model predicts random noise. They quantify randomness-deficiency as Kolmogorov-complexity, i.e. the "compressability" of a signal. Thus, the better a signal can be compressed in relation to its a-priori probability, the more randomness-deficient it is, and the more surprise it will evoke in an observer. As Maguire et al. show, this definition of surprise works very neatly with sequences of digits (e.g., lottery results) that are perceived all at once and subsequently judged to be surprising or random. However, the authors do not make any further statements about how a-priori probabilities are estimated in examples that are not as well defined as lottery drawings, nor do they suggest a way to apply their surprise measure to sequentially observed events, which is why we do not consider randomness-deficiency in Chapter 2. Nonetheless, Maguire et al. state that their approach is equal to a Bayesian model at the limit.

1.2.2 Surprise as a Model for Brain Signals

Empirical evidence from human subjects through noninvasive studies is of crucial importance to make generalizable claims about brain functioning, as is made in the BBH and PC. Here, I review evidence from two surprise-related potentials from electroencephalography (EEG), the mismatch negativity (MMN) and the P300.

While the MMN refers to a preattentive negative potential for unexpected stimuli within a sequence $100 - 200$ ms post-stimulus, the P300 can be measured 300 ms after a to-be-detected stimulus, and thus requires top-down attention toward the respective perceptual input. Because of its high temporal resolution, EEG is especially apt to test timing-related predictions from BBH and PC theories.

For both the MMN as well as the P300, there is some evidence that their amplitude expresses the amount of surprise under SBL conditions (Lieder et al., 2013a; Kolossa et al., 2015). However, concrete evidence remains scarce. This is due to two problems with existing studies: First, classical analysis of event-related potentials (ERP) in EEG rely on averaging across many trials to remove noise in the signal. However, averaging can also remove relevant (and sometimes, on the basis of models, expected) fluctuations in response amplitudes grossly classified into one condition (Blankertz et al., 2011; Mars et al., 2012). Particularly, the BBH and PC rely on computations that are different for every trial and impossible to express in terms of fixed conditions across which one could average. Second, since the model-based single-trial analysis which mitigates these problems is a fairly recent development in EEG research, a space of likely, but distinguishable SBL models in combinations with different surprise functions has not been exhaustively tested. In the following, I review the existing evidence and point out possible gaps.

For an MMN (Näätänen et al., 1978), the change in a stimulus attribute of a previously repetitive (i.e., standard) auditory stimulus elicits a more negative potential $100 - 250$ ms after the deviant stimulus. This difference between standard and deviant waveforms is usually largest over centro-frontal electrodes.

First found in the auditory domain, analogues have since been discovered in the visual (Tales et al., 1999; Czigler, 2007), somatosensory (Kekoni et al., 1997; Restuccia et al., 2007; Spackman et al., 2010), and possibly even olfactory (Krauel et al., 1999; Sabri et al., 2005) modality. Importantly, the MMN is conceived as a preattentive response, meaning that it can be evoked without any top-down attention towards the stimulus (Näätänen, 1992) and thus is an indication of an automatic information processing mechanism of the brain (but see Auksztulewicz and Friston, 2015).

For the auditory, visual and somatosensory domains, there is considerable evidence for Baysian learning and PC playing a role in the formation of the MMN (Garrido et al., 2009; Wacongne et al., 2012; Lieder et al., 2013a; Stefanics et al., 2018; Ostwald et al., 2012). The apparent independence

of perceptual domain is another strong lead in favor of theories for unified brain mechanisms such as the BBH and PC. However, all of these studies used different models to explain the MMN. Notably, these models can possibly make contradictory predictions about the amount of surprise at a given trial, which I will show in Chapter 2.

In an attempt to unify the previous competing theories for MMN formation, namely adaptation and model-adjustment, Garrido et al. (2009) propose PC as a mechanism behind the MMN. According to the authors, while adaptation cannot explain all MMN effects (such as the susceptibility to more abstract rule violations, see Horváth et al., 2001), thus making a model-adjustment process more likely, both competing theories can be explained by predictive coding, just at different hierarchy levels (with adaptation being explained by lower-level predictions and abstract rule violations by higher-level predictions).

In more detail, Wacongne et al. (2012) specify a precise neuronal model of realistic synapses, receptors, and spiking neurons for the MMN using PC. In their model, the authors assume separate neuronal populations for prediction errors and predictions as well as a memory trace, with predictions being influenced both by prediction errors and the memory trace. The results simulated with the model can account for several MMN effects. Nevertheless, it is opaque in terms of its description of Bayesian learning, since the actual computations carried out by the neuronal model are not very easily accessible. Also, Wacongne et al. did not compare their model to other, possibly simpler neuronal networks.

Lieder et al. (2013a) made a first broad attempt at comparing many different hypotheses of MMN formation available from the literature on a single-trial basis. In their model comparison, simple models of change detection and adaptation were tested against nine variants of FEP-based models relying on an SBL with Gaussian distributions. Among FEP-based models a single precision-weighted prediction error model explained the MMN amplitude variation best. However, model-adjustment theories fared better as a model-family. Thus, whereas these results clearly speak in favor of Bayesian learning underlying surprise in sequence perception, the exact computational model to best explain the MMN has not been discovered yet. Furthermore, even though the model comparison of Lieder et al. was set up very broadly, they did not compare different forms of surprise (their "novelty-detection models" only contained expressions derived from PS), and their underlying SBL model was the same for all tested FEP-based models with no competing Bayesian model.

Connected to their data-based model comparison study, Lieder et al. (2013b) also published a neurocomputational model of the MMN, designed to predict the whole time course of the auditory MMN based on neuroanatomical data and prediction error computations. To the best of my knowledge, however, there has not been a replication of either the model by Lieder et al., nor by Wacongne et al., or a test against other possible models since.

In the visual domain, Stefanics et al. (2018) have recently conducted a similar trial-wise analysis of the MMN as generated by precision-weighted prediction error based on a Gaussian SBL model (using the hierarchical Gaussian filter scheme by Mathys et al., 2011, 2014). Even though this PC model explained the visual MMN effect well in their data, no comparison to other SBL models or surprise functions was attempted in this study.

The only single-trial analysis for somatosensory stimulation was conducted by Ostwald et al. (2012). Strictly speaking, their study does not qualify as testing a classical MMN, since participants paid attention to the stimuli to detect changes in the median-nerve stimulation amplitude, and the MMN is defined as purely preattentive to avoid confounding effects of top-down attention on the ERP-amplitude (Näätänen, 1992). Nonetheless, their data exhibit an MMN-like effect on the averaged ERP amplitude that was best explained by BS in an SBL model that employed a Beta distribution over stimulus probabilities and exponential forgetting of past events (see Section 2.3 for a detailed documentation). Although they compared their model to a simple change detection model, again, no other SBL model or surprise function were tested on the same data. Hence, while there is considerable evidence for some form of Bayesian learning and PC underlying the MMN, the exact computational dynamics of it remain elusive.

The association of the P300 with cognitive states was first described more than a decade before the discovery of the MMN (Sutton et al., 1965). The P300 signifies a positive deflection $300 - 600$ ms after an unexpected or rare target stimulus over centro-parietal electrodes, with its amplitude being inversely correlated with the stimulus probability (Duncan-Johnson and Donchin, 1977). Although the potential was characterized in terms of *uncertainty* (Sutton et al., 1965) and *surprise* (Duncan-Johnson and Donchin, 1977), Kopp (2007) was the first to cast the P300 in a Bayesian framework. He assumes that the P300 amplitude (A^{P300}) should vary as a function of the difference of prior and posterior, i.e.,

$$A_t^{P300} = f(p(s_t|o_t) - p(s_t)). \tag{1.12}$$

In other words, Kopp proposes a surprise function that quantifies surprise as the difference between

prior and posterior which should be measurable as P300.

Mars et al. (2008) have put a version of his theory to the test in a visual detection task with different item probabilities, i.e. a more or less predictable experimental environment. In a single-trial analysis, they tested a hierarchical Gaussian SBL model with PS and BS, as well as a traditional categorical model that parametrically coded the stimulus identity, and could show that PS of the SBL model was substantially better in explaining P300 variations than all other models.

More recently, Meyniel et al. (2016) formulated an SBL model based on Beta distributions for transition probabilities (TP) between two different items occurring in a sequence. In their model, they introduced exponential down-weighting of past events (as done before in Ostwald et al., 2012) and compared it with more simple models of item- (c.f., Ostwald et al., 2012) or alternation-probabilities as well as models without forgetting. PS of the TP model with implemented forgetting (half-life of $\approx$ 11 items) could explain P300 data from previous studies such as Squires et al. (1976) exceptionally well. Furthermore, their TP model can account for behavioral effects such as reaction times in detection experiments. Summed up, the study of Meyniel et al. provides a convincing account for Bayesian learning of TPs as well as an exhaustive SBL model formulation and comparison. However, the authors only considered PS while neglecting other possible surprise measures such as BS.

As a more cognitively demanding experiment, Kolossa et al. (2015) conducted an urn-ball-task, in which participants were shown a sample of four balls of two possible colors sequentially, and had to guess which urn the sample was drawn from (i.e., which composition of colored balls the respective urn had), while measuring EEG. Here, urns were defined as hidden states (because the urn from which the balls were drawn was hidden from the participants), whereas balls were called observable states. To model the amount of surprise at a given trial, they used point estimates in an SBL scheme with PS and BS. In their single-trial analysis, they showed that BS about the hidden states (urns) explained the P3a, an early sub-component of the P300, while the later P3b was best explained by PS about the observable ball color. This work is seminal in that it directly connects computational surprise measures to higher order cognitive processes such as inferential decision making, as well as to the subcomponents of the P300.

In conclusion, evidence for Bayesian learning and PC has been provided for both the MMN and the P300. While we find that computational models for the MMN are more often cast in a PC framework, SBL models seem to be particularly valuable in explaining P300 effects. Future research

should concentrate on exhaustive model-comparisons with both electrophysiological potentials, since their different roles in perception and perceptual decision making are not yet fully understood. For example, Ostwald et al. (2012) recorded an MMN-like potential as well as a P300 in their study, however, they only considered BS of item probability SBL models. According to Meyniel et al., however, a TP model could possibly explain the P300 more accurately. In Chapter 2, I will contribute to this research of modeling surprise in two-item-sequences in setting up a larger[2] model space for SBL models in combination with three different surprise functions (i.e., PS, BS, and CS). Chapter 3 then provides a test of these computational surprise models with the somatosensory MMN (sMMN) as a proxy.

1.3 The Somatosensory System

For a better grasp on information processing in the somatosensory system, which is studied in Chapter 3, here I will briefly review its cortical architecture and information processing characteristics. The somatosensory system is possibly the most diverse perceptual system of the human body. Not only does it contribute to the creation of a conscious percept of events happening on the surface of the skin, but it also yields information about where the body and its parts are located in space, and brings feedback control to the motor system for coordination of movements (Hendry and Hsiao, 2008, p. 581). Hence, we can perhaps expect the cortical structure of the somatosensory system to be more heterogeneous than that of the auditory and visual perceptual domains.

1.3.1 Architecture of Somatosensory Cortex

The primary somatosensory cortex (SI) is located at the postcentral sulcus of the brain (c.f. Figure 1.1a) and receives its input from the ventral posterolateral nucleus of the thalamus. SI consists of the subdivisions Brodman area (BA) 3 (which is again subdivided into BA3a and BA3b), BA1, and BA2, with increasing cortical hierarchy levels and complexity of information processing (Felleman and Van Essen, 1991, p. 36; see Figure 1.1b).

While BA3a and BA2 mainly receive proprioceptive input from muscles and joints, BA3b and BA1 receive their input from receptors in the skin (Kaas, 1993). All four areas are highly intercon-

[2] As we shall see in Chapter 2, the model space for SBL models is infinitely large and can only be considered in segments for specific forgetting- and volatility parameters

nected, creating the possibility of both serial and parallel processing for higher order transformation of sensory information (Gardner and Kandel, 2000).

Somatosensory information is then passed on to secondary somatosensory cortex (SII) situated ventrally from SI at the upper bank of the Sylvian fissure (Eickhoff et al., 2007). Both SI and SII have a full-body representation (Blankenburg et al., 2003; Eickhoff et al., 2007, respectively), but while SI represents simple feature orientation (Bensmaia et al., 2008), SII has furthermore been implicated in the perception of light touch and pain (Eickhoff et al., 2006), tactile attention (Burton and Sinclair, 2000), and awareness of tactile stimulation (Auksztulewicz et al., 2012).

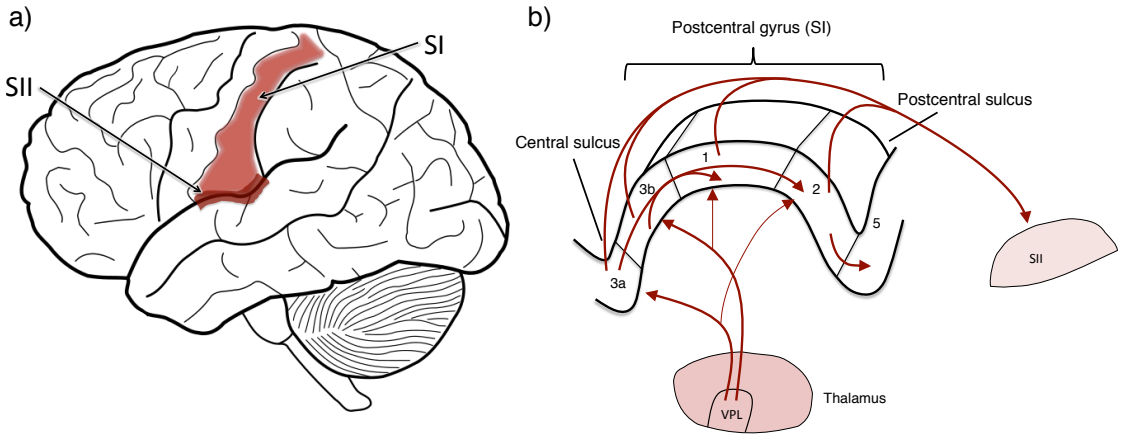

Figure 1.1: Somatosensory cortex of the human brain. a) Primary (SI) and secondary (SII) somatosensory cortices on a whole brain surface. SII extends to the ceiling of the lateral sulcus which is largely not visible from the surface shown here. b) Closer view of SI in a radial cut of the postcentral gyrus, including bottom-up information-processing pathways (red arrows) from ventral posterolateral nucleus (VPL) and to SII. Adopted from Gardner and Kandel (2000).

1.3.2 Perception of Somatosensory Stimuli

An effective way to stimulate the somatosensory cortex is to electrically stimulate the median or ulnar nerves, which innervate thumb and digit, or third, fourth and fifth fingers, respectively. Above-threshold stimulation usually leads to a somatosensory evoked potential (SEP), which is measurable with EEG over contralateral somatosensory cortices (Allison et al., 1991). Stimuli of detectable strength usually lead to a desynchronization of the cortical μ-rhythm around 10 Hz (Schubert et al., 2009).

When stimuli far below the subjective perception threshold are given, the SEP will show a cortical evoked potential (P1) at 60 ms (Nierhaus et al., 2015). However, Nierhaus et al. could not

measure any later potentials, speaking for an end of the feedforward sweep that would otherwise have rendered a conscious percept of the stimulus.

Indeed, using dynamic-causal modeling of their EEG data, Auksztulewicz et al. (2012) could show that consciously perceiving peri-threshold stimuli critically involves feedback from SII. Seen in the PC-framework, the recurrent processing from SII to SI could consist of a "postdiction" of stimulus activity, reassuring that the input was real. In that way, a backpropagation of a prediction (or postdiction) could be a necessity for a conscious percept in line with the hypothesis by Lamme and Roelfsema (2000); Lamme (2006), stating that feedback from secondary to primary sensory cortices is a necessary condition for the experience of a conscious percept.

A recent study measuring single-cell activity from thalamus and SI in non-human primates during vibrotactile stimulation in a detection task could show a task-set dependency of talamocortical feedforward connections (Campo et al., 2018). When the same stimuli from the detection task were administered in a passive condition without a response-requirement from the animal, feedforward interactions between thalamus and SI neurons of corresponding receptive fields were significantly reduced.

An influential study of the perception of somatosensory stimuli embedded in a sequence comes from Ostwald et al. (2012). In their study, the participants consecutively received median nerve stimulation of two distinguishable intensities and were asked to count the number of times the stimulus intensity alternated (i.e., changed from high to low or low to high intensity) during a sequence. SEP results revealed effects of alternation as early as 37 ms, as well as in the form of an sMMN-like component around 100-250 ms, and a clear P300 for alternated (i.e., deviant) stimuli. Moreover, source reconstruction revealed a network of contralateral SI, bilateral SII as well as left inferior frontal gyrus (IFG) and cingulate cortex to be the source of this effect. In addition, the authors conducted a single-trial analysis on the source level using an SBL model with Beta distributions for the stimulus frequency in combination with BS and several parameters for exponential forgetting. Here, they could show that BS in their model with implemented forgetting could explain alternation effects in contralateral SII during an early MMN time window, right IFG in a later MMN time window, and cingulate cortex for the P300 effect. In connection to their work, in Chapter 3, I will apply a broad array of SBL models combined with three surprise functions to EEG data from an sMMN paradigm.

1.4 General Aim of this Thesis

As I will show in this thesis, identifying any Bayesian updating model without a corresponding surprise function, as well as any surprise function without specification of an underlying model this function is based on, are futile. It is only in combination that these computational models of perceptual learning become meaningful, testable hypotheses.

In the following Chapter 2, I specify a broad array of Bayesian models for belief-updating or perceptual inference, point out their free parameters and derive distinct functions for predictive, Bayesian, and confidence-corrected surprise. I build regressors for each of these combinations of SBL models and surprise functions by pairing them with input sequences and explain their time courses. In Chapter 3, I put these models to the test in a somatosensory MMN paradigm. Distinctly, I applied two different amplitudes of electrical median nerve stimulation according to a Markov-chain roving-like paradigm and measured EEG responses. To my knowledge, this is the first test of a pure (i.e., attention-free) MMN in the somatosensory domain using a roving paradigm. Our results show a small mismatch effect, while the single-trial analysis of MMN-amplitude does not favor any computational model or surprise function reliably over the others. However, we find a trend for confidence-corrected surprise in Beta-Bernoulli (BB) models for stimulus probability (SP) and transitional probabilities (TP). While I discuss theoretical and empirical findings separately in Sections 2.5 and 3.4, respectively, I examine them on a more global scale in Chapter 4.

Chapter 2

Theoretical Modeling: Probabilistic Computational Models for Neural Surprise Signals

This chapter reviews several possible sequential Bayesian learning (SBL) schemes and surprise functions for investigating neural surprise signals. It starts with a brief introduction into the current literature on SBL modeling efforts and the necessity for exact model specifications. After describing a general process of defining and generating a two-item sequence probabilistically, an example paradigm according to a roving-like Markov-chain with two hidden states (i.e., a fast and a slow regime) is introduced.

Then, two major classes of SBL models, namely Beta-Bernoulli (BB, assuming static parameters) and Gaussian random walk (GRW, assuming dynamic parameters), are written as special cases of a general probabilistic generative model and formulated to learn different sequence features (stimulus, alternation, or transitional probabilities). From all models, specific PS, BS, and CS functions are derived that are applicable to any two-item sequence. As an illustration of the properties of the SBL model, its free model parameters (exponential forgetting and volatility) as well as surprise function in relation to the sequence, surprise regressors relating to an example sequence are shown and their inter-correlations reviewed.

By exemplifying the multitude of the resulting surprise trajectories, the chapter emphasizes

that neural surprise processes are better understood through determining exact model parameters and surprise functions than by classifying them as evidence for Bayesian learning.

2.1 Introduction

The Bayesian brain hypothesis (BBH) supposes that the human brain conducts Bayesian inference over prior and posterior probability distributions of sensory causes (Knill and Pouget, 2004). It stresses the notions that our brain operates with probability distributions rather than fixed estimates, and that it takes prior experience into account when faced with new information.

Neural responses to sequence perception provide researchers with a useful foundation to put Bayesian inference in the brain to the test. At each stage t of the sequence, the brain can compute Bayesian inference by using the posterior obtained from the previous stage $t-1$ as a prior for the current one. Neural responses signifying surprise or expectation-violation then serve as a proxy for the experimenter to investigate expectations or predictions made by the nervous system, thereby opening the possibility to examine Bayesian inference itself.

In EEG research, a multitude of Bayesian and non-Bayesian models have been applied to investigate two well-known expectation-violation or surprise potentials, the mismatch negativity (MMN; Näätänen et al., 2011) and the P300 (Polich, 2007). Importantly, the size of MMN and P300 amplitudes increase with stronger prediction violation.

Knowing the stimulus sequence and using the principle of sequential Bayesian learning (SBL), one can build a model predicting the relative amplitude size of an EEG surprise-component on a single-trial basis. Models for surprise hinge on several momentous decisions, such as the relevant statistical feature of the sequence (c.f., Meyniel et al., 2016), the class of probability distributions, and the kind of surprise function to extract the amount of surprise from the current distribution at each trial t of the sequence. To our knowledge, only few studies have conducted such analyses for Bayesian models with MMN (Ostwald et al., 2012; Lieder et al., 2013a; Stefanics et al., 2018) and P300 (Seer et al., 2016; Kolossa et al., 2015; Mars et al., 2008).

In these studies, several versions of sequential Bayesian learner models and surprise functions were compared. Importantly, Meyniel et al. (2016) have reanalyzed data from P300 components and emphasized the explanatory power and simplicity of Bayesian learner models in predicting EEG amplitudes. In fact, different types of Bayesian learner models yielded contrasting model evidence regarding their explanatory power of neurophysiological signals. Since some Bayesian learner

models and surprise functions are better than others in explaining the amplitude of expectation-violation components, it is sensible to look at different surprise trajectories in more detail and assess differences relating to model parameters.

In spite of converging evidence for Bayesian learning playing a major role in sequence perception, it is not yet established what *kind* of Bayesian learning and surprise function human surprise responses follow. This chapter makes three key contributions to the ongoing research on Bayesian sequence learning:

I We provide a systematic taxonomy of sequential Bayesian learner models for two-item sequences. We structure these models along three degrees of freedom for a Bayesian learner model: probability distribution family, sequence feature, and volatility. We then formulate three functions that quantify surprise from these models: predictive, Bayesian, and confidence-corrected surprise.

II To assess the sensitivity of different Bayesian learner models, we establish an example sequence from known probabilities and compute the respective surprise regressors yielded by the various models. This allows us to assess the impact of model specifications on surprise regressors.

III Based on our evaluation, we discuss the differences and similarities found across the studied model space. We find that in spite of all models having Bayesian learning as their updating principle, they exemplify contrary surprise trajectories.

The findings suggest that there is little to gain from knowing that a neural surprise response could be modeled by a kind of Bayesian learner per se. Instead, it is likely far more effective to look at the actual model parameters and the resulting surprise trajectory that explain neural surprise responses.

2.2 Experimental Stimulation Paradigms

A stimulus paradigm that allows for tracking neural surprise responses establishes a certain rule or expectation, the deviation of which is then the cause for surprise. For paradigms eliciting an MMN, researchers traditionally used *oddball* stimulation, wherein frequent standard stimuli are interspersed with rare deviants (see Figure 2.1a). A deviant stimulus then leads to a more negative course of the event related potential (ERP) 100-250 ms post stimulus presentation than a standard.

Such effects were discovered first in the auditory (Näätänen et al., 1978), later in the visual (Tales et al., 1999) and somatosensory domain (Kekoni et al., 1997) and do not require the participant to attend to the presented stimuli.

In oddball paradigms, physical stimulus properties and frequency of appearance are inextricably tied to their role as a standard or deviant. This is why more recent MMN-research focused on *roving paradigms* (Cowan et al., 1993). Here, any stimulus can be a deviant or standard depending on the position in the sequence (see Figure 2.1b). This isolates effects specific to physical stimulus properties and stimulus frequency from the MMN. For an overview over other sequential stimulus paradigms for MMN or predictability-effects and control-conditions, see the recent review of Heilbron and Chait (2017).

Both oddball and roving paradigms can consist of more than two different stimuli, although here, we focus on the most simple form with two different stimuli. The key difference between oddball and roving paradigms can be framed with the probability with which a given stimulus can occur, i.e., the stimulus probability (SP). While deviant stimuli in an oddball paradigm mostly have an SP of less than 0.2, roving paradigms usually assign equal probabilities to each stimulus, meaning 0.5 for the simplest case with two different stimuli. However, the probability for an alternation to occur is considerably below 0.5 in a roving paradigm, to allow for longer trains of equal stimuli.

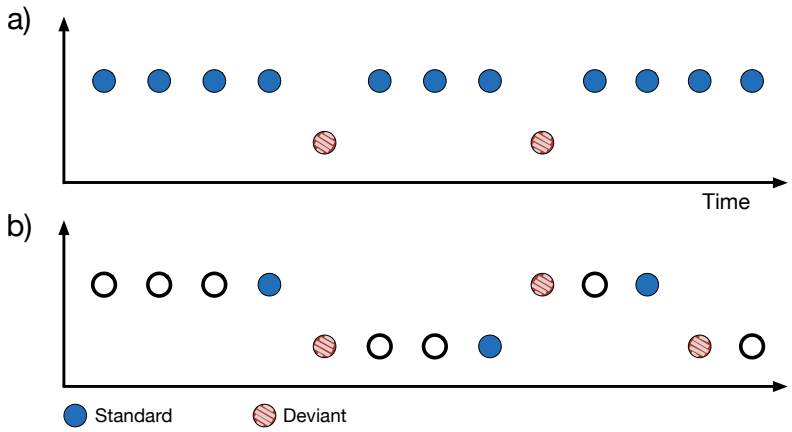

Figure 2.1: MMN Paradigm Scheme. Two different stimuli vary along one or more feature dimensions represented by y-axis. a) Oddball paradigm, many standard stimuli interspersed with few deviants. b) Roving paradigm, last stimulus of a train of equal stimuli is standard, first stimulus after a feature switch is deviant.

For the purpose of this chapter, which aims to comprehensively show the relationship between input sequence, SBL model, and surprise functions, we concentrate on a roving paradigm with two

different stimuli generated by a hierarchical first-order Markov-chain.

A first-order Markov chain is able to capture the form of an oddball as well as a roving paradigm, as it allows for decoupling of SPs from transition probabilities (TPs; Mill et al., 2011). Here, we offer an example of a roving paradigm that can switch between two different first-order Markov-chains with inverted TP matrices. This enables us to analyze the sensitivity of SBL models and surprise functions to changes in first-order sequence properties.

It is important to note that, while we formally describe the generative process for a sequence in Sections 2.2.1 and 2.2.2, the SBL models in Section 2.3 are framed from an observer's point of view. Specifically, the *generative process* is described from the experimenter's perspective, as it relies on the definition of TPs for stimuli o_t in the sequence, formally described as $p(o_t|o_{t-1})$. From these predefined probabilities, a sequence of stimuli $o_{1:T}$ can be sampled and observed in an experiment by a participant. The computational SBL models in Section 2.3 then capture possible ways in which an observer could infer hidden states s_t from observations $o_{1:t}$. These parameters s_t do not necessarily have to relate to the *true* generative process described below.

2.2.1 Stimulus Sequence Generation

We define a stimulus sequence based on two stimuli differing in any feature (e.g. loudness, luminosity, spatial orientation etc.). Those stimuli are denoted here by the possible outcomes 0 and 1, inducing an outcome space $O \coloneqq \{0, 1\}$.

Formally, a stimulus sequence of length $T \in N$ trials derived from a first-order Markov-chain may be expressed as the factorization of the joint probability over the random variable set $o_{1:T} \coloneqq \{o_1, o_2, \ldots, o_T\}$ as follows

$$p\left(o_{1:T}\right) = p\left(o_1\right) \prod_{t=2}^{T} p\left(o_t|o_{t-1}\right) \tag{2.1}$$

Note that the random variable o_t models the stimulus to be observed at trial $t\,(t = 1, \ldots, T)$ and can assume the two outcome values 0 and 1 with a probability that for $t = 2, \ldots, T$ depends only on the state of the random variable o_{t-1} modeling the stimulation at the previous trial. A stationary first order Markov-chain of the form of Equation (2.1) is defined by means of an *initial distribution* π over states and a *transition probability matrix* A for all possible transitions. For the case of two

possible outcomes, these parameters correspond to the vector

$$\pi \coloneqq (p(o_1 = i))_{0 \leq i \leq 1} = \begin{pmatrix} p(o_1 = 0) \\ p(o_1 = 1) \end{pmatrix} \in R^2 \tag{2.2}$$

with

$$0 \leq p(o_1 = i) \leq 1 \text{ and } \sum_{i=0}^{1} p(o_1 = i) = 1 \tag{2.3}$$

and the matrix

$$\begin{aligned} A &\coloneqq (p(o_t = i | o_{t-1} = j))_{0 \leq i,j \leq 1} \\ &= \begin{pmatrix} p(o_t = 0 | o_{t-1} = 0) & p(o_t = 1 | o_{t-1} = 0) \\ p(o_t = 0 | o_{t-1} = 1) & p(o_t = 1 | o_{t-1} = 1) \end{pmatrix} \in \mathbb{R}^{2 \times 2} \end{aligned} \tag{2.4}$$

with

$$\begin{aligned} 0 \leq p(o_t = i | o_{t-1} = j) \leq 1 \, (0 \leq i, j \leq 1) \text{ and} \\ \sum_{i=0}^{1} p(o_t = i | o_{t-1} = j) = 1 \quad (0 \leq j \leq 1) \end{aligned} \tag{2.5}$$

or "row-wise addition to 1" .

2.2.2 Example Paradigm

As an example paradigm, we use a hierarchical Markov-chain as depicted in Figure 2.2 which can take on two different outcomes under two different *regimes* of slow or fast stimulus changes. In this paradigm, a first level determines the regime, while the second level sets the observable outcome. The outcomes correspond to two different kinds of stimulation in an MMN-paradigm, while the regime-states remain hidden and can only be inferred through the frequency of transitions. Because the whole paradigm should follow the form of a roving paradigm, equal initial distributions for both chains as well as symmetrical transition probability matrices are chosen. In doing so, the overall SP stays 0.5 throughout the sequence. Figure 2.2 depicts TPs from all to all possible outcomes of the Markov-chain. Notably, by defining two different regimes, we introduce distinct hidden states that depend on TPs. Thus, an observer who wishes to make inferences on the regime she is currently observing needs to track transitions between the two consecutive stimuli.

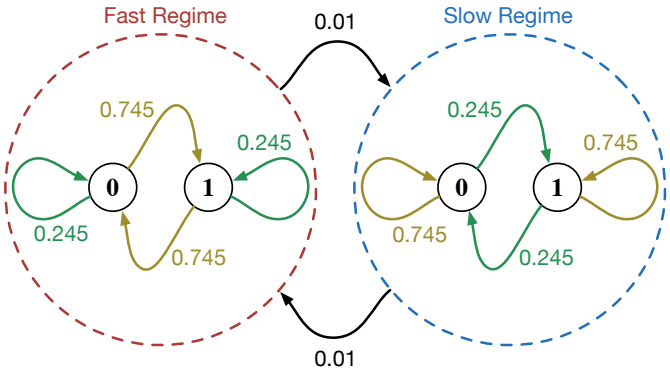

Figure 2.2: Markov Model Schema for the Markov-chain roving-like paradigm. For each outcome, transition probabilities are defined by the regime. In addition, at each stage of the sequence there is a probability of 0.01 for a regime change. In the fast regime, a change in outcome is more likely than a repetition, while in the slow regime, the opposite holds true.

The first level TP matrix determining a regime state is initiated by

$$\pi_1 \coloneqq \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \text{ with } A_1 \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{2.6}$$

guaranteeing a regime change whenever this level is invoked. On the second level, probabilities in the first two rows and columns set the outcomes, while the third column invokes the first level TP matrix A_1, which in turn directs switches between second-level matrices $A_{2.1}$ and $A_{2.2}$:

$$\begin{aligned} \pi_2 &\coloneqq \begin{pmatrix} 0.5 & 0.5 & 0 \end{pmatrix} \text{ with} \\ A_{2.1} &\coloneqq \begin{pmatrix} 0.745 & 0.245 & 0.01 \\ 0.245 & 0.745 & 0.01 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } A_{2.2} \coloneqq \begin{pmatrix} 0.245 & 0.745 & 0.01 \\ 0.745 & 0.245 & 0.01 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned} \tag{2.7}$$

Because the parameters $\pi_1 A_{2.1}$ induce sequences with relatively long subsequences of identical outcomes, this first-order Markov-chain regime is referred to as the *slow change* regime, while the parameters $\pi_2 A_{2.2}$ induce a much more frequent switching between stimuli, thusly called the *fast change* regime. Figure 2.3 depicts example realizations for $T = 200$ from both parameter regimes.

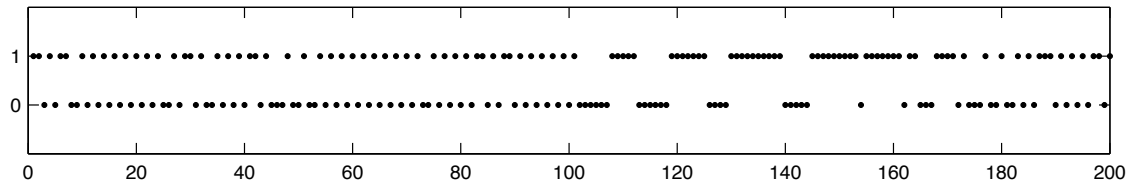

Figure 2.3: Example sequence.

2.3 Computational Models and Surprise Functions

In this section, we describe five classes of computational models for a Bayesian learner who, by observing o_t, makes inferences about possible hidden states s_t. In more detail, we show how these model classes can be understood as special cases of a general probabilistic model structure, document their respective predictive and filter distribution, and the resulting predictive, Bayesian, and confidence-corrected surprise functions. After a brief description of the general model structure and surprise functions, this section considers specific aspects of each SBL model.

Specifically, for trials $t = 0, 1, ..., T$ we consider probabilistic models over (hidden) states $s_t \in S$ and observations $o_t \in \{0, 1\}$ of the general form

$$p(s_{0:T}, o_{1:T}) = p(s_0)p(s_1|s_0)p(o_1|s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(o_t|o_{t-1}, s_t) \tag{2.8}$$

The filter distribution $p(s_t|o_{1:t})$ specifies how the state distribution s_t will be computed given observations $o_{1:t}$. The predictive distribution will follow the form $p(o_t|o_{1:t-1})$ and thus denotes the likelihood of observation o_t given all previous observations. The three surprise functions are derived from their general forms:

$$\begin{aligned} PS(o_t) &\coloneqq -\ln(p(o_t|s_t)) \\ BS(o_t) &\coloneqq KL(p(s_t)||p(s_t|o_t)) \\ CS(o_t) &\coloneqq KL\left(p(s_t)||\hat{p}(s_t|o_t)\right) \end{aligned} \tag{2.9}$$

Predictive surprise $PS(o_t)$ corresponds to the negative log probability of the current observation o_t given the posterior belief distribution over state s at trial $t-1$ (i.e., the prior for computing s_t) and can be traced back to Shannon's concept of information (1948). Bayesian surprise $BS(o_t)$ is computed using the Kullback-Leibler divergence (KLD, KL in math notation) of prior ($p(s_{t-1})$)

and posterior ($p(s_t) = p(s_{t-1}|o_t)$) distributions over s at trial t (as defined in Itti and Baldi, 2009; Baldi and Itti, 2010). Confidence-corrected surprise $CS(o_t)$ is a measure combining the two previous surprise functions linearly, together with a data-dependent constant and a model-commitment term (as defined by Faraji et al., 2018, see derivation below). It is more simply expressed as the the KLD of the prior belief distribution at trial t, $p(s_t)$, and the posterior belief distribution of a naïve observer $\hat{p}(s_t|o_t)$, i.e., an observer who has a flat prior $\hat{p}(s_t)$ and observed o_t. By computing the KLD between the current opinion (the prior $p(s_t)$) and the naïve observer distribution $\hat{p}(s_t|o_t)$, it is ensured that the omitted surprise is scaled not only by the assigned probability of the event, but also by the precision of the current belief distribution.

In the following, we reiterate the definition by Faraji et al. (2018) in our nomenclature and form of sequential Bayesian learner, and point out the link between $CS(t)$ and the other two forms of surprise,$BS(t)$ and $PS(t)$.

A flat prior $\hat{p}(s_t)$ for the model parameter s at timepoint t assumes that all possible values for s from the state space S are equally likely, thus

$$\hat{p}(s_t) = 1/|S| \tag{2.10}$$

is a state-space-dependent constant.

The probability of observation o_t under a flat prior is

$$\hat{p}(o_t) = \int_{s_t} p(o_t|s_t)\hat{p}(s_t)ds_t. \tag{2.11}$$

From Bayes rule, it follows that the posterior belief about model parameter s_t given an observation o_t under the assumption of a flat prior $\hat{p}(s_t)$ is

$$\hat{p}(s_t|o_t) = \frac{p(o_t|s_t)\hat{p}(s_t)}{\int_{s_t} p(o_t|s_t)\hat{p}(s_t)ds_t} = \frac{p(o_t|s_t)}{\int_{s_t} p(o_t|s_t)ds_t} = \frac{p(o_t|s_t)}{\|p_o\|} = \frac{p(o_t|s_t)}{O(t)} \tag{2.12}$$

where $O(t) \coloneqq \|p_o\| = \int_{s_t} p(o_t|s_t)ds_t$ is a data-dependent constant. The marginal probability of a new observation o_t under the current model $p(s_t)$ is written in the Bayesian framework as

$$p(o_t) = \int_{s_t} p(o_t|s_t)p(s_t)ds_t \tag{2.13}$$

Here, $p(s_t)$ summarizes the current probability distribution over the state space before the new data point o_t has occurred, and the integral runs over the whole possible state space. The degree of commitment $C(p(s_t))$ within $p(s_t)$ is defined as the negative entropy this distribution

$$C(p(s_t)) = -H(p(s_t)) = \int_{s_t} p(s_t)\ln(p(s_t))ds_t. \tag{2.14}$$

High commitment thus means low entropy within said distribution. Surprise is a mismatch of a

data point o_t with the current $p(s_t)$. Confidence-corrected surprise is formally defined as

$$CS(o_t) \coloneqq KL\left(p(s_t)\middle|\middle|\hat{p}(s_t|o_t)\right). \tag{2.15}$$

Highlighting its link to predictive and Bayesian surprise, it can also be expressed as

$$CS(o_t) = BS(o_t) + PS(o_t) + C(p(s_t)) + \ln O(t), \text{ where } O(t) \coloneqq \int_s p(o_t|s_t)ds_t \tag{2.16}$$

To see this, consider

$$\begin{aligned}
CS(o_t) &= KL(p(s_t)||\hat{p}(s_t|o_t)) \\
&= \int_{s_t} p(s_t)\ln(\frac{p(s_t)}{\hat{p}(s_t|o_t)})ds_t \\
&= \int_{s_t} p(s_t)\big(\ln(p(s_t)) - \ln(\hat{p}(s_t|o_t))\big)ds_t \\
&= \int_{s_t} p(s_t)\ln p(s_t)ds_t - \int_{s_t} p(s_t)\ln\hat{p}(s_t|o_t)ds_t \\
&= \int_{s_t} p(s_t)\ln p(s_t)ds_t - \int_{s_t} p(s_t)\ln\left(\frac{p(o_t|s_t)}{O(t)}\right)ds_t \\
&= Cp(s_t)) - \int_s p(s_t)\ln p(o_t|s_t)ds_t + \ln O(t) \\
&= -\int_{s_t} p(s_t)\ln\frac{p(s_t|o_t)\int_s p(o_t|s_t)p(s_t)ds}{p(s_t)}ds_t + C(p(s_t)) + \ln O(t) \\
&= -\left(\int_{s_t} p(s_t)\ln(\frac{p(s_t|o_t)}{p(s_t)})ds_t + \ln(\int_{s_t} p(o_t|s_t)p(s_t)ds_t)\right) + C(p(s_t)) + \ln O(t) \\
&= -\int_{s_t} p(s_t)\ln\left(\frac{p(s_t|o_t)}{p(s_t)}\right)ds_t - \ln(\int_{s_t} p(o_t|s_t)p(s_t)ds_t) + Cp(s_t)) + \ln O(t) \\
&= \int_{s_t} p(s_t)\ln\left(\left(\frac{p(s_t|o_t)}{p(s_t)}\right)^{-1}\right)ds_t - \ln(\int_{s_t} p(o_t|s_t)p(s_t)ds_t) + C(p(s_t)) + \ln O(t) \\
&= \int_{s_t} p(s_t)\ln\left(\frac{p(s_t)}{p(s_t|o_t)}\right)ds_t - \ln(\int_s p(o_t|s_t)p(s_t)ds_t) + C(p(s_t)) + \ln O(t) \\
&= KL(p(s_t)||p(s_t|o_t)) - \ln(\int_s p(o_t|s_t)p(s_t)ds_t) + C(p(s_t)) + \ln O(t) \\
&= BS(o_t) + PS(o_t) + C(p(s_t)) + \ln O(t)
\end{aligned} \tag{2.17}$$

where $p(o_t|s_t)$ can be substituted with $\frac{p(s_t|o_t)\int_{s_t} p(o_t|s_t)p(s_t)ds_t}{p(s_t)}$ because of the Bayesian sequential updating scheme (cf. (2.18)). Thus, $CS(o_t)$ increases with Bayesian as well as predictive surprise. □

In the following, we assume an SBL scheme, where the posterior of trial t is used as the prior of trial $t+1$:

$$p(s_{t+1}) \coloneqq p(s_t|o_t) = \frac{p(o_t|s_t)p(s_t)}{p(o_t)} = \frac{p(o_t|s_t)p(s_t)}{\int_{s_t} p(o_t|s_t)p(s_t)ds_t} \tag{2.18}$$

Table 2.1 gives an overview over all computational models presented in this paper, sorted by

probability distribution family and sequence feature.

In addition to the predictive and filter distributions as well as surprise functions for each of these models, we will denote the BB SP model in terms of a prediction-error correcting update. This is intended to provide further insight into the dynamics of BB models as seen in Section 2.4.

Table 2.1: Computational models by distribution family and sequence feature

Distribution family	Sequence feature probability		
	Stimulus	Alternation	Transition
Beta	BB SP	BB AP	BB TP
Gaussian	GRW SP	GRW AP	-

Note. BB: Beta-Bernoulli. GRW: Gaussian random walk.

2.3.1 Beta-Bernoulli Models

For an observer who makes inferences about the probabilistic structure of a Bernoulli sequence $\text{Bern}(o_t; s_t)$, the analytically optimal solution would be to model the parameter s_t with a Beta distribution. In such a Beta-Bernoulli (BB) model, parameter s_t is considered as static, with a steep learning-curve in the beginning and converging to zero later in the sequence. However, to insure learning at all stages of the sequence and mimic a decaying memory-trace of events, an exponential forgetting parameter τ can be introduced, which weights past observations according to their temporal proximity to the current observation o_t (as proposed by Ostwald et al., 2012; Harrison et al., 2011).

The roots of this model class can be traced back to Raïffa and Schlaifer (1961), where Bayesian inference is seen from a purely analytic-mathematical viewpoint. Hence, BB models are established mainly to be mathematically optimal, and not to be neuro-scientifically plausible. In the following, we review BB models estimating hidden states for SP, AP, and TP sequence features.

2.3.1.1 Beta-Bernoulli stimulus probability model

For the BB SP model, we define

$$\begin{aligned} S &\coloneqq [0,1] \\ p(s_0) &\coloneqq \text{Beta}(s_0; \alpha_0, \beta_0) \text{ with } \alpha_0 = \beta_0 = 1 \\ p(s_t|s_{t-1}) &\coloneqq \delta_{s_{t-1}}(s_t) \ \forall t = 1, \dots, T \\ p(o_t|o_{t-1}, s_t) &\coloneqq p(o_t|s_t) \coloneqq \text{Bern}(o_t; s_t) \end{aligned} \tag{2.19}$$

In other words, the state space is given by the closed interval $[0, 1]$, the prior state distribution is given by a Beta distribution with prior parameters α_0 and β_0, the state transition distribution is given by the Dirac distribution and hence $s_t = s_{t-1}$ for all t, and the emission distribution is given by a Bernoulli distribution with expectation parameter s_t. Note that in the current case, the distribution of o_t is conditionally independent of o_{t-1} given s_t. A prediction-error correcting update for the state expectation is given by

$$\mu_t = \mu_{t-1} + \frac{1}{t+1}(o_t - \mu_{t-1}), \text{ with } \mu_{t-1} = E_{Beta(s_{t-1};\alpha_{t-1},\beta_{t-1})}. \tag{2.20}$$

We show that equation (2.20) holds true for the Beta-Binomial model by postulating a theorem and proving it by induction.

Theorem. For $t = 1, \dots i, T$, let

$$\mu_t = E_{Beta(s_t;\alpha_t,\beta_t)} \tag{2.21}$$

and

$$\phi_t = \phi_{t-1} + \frac{1}{t+1}(o_t - \phi_{t-1}) \tag{2.22}$$

Set $\alpha_0 \coloneqq 1$, $\beta_0 \coloneqq 1$, and $\phi_0 \coloneqq 0.5$. Then

$$\phi_t = \mu_t \text{ for } t = 1, \dots T \tag{2.23}$$

Proof. Note that

$$o_t = 0 \Rightarrow \mu_t = \frac{\alpha_{t-1}}{\alpha_{t-1} + \beta_{t-1} + 1} \text{ and } o_t = 1 \Rightarrow \mu_t = \frac{\alpha_{t-1} + 1}{\alpha_{t-1} + \beta_{t-1} + 1} \tag{2.24}$$

Base Case. Let $t = 0$. Then

$$\mu_0 = E_{Beta(s_0;\alpha_0,\beta_0)}(s_0) = \frac{1}{1+1} = 0.5 \tag{2.25}$$

and $\phi_0 = 0.5$ by definition. Thus $\phi_0 = \mu_0$.

Induction step. Assume that $\phi_{t-1} = \frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}$. For convenience of notation, define $a \coloneqq \alpha_{t-1}$ and $b \coloneqq \beta_{t-1}$. We consider the cases $o_t = 0$ and $o_t = 1$ subsequently. Case (1). Let $o_t \coloneqq 0$. Then

$$\begin{aligned}
\phi_t &= \phi_{t-1} + \frac{1}{t+1}(o_t - \phi_{t-1}) \\
&= \frac{a}{(a+b)} + \frac{1}{a+b+1}\left(0 - \frac{a}{(a+b)}\right) \\
&= \frac{a}{(a+b)} - \frac{a}{(a+b+1)(a+b)} \\
&= \frac{a(a+b+1)-a}{(a+b+1)(a+b)} \\
&= \frac{a^2+ab+a-a}{(a+b+1)(a+b)} \\
&= \frac{a^2+ab}{(a+b+1)(a+b)} \\
&= \frac{a(a+b)}{(a+b+1)(a+b)} \\
&= \frac{a}{a+b+1} \\
&= \frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}+1} \\
&= \mu_t
\end{aligned} \tag{2.26}$$

Case (2). Let $o_t \coloneqq 1$. Then

$$\begin{aligned}
\phi_t &= \phi_{t-1} + \frac{1}{t+1}(o_t - \phi_{t-1}) \\
&= \frac{a}{(a+b)} + \frac{1}{a+b+1}\left(1 - \frac{a}{(a+b)}\right) \\
&= \frac{a}{(a+b)} + \frac{1}{a+b+1} - \frac{a}{(a+b+1)(a+b)} \\
&= \frac{a^2+ab+a+b}{(a+b+1)(a+b)} \\
&= \frac{(a+b)(a+1)}{(a+b+1)(a+b)} \\
&= \frac{a+1}{a+b+1} \\
&= \frac{\alpha_{t-1}+1}{\alpha_{t-1}+\beta_{t-1}+1} \\
&= \mu_t
\end{aligned}$$

Thus, $\phi_t = \mu_t$ in all possible cases and the theorem follows by induction. □

Based on (2.19), the filter and predictive distributions are

$$\begin{aligned} p(s_t|o_{1:t}) =& \text{Beta}(s_t; \alpha_t, \beta_t), \\ & \text{where } \alpha_t = 1 + \sum_{k=1}^{t} o_k \text{ and } \beta_t = 1 + \sum_{k=1}^{t}(1 - o_k) \\ p(o_t|o_{1:t-1}) =& \left(\frac{\alpha_{t-1}}{\alpha_{t-1} + \beta_{t-1}}\right)^{o_t} \left(1 - \frac{\alpha_{t-1}}{\alpha_{t-1} + \beta_{t-1}}\right)^{1-o_t}. \end{aligned} \tag{2.27}$$

The predictive distribution $p(o_t|o_{1:t-1})$ is derived by considering the following: For observations o_t and the distribution over possible states s_t taking on values in $[0, 1]$, we have

$$\begin{aligned} p(o_t|o_{1:t-1}) &= \int_0^1 p(o_t, s_t|o_{1:t-1}) ds_t \\ &= \int_0^1 p(o_t|s_t, o_{1:t-1}) p(s_t|o_{1:t-1}) ds_t \\ &= \int_0^1 p(o_t|s_t) p(s_t|o_{1:t-1}) ds_t, \end{aligned} \tag{2.28}$$

where the third equality follows with definition of the Beta-Bernoulli model, i.e. the conditional independence of the random variable o_t of all other random variables, given s_t. For $o_t = 1$, we thus obtain with the definition of $p(o_t|s_t)$ (cf. (2.19))

$$p(o_t = 1|o_{1:t-1}) = \int_0^1 s_t^1 (1 - s_t)^0 p(s_t|o_{1:t-1}) ds_t = E_{p(s_t|o_{1:t-1})}(s_t) \tag{2.29}$$

In words, the probability of the observation o_t to take on the value 1 at trial t given observations $o_{1:t-1}$ corresponds to the expectation E of the parameter s_t with respect to the conditional distribution $p(s_t|o_{1:t-1})$, i.e., the (Beta) posterior distribution formed by all observations up to trial $t-1$. This expectation can be expressed in terms of the parameters of the Beta distribution as in Barber (2011, p. 166)

$$E_{p(s_t|o_{1:t-1})}(s_t) = \frac{\alpha_{t-1}}{\alpha_{t-1} + \beta_{t-1}}. \tag{2.30}$$

□

The predictive, Bayesian, and confidence-corrected surprise functions evaluate to

$$\begin{aligned} PS(o_t) =& -\ln\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{o_t}\left(1-\frac{\alpha}{\alpha+\beta}\right)^{1-o_t}\right) \\ BS(o_t) =& \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) - \ln\left(\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+o_t)\Gamma(\beta+1-o_t)}\right) \\ & - o_t\Psi(\alpha) + (o_t-1)\Psi(\beta) + \Psi(\alpha+\beta) \\ CS(o_t) =& \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) - \ln(2) + (\alpha-1-o_t)\Psi(\alpha) \\ & + (\beta-2+o_t)\Psi(\beta) + (3-\alpha-\beta)\Psi(\alpha+\beta), \end{aligned} \tag{2.31}$$

where α and β are short for α_{t-1} and β_{t-1}, respectively.

Derivation of (2.31)*,* $PS(o_t)$

Because $p(o_t = 0|o_{1:t-1})$ thus equals $1 - \frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}$, for the predictive surprise in (2.31) we obtain

$$-\ln p(o_t|o_{1:t-1}) = -\ln\left(\left(\frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}\right)^{o_t}\left(1-\frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}\right)^{1-o_t}\right). \tag{2.32}$$

□

Derivation of (2.31)*,* $BS(o_t)$

Since Bayesian Surprise is defined as the Kullback-Leibler divergence of prior (i.e., posterior at $t-1$) and posterior distributions over the state s at trial t (Baldi and Itti, 2010; Itti and Baldi, 2009), we have

$$BS(o_t) = KL(p(s_t)||p(s_t|o_t)) = \int_0^1 p(s_t)\log\frac{p(s_t)}{p(s_t|o_t)}ds \tag{2.33}$$

According to the definition in (2.19), we obtain

$$BS(o_t) = \int_0^1 \text{Beta}(s_{t-1};\alpha_{t-1},\beta_{t-1})\log\frac{\text{Beta}(s_{t-1};\alpha_{t-1},\beta_{t-1})}{\text{Beta}(s_t;\alpha_t,\beta_t)}ds \tag{2.34}$$

As proven in Liu et al. (2006), the KL divergence of two Beta distributions with parameters α_{t-1},β_{t-1} and α_t,β_t, respectively, amounts to

$$\begin{aligned} \text{KL}(\text{Beta}(s_{s_t},\alpha_{t-1},\beta_{t-1})||\text{Beta}(s_{s_t},\alpha_t,\beta_t) =& \ln\left(\frac{\Gamma(\alpha_{t-1}+\beta_{t-1})}{\Gamma(\alpha_{t-1})\Gamma(\beta_{t-1})}\right) - \ln\left(\frac{\Gamma(\alpha_t+\beta_t)}{\Gamma(\alpha_t)\Gamma(\beta_t)}\right) \\ & + (\alpha_{t-1}-\alpha_t)\Psi(\alpha_{t-1}) + (\beta_{t-1}-\beta_t)\Psi(\beta_{t-1}) \\ & + (\alpha_t-\alpha_{t-1}+\beta_t-\beta_{t-1})\Psi(\alpha_{t-1}+\beta_{t-1}) \end{aligned} \tag{2.35}$$

from which, with $\alpha_t \coloneqq \alpha_{t-1} + o_t$ and $\beta_t \coloneqq \beta_{t-1} + 1 - o_t$ it follows that

$$\begin{aligned} &\mathrm{KL}(\mathrm{Beta}(s_{s_t}, \alpha_{t-1}, \beta_{t-1}) \big|\big| \mathrm{Beta}(s_{s_t}, \alpha_{t-1} + o_t, \beta_{t-1} + 1 - o_t) \\ &= \ln\left(\frac{\Gamma(\alpha_{t-1} + \beta_{t-1})}{\Gamma(\alpha_{t-1})\Gamma(\beta_{t-1})}\right) - \ln\left(\frac{\Gamma(\alpha_{t-1} + \beta_{t-1} + 1)}{\Gamma(\alpha_{t-1} + o_t)\Gamma(\beta_{t-1} + 1 - o_t)}\right) \\ &\quad -o_t\Psi(\alpha_{t-1}) + (o_t - 1)\Psi(\beta_{t-1}) + \Psi(\alpha_{t-1} + \beta_{t-1}). \end{aligned} \tag{2.36}$$

□

Derivation of (2.31)*,* $CS(o_t)$

Based on equation (2.27), the filter distribution at $t-1$ is given by

$$p(s_{t-1}|o_{1:t-1}) = \mathrm{Beta}(s_t; \alpha_{t-1}, \beta_{t-1}) \tag{2.37}$$

In the BB SP model, since a flat prior is denoted by $Beta(s_t, 1, 1)$, we can write the parameters of the updated *Beta*-distribution $\hat{\alpha_t}$ and $\hat{\beta_t}$ as $1 + o_t$ and $1 + (1 - o_t) = 2 - o_t$, respectively. The KL-divergence between those two Beta distributions then amounts to (again, according to Liu et al., 2006)

$$\begin{aligned} \mathrm{KL}(\mathrm{Beta}(s_t; \alpha_{t-1}, \beta_{t-1}) \big|\big| \mathrm{Beta}(s_t, \hat{\alpha_t}, \hat{\beta_t})) = &\ln\frac{\Gamma(\alpha_{t-1} + \beta_{t-1})}{\Gamma(\alpha_{t-1})\Gamma(\beta_{t-1})} - \ln\frac{\Gamma(\hat{\alpha_t} + \hat{\beta_t})}{\Gamma(\hat{\alpha_t})\Gamma(\hat{\beta_t})} \\ &+ (\alpha_{t-1} - \hat{\alpha}_t)\Psi(\alpha_{t-1}) + (\beta_{t-1} - \hat{\beta}_t)\Psi(\beta_{t-1}) \\ &+ (\hat{\alpha_t} - \alpha_{t-1} + \hat{\beta_t} - \beta_{t-1})\Psi(\alpha_{t-1} + \beta_{t-1}) \\ = &\ln\frac{\Gamma(\alpha_{t-1} + \beta_{t-1})}{\Gamma(\alpha_{t-1})\Gamma(\beta_{t-1})} - \ln(2) \\ &+ (\alpha_{t-1} - 1 - o_t)\Psi(\alpha_{t-1}) + (\beta_{t-1} - 2 + o_t)\Psi(\beta_{t-1}) \\ &+ (3 - \alpha_{t-1} - \beta_{t-1})\Psi(\alpha_{t-1} + \beta_{t-1}) \end{aligned} \tag{2.38}$$

To write $CS(o_t)$ as a combination of $BS(o_t)$, $PS(o_t)$, $C(p(s_t))$ and $\ln(O(t))$, we consider $C(p(s_t))$ and $\ln(O(t))$ for the Beta-Bernoulli case. Since our model space is $S = [0, 1]$, integrals over variable s will run from 0 to 1.

$$\begin{aligned} C(p(s_t)) &= \int_0^1 p(s_t)\ln(p(s_t))ds_t \\ &= \int_0^1 \mathrm{Beta}(s_t; \alpha_{t-1}, \beta_{t-1}\ln(\mathrm{Beta}(s_t; \alpha_{t-1}, \beta_{t-1}))ds_t \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha-1}(1-s)^{\beta-1} \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha-1}(1-s)^{\beta-1}\right) ds_t \\ &= \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 s^{\alpha-1}(1-s)^{\beta-1} \ln\left(s^{\alpha-1}(1-s)^{\beta-1}\right) ds_t \end{aligned} \tag{2.39}$$

For the data-dependent constant $O(t)$, we have

$$O(t) = \int_0^1 p(o_t|s_t)ds_t = \int_0^1 s_t^{o_t}(1-s_t)^{(1-o_t)}ds_t, \tag{2.40}$$

which equals 0.5 for both $o_t = 1$ as well as for $o_t = 0$. □

With an implemented exponential forgetting of a rate determined by τ, Beta distribution parameters α and β are given by

$$\alpha_t = 1 + \sum_{k=1}^{t} e^{-\tau(t-k)} o_k \text{ and } \beta_t = 1 + \sum_{k=1}^{t} e^{-\tau(t-k)}(1 - o_k). \tag{2.41}$$

Implemented forgetting for the BB SP (as well as the following AP and TP) model according to Ostwald et al. (2012) follows an exponential down-weighting of past events of the form

$$e^{-\tau(t-k)}. \tag{2.42}$$

Here, k counts from the first to current trial t, and τ specifies the rate of forgetting, with $\tau \in \mathbb{R}$ for $0 \leq \tau < 1$. No downweighting occurs by setting $\tau = 0$. This weighting is applied to the summation of events for the α- and β-parameters of the Beta distribution, yielding (2.41). □

2.3.1.2 Beta-Bernoulli alternation probability model

For the BB AP model, we define

$$\begin{aligned} S &\coloneqq [0, 1] \\ p(s_0) &\coloneqq \text{Beta}(s_0; \alpha_0, \beta_0) \text{ with } \alpha_0 = \beta_0 = 1 \\ p(s_t|s_{t-1}) &\coloneqq \delta_{s_{t-1}}(s_t) \; \forall t = 1, \ldots, T \\ p(o_t|o_{t-1}, s_t) &\coloneqq \begin{pmatrix} s_t & 1 - s_t \\ 1 - s_t & s_t \end{pmatrix} \end{aligned} \tag{2.43}$$

In this model, the state space is given by the closed interval $[0, 1]$, the prior state distribution is given by a Beta distribution with prior parameters α_0 and β_0, the state transition distribution is given by the Dirac distribution and hence $s_t = s_{t-1}$ for all t, and the emission distribution is assumed to be be specified non-parametrically in terms of the matrix indicated in (2.43). Here, the two combinations of realizations for o_t and o_{t-1} are estimated with an identical parameter s_t, namely $o_t = 0|o_{t-1} = 0$ and $o_t = 1|o_{t-1} = 1$. Conversely, the probability of $o_t = 0|o_{t-1} = 1$ and $o_t = 1|o_{t-1} = 0$ is represented by $1 - s_t$. Thus, it is constrained by the notion that any stimulus change, regardless of the stimulus property itself (and hence stimulus equality) between o_t and o_{t-1} lead to the same outcome.

Based on (2.43), the filter and predictive distributions evaluate to

$$\begin{aligned} p(s_t|o_{1:t}) &= \text{Beta}(s_t;\alpha_t,\beta_t), \text{ where} \\ &\beta_1 = \alpha_1 = 1, \\ &d_t \coloneqq [o_{k-1} = o_k] \\ &\alpha_t = 1 + \sum_{k=2}^{t} d_k \text{ and } \beta_t = 1 + \sum_{k=2}^{t}(1-d_k) \\ p(o_t|o_{1:t-1}) &= \left(\frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}\right)^{d_t}\left(1-\frac{\alpha_{t-1}}{\alpha_{t-1}+\beta_{t-1}}\right)^{(1-d_t)} \end{aligned} \tag{2.44}$$

and the predictive, Bayesian, and confidence-corrected surprise functions are analogous to (2.31) and evaluate to

$$\begin{aligned} PS(o_t) =& -\ln\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{d_t}\left(1-\frac{\alpha}{\alpha+\beta}\right)^{(1-d_t)}\right) \\ BS(o_t) =& \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) - \ln\left(\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+d_t)\Gamma(\beta+1-d_t)}\right) - d_t\Psi(\alpha) \\ &+ (d_t-1)\Psi(\beta) + \Psi(\alpha+\beta) \\ CS(o_t) =& \ln\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} - \ln(2) + (\alpha-1-d_t)\Psi(\alpha) + (\beta-2+d_t)\Psi(\beta) \\ &+ (3-\alpha-\beta)\Psi(\alpha+\beta), \end{aligned} \tag{2.45}$$

where α and β are short for α_{t-1} and β_{t-1}, respectively.

With an implemented exponential forgetting of a rate determined by τ, Beta distribution parameters α and β are given by

$$\alpha_t = 1 + \sum_{k=2}^{t} e^{-\tau(t-k)} d_k \text{ and } \beta_t = 1 + \sum_{k=2}^{t} e^{-\tau(t-k)}(1-d_k). \tag{2.46}$$

2.3.1.3 Beta-Bernoulli transition probability model

For the BB TP model, we define

$$\begin{aligned} S &\coloneqq [0,1]^2 \text{ with } s \coloneqq (s^{(0)}, s^{(1)}) \\ p(s_0^{(i)}) &\coloneqq \prod_{i=0}^{1} \text{Beta}(s_0^{(i)}; \alpha_0^{(i)}, \beta_0^{(i)}) \text{ with } \alpha_0^{(i)} = \beta_0^{(i)} = 1, i \in \{0,1\} \\ p(s_t^{(i)}|s_{t-1}^{(i)}) &\coloneqq \delta_{s_{t-1}^{(i)}}(s_t^{(i)}) \; \forall t = 1, \ldots, T \\ p(o_t|o_{t-1}, s_t) &\coloneqq \begin{pmatrix} 1 - s_t^{(0)} & s_t^{(0)} \\ 1 - s_t^{(1)} & s_t^{(1)} \end{pmatrix} \end{aligned} \tag{2.47}$$

Here, all possible transitions from $t-1$ to t are modeled separately. The state space is given by a 2-dimensional closed interval $[0,1]$, because there are 2×2 possible transitions. Thus, the state parameter s has 2 dimensions. The prior state distribution for $s_0^{(i)}$ is given by factorized Beta distributions (Strelioff et al., 2007) with prior parameters $\alpha_0^{(i)}$ and $\beta_0^{(i)}$, the state transition distribution is given by the Dirac distribution and hence $s_t^{(i)} = s_{t-1}^{(i)}$ for all t, and the emission distribution is assumed to be specified non-parametrically in terms of the matrix indicated in (2.47). We assume that $p(o_1)$ is known so that every o_{t-1} is defined.

Based on (2.47), the filter and predictive distributions evaluate to

$$\begin{aligned} p(s_t|o_{1:t}) &= \prod_{i=0}^{1} \text{Beta}(s_t^{(i)}; \alpha_t^{(i)}, \beta_t^{(i)}), \text{ where} \\ &\alpha_t^{(i)} = 1 + \sum_{k=2}^{t} o_k [o_{k-1} = i] \text{ and } \beta_t^{(i)} = 1 + \sum_{k=2}^{t} (1 - o_k)[o_{k-1} = i] \\ &\text{with } i \in \{0,1\} \\ p(o_t|o_{1:t-1}) &= \prod_{i=0}^{1} \left(\left(\frac{\alpha_{t-1}^{(i)}}{\alpha_{t-1}^{(i)} + \beta_{t-1}^{(i)}} \right)^{o_t} \left(1 - \frac{\alpha_{t-1}^{(i)}}{\alpha_{t-1}^{(i)} + \beta_{t-1}^{(i)}} \right)^{1-o_t} \right)^{[o_{t-1}=i]}, \end{aligned} \tag{2.48}$$

where, in the predictive distribution, the first factor of the product is 1 if $o_{t-1} = 1$, and thus, $\alpha_{t-1}^{(0)}$ and $\beta_{t-1}^{(0)}$ are irrelevant, while the reverse holds true for the second factor. Consequently, the predictive distribution depends on the stimulus outcome o_{t-1}.

The predictive, Bayesian, and confidence-corrected surprise functions evaluate to

$$
\begin{aligned}
PS(o_t) =& -\ln \prod_{i=0}^{1} \left(\left(\frac{\alpha^{(i)}}{\alpha^{(i)}+\beta^{(i)}} \right)^{o_t} \left(1 - \frac{\alpha^{(i)}}{\alpha^{(i)}+\beta^{(i)}} \right)^{1-o_t} \right)^{[o_{t-1}=i]} \\
BS(o_t) =& \sum_{i=0}^{1} \Bigg(\ln \left(\frac{\Gamma(\alpha^{(i)}+\beta^{(i)})}{\Gamma(\alpha^{(i)})\Gamma(\beta^{(i)})} \right) - \ln \left(\frac{\Gamma(\alpha^{(i)}+\beta^{(i)}+1)}{\Gamma(\alpha^{(i)}+o_t)\Gamma(\beta^{(i)}+1-o_t)} \right) \\
& - o_t \Psi(\alpha^{(i)}) + (o_t - 1)\Psi(\beta^{(i)}) + \Psi(\alpha^{(i)}+\beta^{(i)}) \Bigg) [o_{t-1}=i] \\
CS(o_t) =& \sum_{i=0}^{1} \Bigg(\ln \left(\frac{\Gamma(\alpha^{(i)}+\beta^{(i)})}{\Gamma(\alpha^{(i)})\Gamma(\beta^{(i)})} \right) - \ln(2) + (\alpha^{(i)} - 1 - o_t)\Psi(\alpha^{(i)}) \\
& + (\beta^{(i)} - 2 + o_t)\Psi(\beta^{(i)}) + (3 - \alpha^{(i)} - \beta^{(i)})\Psi(\alpha^{(i)}+\beta^{(i)}) \Bigg) [o_{t-1}=i],
\end{aligned}
\tag{2.49}
$$

where α and β are short for α_{t-1} and β_{t-1}, respectively, and the summation over $i \in 0, 1$ in $BS(o_t)$ and $CS(o_t)$ follows from the KLD of products. Intuitively, it can also be seen since the cases of $o_{t-1} = 0$ and $o_{t-1} = 1$ are independent.

With an implemented exponential forgetting of a rate determined by τ, Beta distribution parameters α and β are given by

$$
\begin{aligned}
\alpha_t^{(i)} =& 1 + \sum_{k=2}^{t} e^{-\tau(t-k)} o_k [o_{k-1}=i] \text{ and} \\
\beta_t^{(i)} =& 1 + \sum_{k=2}^{t} e^{-\tau(t-k)} (1-o_k) [o_{k-1}=i] \text{ with } i \in \{0,1\}.
\end{aligned}
\tag{2.50}
$$

When down-weighting past events for $\alpha^{(i)}$ and $\beta^{(i)}$ in the BB TP model, counting of k starts at 2, because summation of events always depends on o_{t-1}:

$$
\begin{aligned}
\alpha_t^{(i)} =& 1 + \sum_{k=2}^{t} e^{-\tau(t-k)} o_k [o_{k-1}=i] \text{ and} \\
\beta_t^{(i)} =& 1 + \sum_{k=2}^{t} e^{-\tau(t-k)} (1-o_k) [o_{k-1}=i] \text{ with } i \in \{0,1\}.
\end{aligned}
\tag{2.51}
$$

The same holds true for exponential down-weighting of the BB AP model, Equation (2.46).

2.3.2 Gaussian Random Walk Models

In a Gaussian random walk (GRW) model, the observer assumes parameters s_t to change dynamically according to a Gaussian normal distribution. Notably, here, the learning rate is governed by variance σ^2 of the Gaussian distribution. Because Bayesian learning of Gaussian distributions

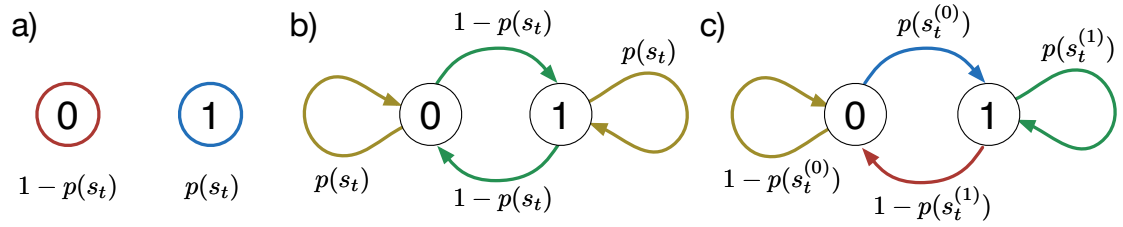

Figure 2.4: Markov model schemata for the Beta-Bernoulli models. **a)** SP model. Assuming that all states are independent of previous outcomes, only one parameter $p(s_t)$ for the probability of state $o_t = 1$ is estimated. As there are only two possible outcomes, $1 - p(s_t)$ represents the probability for state $o_t = 0$. **b)** AP model. Transitions between equal outcomes (1 to 1 or 0 to 0) are subsumed under parameter s_t, so the implicit assumption of this model is that repetitions of outcomes are equally likely. **c)** TP model. Each possible transition is regarded separately. Because TPs from one outcome at $t-1$ to the two possible outcomes at t have to sum up to 1, two probability distributions of parameters $s_t^{(0)}$ and $s_t^{(1)}$ are estimated for transitions from 0 to 1 and 1 to 1, respectively.

involves analytically intractable integrals, we describe a comprehensive numerical inversion scheme (as in Behrens et al., 2007) in the following GRW models for SP and AP sequence features. Other comparable models employ a variational-inversion approach (e.g., Mathys et al., 2011).

2.3.2.1 Gaussian random walk stimulus probability model

For the GRW SP model, we define

$$\begin{aligned} S &\coloneqq \mathbb{R} \\ p(s_0) &\coloneqq \mathrm{N}(s_0; \mu_0, \sigma_0^2) \\ p(s_t|s_{t-1}) &\coloneqq \mathrm{N}(s_t; s_{t-1}, \sigma^2)\, \forall t = 1, \ldots, T \\ p(o_t|o_{t-1}, s_t) &\coloneqq p(o_t|s_t) \coloneqq \mathrm{Bern}(o_t; l(s_t)) \end{aligned} \tag{2.52}$$

where

$$f : \mathbb{R} \to]0, 1[, x \mapsto l(x) \coloneqq (1 + e^{-x})^{-1} \tag{2.53}$$

denotes the standard logistic function. In other words, the state space is given by the real line, the prior state distribution corresponds to a univariate normal distribution with expectation parameter μ_0 and variance parameter $\sigma_0^2 > 0$, the state transition distribution corresponds to a univariate normal distribution with expectation parameter s_{t-1} (i.e., the value of the previous state realization) and variance parameter $\sigma_0^2 > 0$, and the emission distribution is given by a Bernoulli distribution with expectation parameter given by a nonlinear transformation of the state

s_t. Note that in the current case, the distribution of o_t is conditionally independent of o_{t-1} given s_t.

Numerical Inversion

Based on (2.52), the filter and predictive distributions evaluate to the recursive numerical integration of the proportionality relations

$$\begin{aligned} p(s_t|o_{1:t}) = \pi_t &= \frac{\sum_{j=1}^{n}(P_{o_t})_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{ij}} \\ p(o_t|o_{1:t-1}) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{i,j}}{\sum_{k=0}^{1}\sum_{i=1}^{n}}. \end{aligned} \tag{2.54}$$

Derivation of (2.54)

For $t = 2, 3, \ldots, t$, equation (2.54) provides a recursive expression for the filter distribution $p(s_t|o_{1:t})$ in terms of the integral with respect to s_{t-1} of the product of the preceding filter distribution $p(s_{t-1}|o_{1:t-1})$, the state transition distribution $p(s_t|s_{t-1})$, and the emission distribution $p(o_t|s_t)$. Here, we derive this expression based on the definition of the joint distribution

$$p(s_{0:t-1}, o_{1:t-1}) = p(s_0)\prod_{k=1}^{t-1} p(s_k|s_{k-1})p(o_k|s_k) \tag{2.55}$$

implicit in (2.52). We first note that by the definition of conditional probability, we have

$$p(s_{0:t-1}|o_{1:t-1}) = \frac{p(s_0)\prod_{k=1}^{t-1} p(s_k|s_{k-1})p(o_k|s_k)}{p(o_{1:t-1})} \tag{2.56}$$

and hence

$$p(s_{0:t-1}|o_{1:t-1}) \propto p(s_0)\prod_{k=1}^{t-1} p(s_k|s_{k-1})p(o_k|s_k). \tag{2.57}$$

Integration over $s_{0:t-2}$ then yields the following expression for the filter distribution at time $t-1$

$$p(s_{t-1}|o_{1:t-1}) \propto \int\cdots\int p(s_0)\prod_{k=1}^{t-1} p(s_k|s_{k-1})p(o_k|s_k)\, ds_0 \cdots ds_{t-2}. \tag{2.58}$$

Next, we note that based on (2.57) we have a similar expression for the filter distribution at time

t, which we can rewrite in terms of $p(s_{t-1}|o_{1:t-1})$ as follows

$$\begin{aligned}
p(s_t|o_{1:t}) &\propto \int \cdots \int p(s_0) \prod_{k=1}^{t} p(s_k|s_{k-1}) p(o_k|s_k)\, ds_0 \cdots ds_{t-1} \\
&= \int \cdots \int p(s_0) \left(\prod_{k=1}^{t-1} p(s_k|s_{k-1}) p(o_k|s_k) \right) p(s_t|s_{t-1}) p(o_t|s_t)\, ds_0 \cdots ds_{t-1} \\
&= \int \left(\int \cdots \int p(s_0) \left(\prod_{k=1}^{t-1} p(s_k|s_{k-1}) p(o_k|s_k) \right) p(s_t|s_{t-1}) p(o_t|s_t)\, ds_0 \cdots ds_{t-2} \right) ds_{t-1} \\
&= \int p(s_t|s_{t-1}) p(o_t|s_t) \left(\int \cdots \int p(s_0) \prod_{k=1}^{t-1} p(s_k|s_{k-1}) p(o_k|s_k)\, ds_0 \cdots ds_{t-2} \right) ds_{t-1} \\
&= \int p(s_t|s_{t-1}) p(o_t|s_t) p(s_{t-1}|o_{1:t-1}) ds_{t-1}.
\end{aligned} \tag{2.59}$$

Here, the proportionality statement is a direct consequence of (2.58) after increasing the time index from $t-1$ to t, the first equality results from splitting the product into terms from k to t-1 and t, the second equality merely introduces a more fine-grained notation of the multiple integral, the third equality uses the fact that $p(s_t|s_{t-1}|p(o_t|s_t)$ are constant with respect to the variables of integration $s_0, ..., s_{t-2}$ and the linearity property of integrals, and the last equality, and hence (2.54) results from substitution of the left hand side of (2.58).

To achieve numerical integration of the joint distributions, we set boundaries for the theoretically infinite state space, as well as define a resolution for the integration between these boundaries, making it possible to summate over every "resolution-bin". So for numerical integration, we have

$$\begin{aligned}
S &\coloneqq [s_{min} s_{max}] \\
s_{res} &\coloneqq n \\
s_i &\coloneqq s_{min} + (s_{max} - s_{min}) i/n \\
p(s_0) &\coloneqq 1/n \\
p(s_t|o_{1:t}) &\coloneqq \pi_t \in \mathbb{R}^{1 \times n}
\end{aligned} \tag{2.60}$$

where s_i is a vector of length n with entries from s_{min} to s_{max}, and π_t is the filter distribution vector of length n derived from all observations until t. Consider now the implementation of the joint probability over o_t, s_t, and s_{t-1} given all previous stimuli o_{t-1}

$$p(o_t, s_t, s_{t-1}|o_{t-1}) = Bern(o_t, l(s_t)) N(s_t; s_{t-1}, \sigma^2) p(s_t) \tag{2.61}$$

with $p(s_t)$ being the posterior $p(s_{t-1}|o_{t-1})$ used as the prior at trial t (c.f. equation (2.18)) as a tabular probability mass function. For each trial t and observation o_t, we can calculate matrix $P_{o_t} \in \mathbb{R}^{n \times n}$ with $o_t \in \{0, 1\}$ by plugging s_i into Equation (2.61) for s_t and s_{t-1}. In P_{o_t}, the first dimension i contains the distribution for s_t conditioned on the values of s_{t-1} (second dimension

j).

$$P_{o_t} = \begin{bmatrix} p_{1,1} & \dots & p_{1,j} & \dots & p_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i,1} & \dots & p_{i,j} & \dots & p_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n,1} & \dots & p_{n,j} & \dots & p_{n,n} \end{bmatrix} \tag{2.62}$$

with

$$p_{i,j} = Bern(o_t, l(s_i))N(s_i; s_j, \sigma^2)\pi_{t-1} \tag{2.63}$$

To obtain the numerically integrated filter distribution represented by vector π_t, we summate over dimension j of P_{o_t} for the current observation o_t and normalize by the sum of all elements of P_{o_t}:

$$p(s_t|o_{1:t}) = \pi_t = \frac{\sum_{j=1}^{n}(P_{o_t})_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{ij}} \tag{2.64}$$

yielding filter distribution vector π_t of length n. For the predictive distribution, we divide the summed P_{o_t} by a summation over both possible o_t:

$$p(o_t|o_{1:t-1}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{i,j}}{\sum_{k=0}^{1}\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{k,i,j}} \tag{2.65}$$

□

The predictive, Bayesian, and confidence-corrected surprise functions evaluate to

$$\begin{aligned} PS(o_t) &\coloneqq -\ln\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{i,j}}{\sum_{k=0}^{1}\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{o_t})_{k,i,j}}\right) \\ BS(o_t) &\coloneqq \sum_{i=1}^{n}(\pi_{t-1})_i \ln\left(\frac{(\pi_{t-1})_i}{(\pi_t)_i}\right) \\ CS(o_t) &\coloneqq \sum_{i=1}^{n}(\pi_{t-1})_i \ln\left(\frac{(\pi_{t-1})_i}{(\hat{\pi}_t)_i}\right) \end{aligned} \tag{2.66}$$

For confidence-corrected surprise, we calculate a posterior under a flat prior $\hat{w}(t)$, which assumes a prior of $p(s_t) = 1/n$ for all t and is derived by

$$\hat{\pi}_t = \frac{Bern(o_t, l(s_i))N(s_i; \frac{1}{n}, \sigma^2)\frac{1}{n}}{\sum_{i=1}^{n} Bern(o_t, l(s_i))N(s_i; \frac{1}{n}, \sigma^2)\frac{1}{n}}. \tag{2.67}$$

2.3.2.2 Gaussian random walk alternation probability model

For the GRW AP model, we define

$$\begin{aligned} S &\coloneqq \mathbb{R} \\ p(s_0) &\coloneqq \mathrm{N}(s_0; \mu_0, \sigma_0^2) \\ p(s_t|s_{t-1}) &\coloneqq \mathrm{N}(s_t; s_{t-1}, \sigma^2)\, \forall t = 1, \ldots, T \\ p(o_t|o_{t-1}, s_t) &\coloneqq p(o_t|s_t) \coloneqq \mathrm{Bern}(d_t; l(s_t)) \end{aligned} \tag{2.68}$$

where, again,

$$f : \mathbb{R} \to]0, 1[, x \mapsto l(x) \coloneqq (1 + e^{-x})^{-1} \tag{2.69}$$

denotes the standard logistic function. Similarly to the GRW Bernoulli model, the state space is given by the real line, the prior state distribution corresponds to a univariate normal distribution with expectation parameter μ_0 and variance parameter $\sigma_0^2 > 0$, the state transition distribution corresponds to a univariate normal distribution with expectation parameter s_{t-1} (i.e., the value of the previous state realization) and variance parameter $\sigma_0^2 > 0$, and the emission distribution is given by a Bernoulli distribution with expectation parameter given by a nonlinear transformation of the state s_t. Note that in here, the distribution of o_t is conditionally dependent on o_{t-1} given s_t, since the Bernoulli distribution depends on stimulus change vector d_t given by

$$d_t \coloneqq [o_{t-1} = o_t]. \tag{2.70}$$

Numerical Inversion

Based on (2.68), the filter and predictive distributions evaluate to the recursive numerical integration of the proportionality relations

$$\begin{aligned} p(s_t|o_{1:t}) =& \pi_t = \frac{\sum_{j=1}^{n} (P_{d_t})_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} (P_{d_t})_{ij}} \\ p(o_t|o_{1:t-1}) =& \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (P_{d_t})_{i,j}}{\sum_{k=0}^{1} \sum_{i=1}^{n} \sum_{j=1}^{n} (P_{d_t})_{k,i,j}}. \end{aligned} \tag{2.71}$$

Here, we use the same definitions for S, s_{res}, and s_i. However, the filter distribution vector π_t is

conditioned on the stimulus-change vector $d_t \coloneqq [o_{t-1} = o_t]$, and thus

$$p(s_t|d_t) \coloneqq \pi_t \in \mathbb{R}^{1\times n} \tag{2.72}$$

Elements of Matrix P_{d_t} are given by

$$p_{i,j} = Bern(d_t, l(s_i))N(s_i; s_j, \sigma^2)\pi_{t-1} \tag{2.73}$$

and analogous to (2.64), we calculate the entries of the filter distribution vector by

$$p(s_t|d_{1:t}) = \pi_t = \frac{\sum_{j=1}^{n}(P_{d_t})_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{d_t})_{ij}}, \tag{2.74}$$

consequently yielding the predictive distribution in (2.71).

The predictive, Bayesian, and confidence-corrected surprise functions evaluate to

$$\begin{aligned} PS(o_t) &\coloneqq -\ln\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{d_t})_{i,j}}{\sum_{k=0}^{1}\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{d_t})_{k,i,j}}\right) \\ BS(o_t) &\coloneqq \sum_{i=1}^{n}(\pi_{t-1})_i \ln\left(\frac{(\pi_{t-1})_i}{(\pi_t)_i}\right) \\ CS(o_t) &\coloneqq \sum_{i=1}^{n}(\pi_{t-1})_i \ln\left(\frac{(\pi_{t-1})_i}{(\hat{\pi}_t)_i}\right) \end{aligned} \tag{2.75}$$

The posterior under the flat prior for confidence-corrected surprise is given by

$$\hat{\pi}_t = \frac{Bern(d_t, l(s_i))N(s_i; \frac{1}{n}, \sigma^2)\frac{1}{n}}{\sum_{i=1}^{n} Bern(d_t, l(s_i))N(s_i; \frac{1}{n}, \sigma^2)\frac{1}{n}}. \tag{2.76}$$

MATLAB code for all models and surprise functions documented above, as well as the input sequence of the Markov-chain roving-like paradigm is provided at: `https://github.com/kathrintertel/Sequential_Bayesian_learner`. This code can also be used to create the figures shown in the following Section 2.4.

2.4 Results

Our input sequence (see Figure 2.3) was sampled from the hierarchical Markov-chain described in section 2.2.2, with a probability of $p = 0.01$ to switch between the two regimes. Here, a regime change (i.e., a switch between the two TP matrices) occurs twice in the sequence, at $t = 101$ and $t = 163$, with outcome $o_t = 1$ occurring 108 times out of $T = 200$.

Table 2.2: Surprise regressors as a combination of SBL model and surprise function

SBL model		Surprise functions		
	τ/σ^2	PS(t)	BS(t)	CS(t)
Beta-Bernoulli SP		PS0N	BS0N	CS0N
Beta-Bernoulli SP	$\tau = 0.14$	PS0F	BS0F	CS0F
Beta-Bernoulli AP		PS1N	BS1N	CS1N
Beta-Bernoulli AP	$\tau = 0.14$	PS1F	BS1F	CS1F
Beta-Bernoulli TP		PS1TN	BS1TN	CS1TN
Beta-Bernoulli TP	$\tau = 0.14$	PS1TF	BS1TF	CS1TF
GRW SP	$\sigma^2 = 2.5$	PS0H	BS0H	CS0H
GRW SP	$\sigma^2 = 0.1$	PS0L	BS0L	CS0L
GRW AP	$\sigma^2 = 2.5$	PS1H	BS1H	CS1H
GRW AP	$\sigma^2 = 0.1$	PS1L	BS1L	CS1L

Note. SBL: Sequential Bayesian learner. S/A/TP: Stimulus/alternation/transition probability. GRW: Gaussian random walk.

In this section, we review the surprise regressors built from the computational models and the input sequence. To obtain surprise regressors from the above computational models and surprise functions for a given input sequence $o_1, o_2, \cdots o_T$, we chose an exponential forgetting parameter τ for BB models and volatility parameter σ^2 for GRW models as specified in Table 2.2 for all regressors analyzed in this section. Additionally, for numerical integration of GRW models, we set the initial state space to $S = [-5\ 5]$ with a resolution of $s_{res} = 70$. In addition to absolute regressor values, we also display a normalized regressors (subtraction of mean and division by standard deviation per regressor) to provide easier comparisons between regressors.

2.4.1 Beta-Bernoulli Stimulus Probability Surprise

Figure 2.5 shows surprise regressors for the BB SP model. In this model, each state at trial t is assumed to be independent of all previous trials $1 : t-1$. Thus, the sequential Bayesian learner estimates the probability $p(s_t)$ of outcome 1 to occur (and with $1-p(s_t)$, that of outcome 0). Due to the fast-change regime within the first 100 trials, at each t the Beta distribution counts α_t and

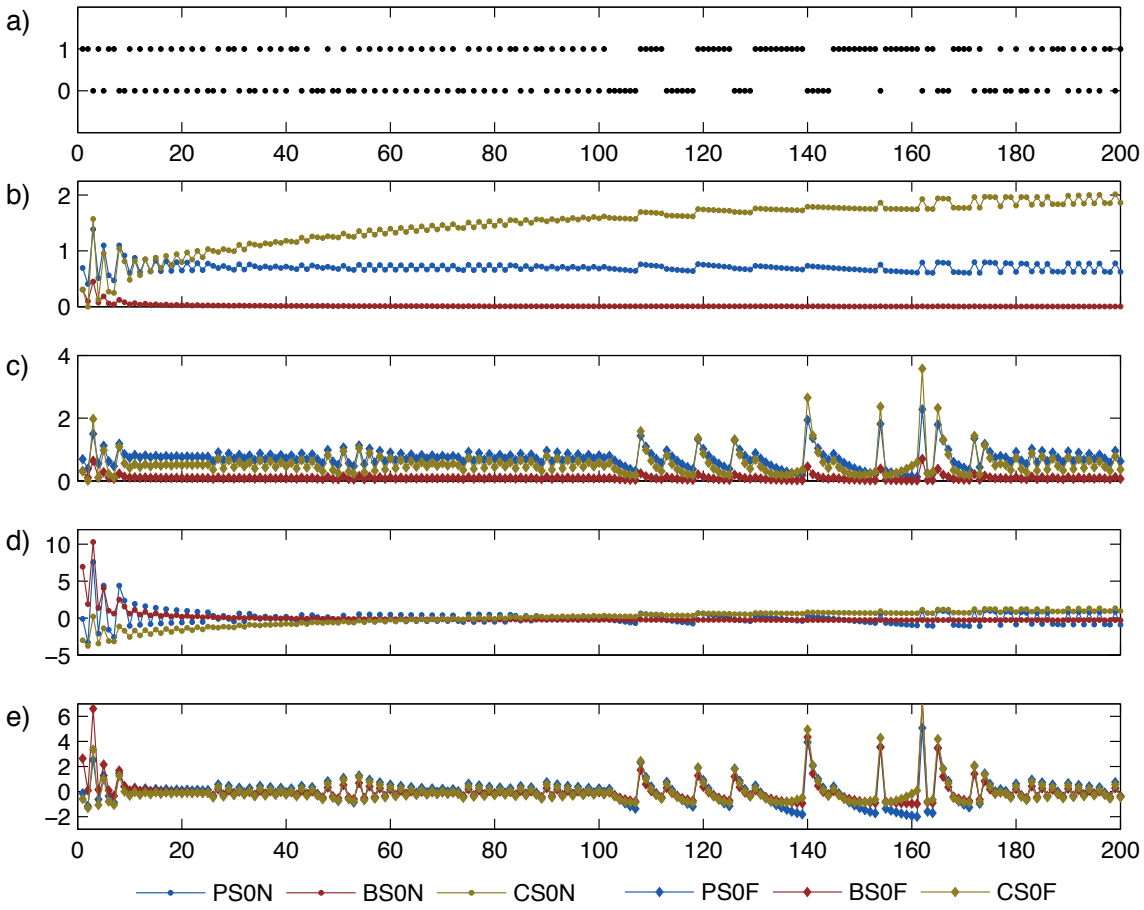

Figure 2.5: BB SP regressors for predictive, Bayesian, and confidence-corrected surprise (PS0, BS0, and CS0, respectively). *a)* Two-state input sequence used for surprise regressors. *b)* Surprise regressors without implemented forgetting (-N for no forgetting). *c)* surprise regressors with implemented forgetting (-F for implemented exponential forgetting with a half-life of $t - 5$). *d)* Normalized -N regressors. *e)* Normalized -F regressors.

β_t are close to equal, resulting in distributions centering around .5.

$PS(o_t)$ does not directly relate to the probability distributions over s_t but to its expectation value at each t, derived from the posterior distribution of $t - 1$. Accordingly, for equal Beta distribution counts, PS0N converges to to $-\ln$ of .5 (.5 being the expectation value of a *Beta* distribution with $\alpha = \beta$), as can be seen in Figure 2.5b. Implemented forgetting for $PS(o_t)$ (PS0F) leads to a similar course during the fast regime, while surprise at stimulus change in the slow regime increases much more sharply in the slow regime after a train of equal stimuli.

$BS(o_t)$ in a BB SP model is much less affected by the early frequent stimulus changes, and even less by the later slow changes. This stems from its dependence only on the shift in distribution (quantified by KLD), which rapidly converges to zero for small differences between α and β in

the first 100 trials, and proportionally changes very little once longer stimulus-trains appear (see BS0N in Figure 2.5). With implemented forgetting (BS0F), however, $BS(o_t)$ reaches larger values, especially at switches after longer trains of equal outcomes, for example at $t = 140$ or $t = 160$, while still falling short of $PS(o_t)$ here. In CS0N as well as CS0F, one can readily recognize it as the linear composition of $BS(o_t)$, $PS(o_t)$ (mathematically shown in Equation (2.17)). The slowly decreasing rise of the regressor is the result of model-commitment $C(s_t)$ (i.e., negative Entropy, c.f. Equation (2.14)) also added at each t. $C(s_t)$ increases with each Beta-update as the distribution gets sharper. In the normalized comparison between all three zero-order regressors without forgetting (Figure 2.5d), BS0N and CS0N are most strongly affected by convergence, while PS0N still emits surprise much later in the sequence. With exponential forgetting, the three surprise functions do not differ much as normalized regressors after the first few trials.

2.4.2 Beta-Bernoulli Alternation Probability Surprise

Figure 2.6 shows surprise regressors for the BB AP model under the assumption that a switch from 1 to 0 is equally likely as a switch from 0 to 1 (see Figure 2.4b). In this model, the *Beta*-distributions estimate only parameter $p(s_t)$ for any stimulus-repetition (and consequently, $1-p(s_t)$ for a stimulus-change). The "true" $p(s_t)$ from which the sequence was sampled indeed changes with each regime change, so surprise values could theoretically pick up these changes. At the same time, because there is only one $p(s_t)$ regardless of o_{t-1} in the symmetrical TP-matrices for sequence generation, this model should be able to map sequence properties better than the previous (unconstrained) one.

Regressor PS1N takes a very similar course to PS1TN (in Figure 2.7), with peaks for stimulus repetitions during the first 100 trials while not fulling to the slow regime later. With exponential downweighting (see Figure 2.6c2), $PS(o_t)$ quickly adjusts to the first regime-change in the sequence with higher surprise for the less frequent stimulus-changes, again similar to the unconstrained model.

BS1N shows the usual quick convergence to zero. However, with implemented forgetting (BS1F, see Figure 2.6c2), it adapts to regime changes similar to $PS(o_t)$.

In $CS(o_t)$, we can observe the influence of the constrained model specifically in the ramping-up due to model-confidence, which is very steady for CS1N during stimulus-changes in the fast regime (the analogous unconstrained regressor showed more zig-zag originating from two Beta

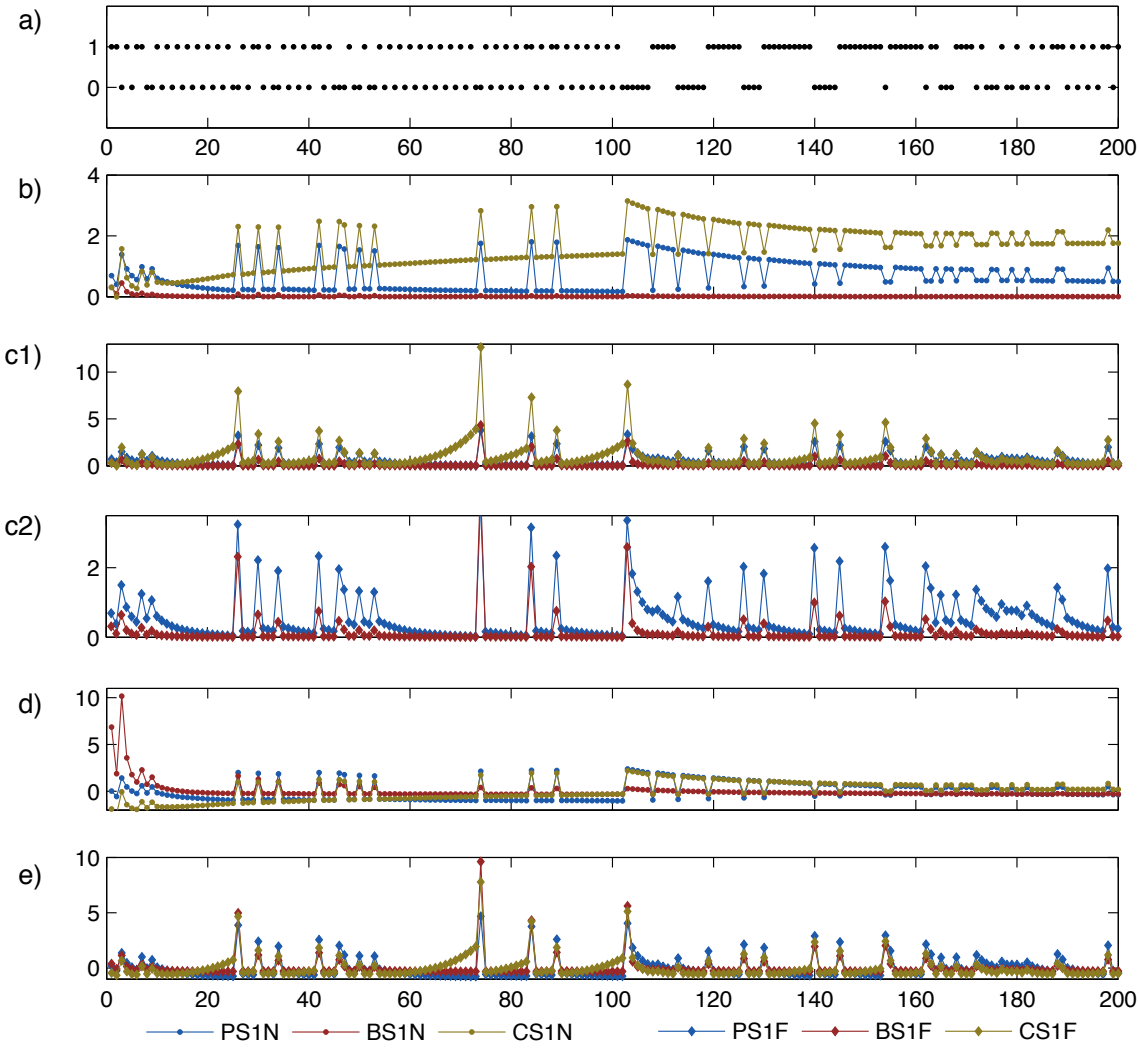

Figure 2.6: BB AP regressors for predictive, Bayesian, and confidence-corrected surprise (PS1, BS1, and CS1, respectively). *a)* Two-state input sequence used for surprise regressors. *b)* Surprise regressors without implemented forgetting (ending -N for no forgetting). *c1)* Regressor CS1F superimposed on PS- and BS1UF (-F for implemented exponential forgetting with a half-life of $t - 5$). *c2)* PS- and BS1F regressors only, with zoomed-in y-scaling. *d)* Normalized -N regressors. *e)* Normalized -F regressors.

distributions with two different $C(o_t)$). Roughly, CS1UN and CS1N take a similar course. With forgetting, $CS(o_t)$ does not rise as strongly in the constrained model, while the counterintuitive increase right before unlikely events remain. In normalized comparison, PS- and CS1N do not differ much, while BS1N converges much more quickly. The forgetting regressors paint a different picture: here, BS1F even surpasses $CS(o_t)$ and $PS(o_t)$ for unlikely events at $t = 74$ and $t = 103$ (see Figure 2.6e).

2.4.3 Beta-Bernoulli Transition Probability Surprise

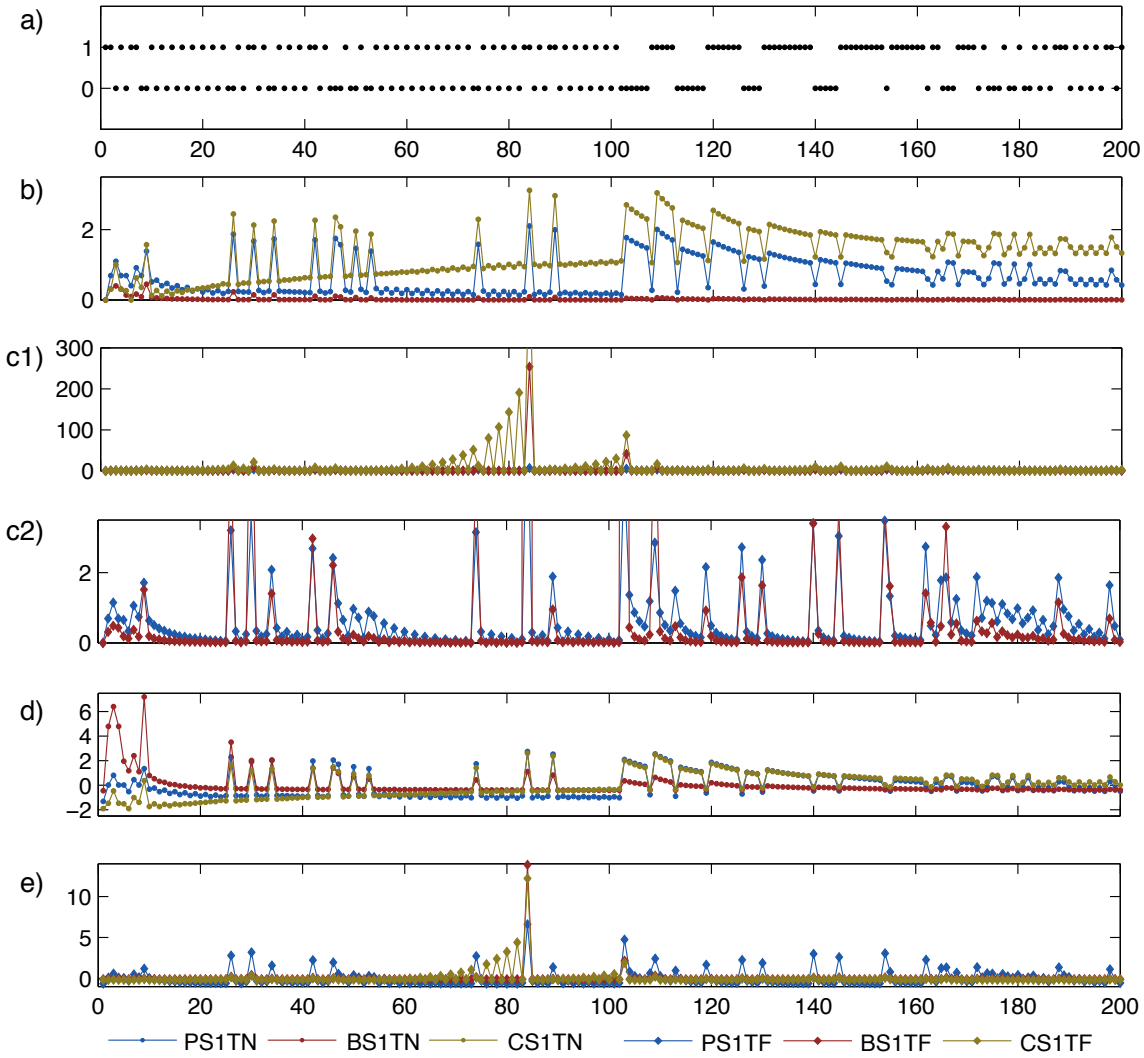

Figure 2.7: BB TP regressors for predictive, Bayesian, and confidence-corrected surprise (PS1T, BS1T, and CS1T, respectively). *a)* Two-state input sequence used for surprise regressors. *b)* Surprise regressors without implemented forgetting (ending -N for no forgetting). *c1)* Regressor CS1TF superimposed on PS- and BS1TF (-F for implemented exponential forgetting with a half-life of $t - 5$). *c2)* PS- and BS1TF regressors only, with zoomed-in y-scaling. *d)* Normalized -N regressors. *e)* Normalized -F regressors.

Figure 2.7 shows surprise regressors for the BB TP model. In this model, separate s_t-parameters are estimated for transitions from two possible outcomes $o_{t-1} = 0$ and $o_{t-1} = 1$, called $s_t^{(0)}$ and $s_t^{(1)}$, respectively (see Figure 2.4c for an overview over estimated parameters).

Regressor PS1TN becomes visibly receptive to the fast regime in the beginning, emitting surprise for equal stimuli and much less for stimulus-changes. After the regime change at $t = 101$, PS1TN slowly adjusts to a higher probability of stimulus-repetitions, with surprise decreasing after

each stimulus-repetition, but still being higher for repetitions than for changes. Before the $PS(o_t)$ regressor can map less surprise for repetitions in the slow regime, it changes again to the fast regime in $t = 163$. With exponential down-weighting, $PS(o_t)$ properly adjusts to regime changes, emitting surprise in the fast regime when states stay the same and in the slow regime when state switches occur (c.f. panel c2 in Figure 2.7).

BS1TN quickly converges to 0, similar to BS0N (note the different y-axis scaling). With exponential forgetting, $BS(o_t)$ can rise much higher, such as, e.g., at $t = 84$. In this case, a transition from 1 to 1 lies much further back than any other possible transition, resulting in a peak in $BS(o_t)$ (Figure 2.7 c2).

Analogous to the BB SP model, CS1TN again represents the its composition of PS1TN and BS1TN, together with a model-commitment part. Because in the BB TP model, transitions from two possible outcomes are modeled separately, model-commitment rises much slower than for the Beta-Bernoulli case. With exponential forgetting, the regressor for $CS(o_t)$ (see panel c1 in Figure 2.7) shows a counterintuitive course at first glance, ramping up in zig-zag to its peak at $t = 84$, where there is no such slow rise for PS- and BS1TF. This phenomenon arises from the fast build-up of model-commitment for $p(s_t^{(1)})$ that is not equally possible for the case of $p(s_t^{(0)})$ or without down-weighting of more distant outcomes.

In normalized comparison, PS1TN and CS1TN do not differ much, and while fittingly adjusting to the fast regime failing to do so for the slow regime later in the sequence. The BS1TN regressor stands out with more extreme surprise in the beginning and much lower surprise with higher t. Surprise regressors with exponential down-weighting look most nuanced and well-adjusted to regime changes for $PS(o_t)$. BS- as well as CS1TF partly present strikingly extreme outliers, while the latter also shows counterintuitive ramps to surprise.

2.4.4 Gaussian Random Walk Stimulus Probability Surprise

Figure 2.8 summarizes random GRW SP surprise regressors for a high ($\sigma^2 = 2.5$) and low ($\sigma^2 = 0.1$) variance parameter, thus accommodating either higher or lower volatility in the sequence. Note that exact regressor results here also depend on the initial state space and filtering resolution (for our example regressors, set to $S = [-5\ \ 5]$ and $s_{res} = 70$, respectively).

All surprise regressors of high volatility (PS0H, BS0H, CS0H) very soon assume almost-static values for equal counts of $o_t = 0$ and $o_t = 1$ during the fast-regime stimulus-changes (e.g., between

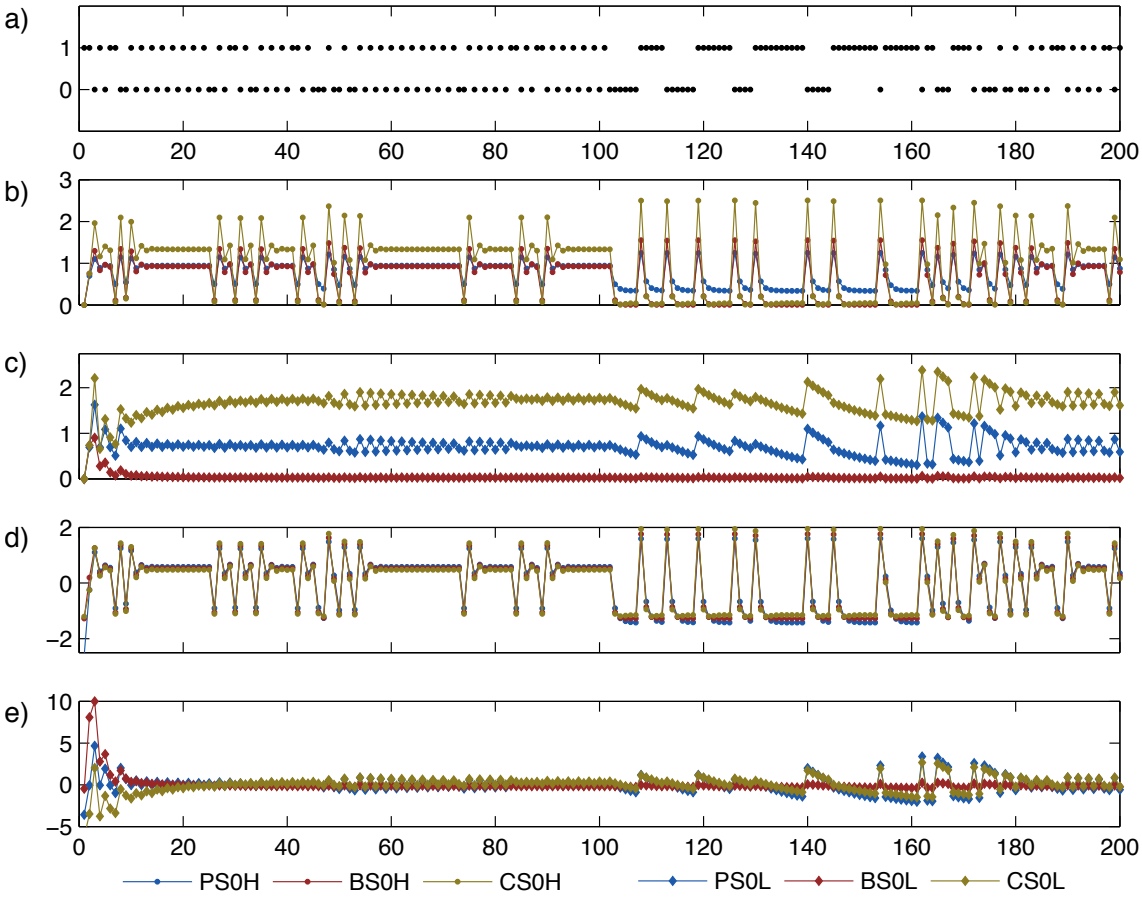

Figure 2.8: GRW SP regressors for predictive, Bayesian, and confidence-corrected surprise (PS0, BS0, and CS0, respectively). *a)* Two-state input sequence used for surprise regressors. *b)* Surprise regressors with high assumed volatility (ending -H for $\sigma^2 = 2.5$). *c)* Surprise regressors with low assumed volatility (ending -L for $\sigma^2 = 0.1$). *d)* Normalized -H regressors. *e)* Normalized -L regressors.

$t = 12$ and $t = 25$). Stimulus-repetitions then elicit a sudden decrease of surprise, followed by a steep increase with the next stimulus change. In absolute numbers, PS0H shows least extreme, CS0H the most extreme deflections. After normalization (plotted in Figure 2.8d), all high-volatility regressors look close to equal. With a low σ^2 (shown in Figure 2.8c), surprise regressors look more similar to conjugate models. Characteristically, $PS(o_t)$ zig-zags in the fast regime, while in the slow regime decreasing slowly with stimulus-repetitions and steeply increasing with stimulus-change. $BS(o_t)$ quickly converges to zero, and $CS(o_t)$ is the composition of the other two surprise functions together with model-confidence increasing with t. These distinctive courses still hold up after normalization, with differences between PS0L and CS0L become smaller with larger t due to decreasing rise of model-confidence (Figure 2.8e).

2.4.5 Gaussian Random Walk Alternation Probability Surprise

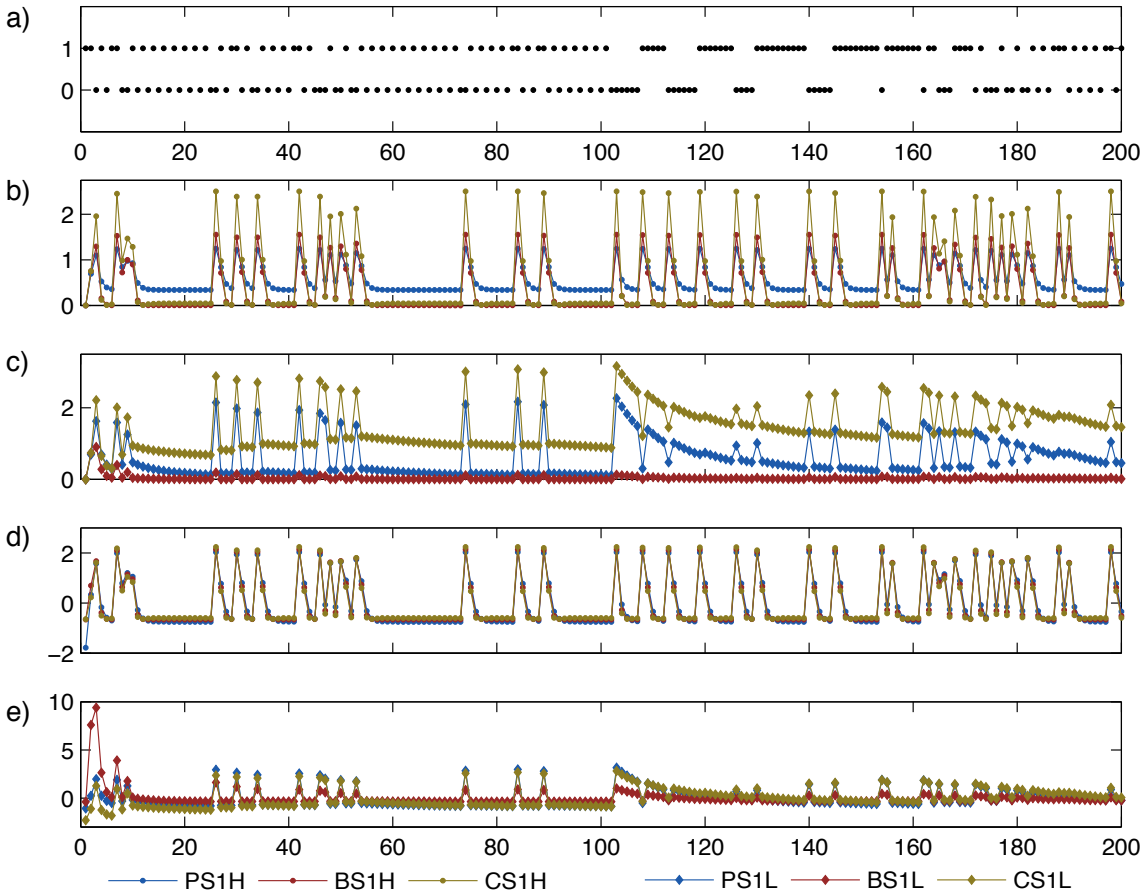

Figure 2.9: GRW AP regressors for predictive, Bayesian, and confidence-corrected surprise (PS1, BS1, and CS1, respectively). *a)* Two-state input sequence used for surprise regressors. *b)* Surprise regressors with high assumed volatility (ending -H for $\sigma^2 = 2.5$). *c)* Surprise regressors with low assumed volatility (ending -L for $\sigma^2 = 0.1$). *d)* Normalized -H regressors. *e)* Normalized -L regressors.

GRW AP surprise regressors with assumed high and low volatility are plotted in Figure 2.9. Like in the Gaussian zero-order case, the three surprise functions are very similar for high-variance models (c.f. Figure 2.9b and d). Remarkably, they adjust to regime-changes very quickly, with surprise peaks for stimulus repetitions during fast regimes and for stimulus-changes in slow regimes. This is due to the assumed high volatility of TPs, making it easy for the model to accommodate sudden changes in the TP-matrix. Normalization again highlights the congruence of surprise regressors in the high-volatility case. With a low σ^2, we observe a slower adjustment to the slow-regime for PS1L as well as CS1L, with first recognizable peaks in surprise after stimulus-changes at $t = 126$ (i.e., 25 observations into the new regime, c.f. Figure 2.9c). Although BS1L also

emits surprise here, it is still considerably smaller than PS1L and CS1L even after normalization (Figure 2.9e). Although model-confidence plays some role in Gaussian $CS(o_t)$, here, we do not see the implausible increase of CS1 regressors before the unlikely event that was present in conjugate first-order models with exponential forgetting, which makes the Gaussian $CS(o_t)$ much more fit to map theoretical first-order surprise.

2.4.6 Correlation Between Models

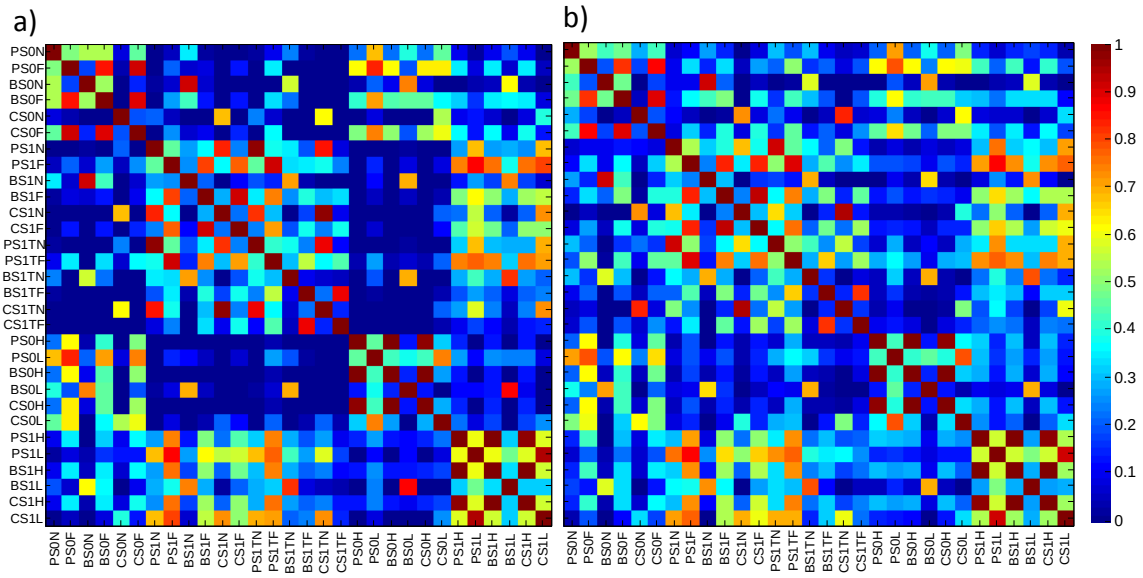

Figure 2.10: Correlation map of all reviewed models for a) the previous example sequence and b) average over 100 sequences sampled using the same hierarchical Markov-chain definition. This average correlation matrix is very similar to our example-sequence-matrix, suggesting that model correlations are fairly robust.

Figure 2.10a depicts correlation matrices for all possible relationships between the regressors we computed for the example sequence (Figure 2.3). To emphasize the robustness of these inter-regressor correlations, in Figure 2.10b we show average correlations from 100 regressors computed for additional sequences derived from the same stimulus-generation algorithm described in Section 2.2.2.

There are no large deviations from the single-sequence matrix, meaning that at least for our chosen Markov-chain paradigm, inter-regressor correlations stay roughly constant for different sequences. We obtain the strongest correlations between regressors of the three surprise functions in GRW models with high variance parameters. This is the case in both zero- and first-order sequence properties (P/B/CS0H, P/B/CS1H). We can explain this connection as follows: A high

variance parameter in a Gaussian model means that a high volatility is assumed by the Bayesian learner. Thus, new observations are taken with higher certainty compared to previous observations, which is represented by narrower distributions. For narrow distributions, there is no big difference between Bayesian and predictive surprise, as well as a high model-confidence from the start. A similar reasoning holds in the conjugate model class, where regressors from the three surprise functions are highly correlated when exponential forgetting is implemented for zero-order properties (P/B/CS0F).

Furthermore, Bayesian surprise correlates with itself across models because of fast convergence to zero, while predictive and confidence-corrected surprise regressors show high correlations within one model-class emanating from the compositionality of $CS(o_t)$ with a larger influence of $PS(o_t)$ (since in most cases, $BS(o_t)$ soon converges to zero). BB models for AP and TP result in high regressor-correlations for predictive and confidence-corrected surprise functions as well.

Despite these exceptions, we find that the Bayesian learner models combined with predictive and Bayesian surprise functions largely result in very heterogeneous surprise trajectories. These correlation results further underline the point that using different models and functions for surprise in many cases leads to vastly disparate assumptions about which events lead to which amount of surprise.

2.5 Discussion

Most applications of the BBH to the perception of stimulus sequences (e.g., Ostwald et al., 2012; Kolossa et al., 2015; Lieder et al., 2013b; Behrens et al., 2007; Iglesias et al., 2013) differ greatly in the *kind* of Bayesian learning that was assumed to be performed by the brain. This chapter aimed at shedding some light on these degrees of freedom an SBL model can have, and how these play out for some of the most simple models and sequences with two different stimuli (see Figure 2.11 for an overview). Specifically, here we documented BB as well as GRW SBL models for SP, AP, and TP sequence characteristics, that are built to learn from sequences with two different possible observations ($o_t = 1$ and $o_t = 0$). Next, we applied surprise functions from the literature for PS, BS, and CS to each of these models. To review regressors that can be derived by these combinations of SBL model and surprise function, we created an example sequence (c.f. Figure 2.3) as input and described each resulting regressor in detail in Section 2.4. It is possible to extend these models to create surprise regressors for any input sequence in infinitely many ways (for

example by using hierarchies of models as in Mathys et al., 2011, 2014). Nonetheless, we believe that it is important to clearly set up, classify and describe these simple models to understand their underlying assumptions before extending them to more complex versions. In any case, when arguing for Bayesian inference processes in the brain, it is important to be aware of the parts of the model that were deliberately chosen by the experimenter and how they influence the surprise regressor.

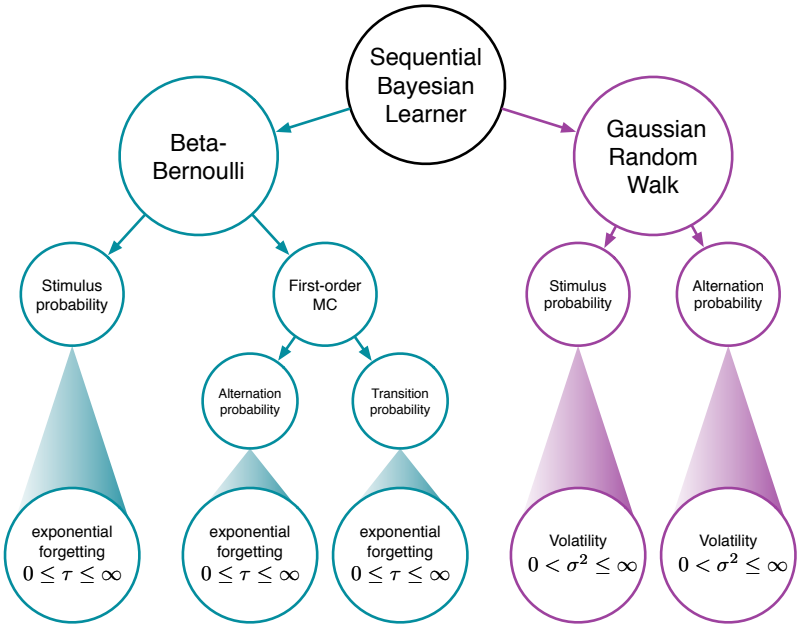

Figure 2.11: Sequential Bayesian learner models. MC: Markov-chain. Color gradients symbolize the continuum of values this parameter can take, while arrows stand for either-or-decisions. Each specific model can then be combined with a predictive, Bayesian, or confidence-corrected surprise function to yield a surprise regressor in order to test its fit to neural surprise signals.

In this chapter, a simple scenario with sequences containing only two different stimuli was considered. Most likely, greater differences in surprise regressors will show when more complex input sequences are concerned. One limitation of the theoretical analysis in this chapter is the use of the same hierarchical Markov-model for generating input sequences. This may have had a sizable influence on the regressor correlation matrix obtained to compare different models and surprise functions (Figure 2.10) and should be considered when extending the models reported above to sequences sampled from different generative models. However, as a test of Bayesian model-heterogeneity, using only one Markov-model specification for the input structure is appropriately strict, as it emphasizes the relevance of sequence property, probability distributions and free parameters when choosing model settings.

The parameters that make for the greatest discrepancies between models are arguably the exponential forgetting and volatility parameters. These refer, in a broader sense, to the learning rate of the organism and could perhaps be dependent on the stimulus modality, the attention and the kind of learning or observing that is taking place. Specifically, while our Gaussian models employ a constant learning rate throughout the whole sequence, learning depends on t in BB models without implemented forgetting, leading to fast learning in the beginning of a sequence and much slower learning with higher t.

It seems most sensible to either make a cognitively- or neurobiologically-informed decision on forgetting and volatility parameters, or to estimate them on the basis of the acquired data, either behavioral or neurophysiological. A necessary prerequisite for the latter approach is using different data sets for training (i.e., estimating the learning rate parameter) and testing (i.e., determining the model fit). As Griffiths et al. (2012) propose, the deviations from the mathematically optimal solution that our brain responses show (e.g., with a lower learning rate or higher volatility) are not arbitrary. Instead, they themselves reveal certain properties of the functioning of the brain.

Surprise trajectories are impacted differently by BB and GRW models as well. BB models without implemented forgetting usually display a sharp decrease in BS, a slower convergence to $ln(0.5)$ in PS, as well as the linear combination of the two together with a slowly increasing model-comitment function in CS. With implemented forgetting, strong distinctions can still be recognized between surprise measures However, in Gaussian models these distinctions vanish after regressor normalization when volatility parameters are sufficiently high.

In summary, among research of sequential model updating and neural surprise responses, a multitude of different models and surprise functions are used. By providing formulations of the most commonly used models for Bayesian sequential perception as well as pointing out their similarities and differences when applied to an example input sequence, we address this heterogeneity in the literature and present a foundation for further discourse.

Chapter 3

Empirical Application: Computational Bayesian Modeling of Trial-by-Trial Somatosensory Mismatch Negativity Effects

This chapter presents an investigation of the somatosensory mismatch negativity (sMMN) in a roving paradigm as well as an application of computational models specified in the previous chapter. After a short introduction into relevant literature and motivation of the study, the EEG-experiment together with the conventional ERP- and a more fine-grained single-trial data-analysis approach is described. Results of a small sMMN effect reported for the first time in a roving paradigm and a non-significant trend for confidence-corrected surprise in a Beta-Bernoulli model are reviewed and discussed in the end. In conclusion, the results in this chapter call the assumption of an underlying modality-independent mechanism for MMN generation into question.

3.1 Introduction

Since the MMN is often interpreted as a neural response related to violations of a sequential regularity or the improbability of a stimulus embedded in a sequence (Garrido et al., 2008), it is a

likely candidate-signal for Bayesian learning.

First found in the auditory perceptual domain (Näätänen et al., 1978), visual (Tales et al., 1999; Czigler, 2007) and somatosensory (Kekoni et al., 1997; Restuccia et al., 2007; Spackman et al., 2010) analogues have been described as well, with evidence for Bayesian learning and predictive coding in all domains (Wacongne et al., 2012; Stefanics et al., 2018; Ostwald et al., 2012). Consequently, a modality-independent preattentive sequential Bayesian learning process in the brain seems plausible.

However, there are several open questions about the less-studied somatosensory domain in particular, before one can assume such modality-independence of the MMN. Firstly, it is not known whether somatosensory MMN (sMMN) occurs independent of the stimulation paradigm used. While MMN was discovered with rare oddball stimuli, it can also be obtained through a roving-paradigm with equal stimulus probability (SP), where the first stimulus of a train of equal stimuli is defined as deviant. This paradigm-independence is a given for auditory (Phillips et al., 2016) and visual (Stefanics et al., 2014) MMN, whereas the preattentive somatosensory analogue has so far only been studied with an oddball paradigm. One roving-paradigm EEG-study with median-nerve stimulation in two different intensities has found potentials resembling the sMMN (Ostwald et al., 2012), but since they also imposed a deviance-counting task-set on the subjects, they could not investigate mismatch responses of a purely preattentive, automatic nature.

The second question seamlessly connects to paradigm-independence, namely, it is about the abstractness of the rules inherent in the sequence that are automatically extracted as exhibited by the MMN. These rules are affected by the probabilities defining the observed stimulus sequences. While oddball stimulation is concerned with the SP per se (i.e., rarer stimulus elicits MMN), roving paradigms rely on a lower transitional probability (TP) for stimulus changes (i.e., the rare event of a stimulus change elicits MMN). The more clearly these rules are expressed, i.e., the closer SPs for oddball stimuli or stimulus-change TPs are to 0, the larger the MMN occurring at these events (Näätänen, 1992), and thus, the system must likely be able to track the certainty of rules as well.

To abstract this even further, an MMN also appears after unexpected auditory (Alain et al., 1994; Nordby et al., 1988; Horváth et al., 2001; Todorovic et al., 2011; Cornella et al., 2012; Mittag et al., 2016) or visual (Czigler et al., 2006) *stimulus repetitions* when TPs for stimulus-changes are very high. Again, such levels of rule abstraction remain unclear for sMMN.

When investigating MMN sensitivity to a more abstract sequence probability structure, we

reach limitations of conventional average-based ERP analyses. Such averaging procedures cannot take into account what an agent has possibly inferred about SPs and TPs at different stages of a stream of stimuli. Rather, one rigid definition for standard and deviant stimuli determines the averaging procedure, while any other possibly relevant information about the sequence environment is lost. Here, Bayesian learning offers an elegant solution, because such a model will sequentially adjust probability distributions over observed sequence characteristics and thus make predictions at each event of the sequence, making use of all trial-by-trial fluctuations in the EEG-signal (Mars et al., 2008). Assuming that expectations are derived by Bayesian inference about probabilities, and that the MMN amplitude is an instance of expectation violation, Bayesian models should be most suitable to predict MMN amplitudes for given sequences.

Nevertheless, the Bayesian brain hypothesis entails many degrees of freedom for its concrete application to data (Friston et al., 2018), which is evidenced by the multitude of Bayesian approaches used for MMN-amplitude-prediction on a single-trial basis (Lieder et al., 2013a; Ostwald et al., 2012; Wacongne et al., 2012; Winkler and Czigler, 2012). Furthermore, there are varying mathematical definitions of surprise functions one can use to extract trial-wise estimates of MMN-amplitude from a Bayesian model. Although predictive and Bayesian surprise (PS and BS, respectively) have been contrasted against each other for the later P300 components (Kolossa et al., 2015; Seer et al., 2016), there are no comparisons of their explanatory power for the MMN-amplitude in the literature.

In summary, auditory and visual MMN respond to improbable events on different abstraction levels beyond pure change detection and most likely exhibit perceptual Bayesian learning processes. Much less is known about the dynamics of sMMN. The study described in this Chapter aims at further completing the picture of a modality-independent automatic novelty detection mechanism which possibly generates the MMN. To investigate to what extent a roving paradigm and different TPs of stimulus-change can elicit an sMMN, we designed a Markov-chain roving-like paradigm with a hierarchical structure, such that phases of high stimulus-change TPs alternated with phases of rare stimulus changes. Using EEG data from 11 subjects, we perform an average ERP analysis to identify the sMMN. Moreover, as an exploratory proof-of-concept analysis, we apply an array of combinations of Bayesian learner models and surprise functions from Chapter 2 to single-trial EEG data. Taken together, the results point out limitations of automatic information processing in the somatosensory domain, as it appears more strongly confined to simpler sequence characteristics, while still going beyond a simple change-detection mechanism.

3.2 Material and Methods

3.2.1 Participants

20 healthy volunteers (10 female) participated in the experiment. All participants gave written informed consent (see Supplement Section 4.4 for participant instructions and consent form). The study corresponded to the Human Subject Guidelines of the Declaration of Helsinki and was approved by the Ethical Committee of the Charité University Hospital.

3.2.2 Stimuli

Electrical stimuli with a duration of 0.2 ms and a constant interstimulus-interval (ISI) of 650 ms were delivered to the left median nerve using adhesive electrodes placed to the wrist. Two levels of intensity (low and high) were determined for each subject individually such that the low intensity stimulation was close to sensory threshold but clearly noticeable for each stimulus repetition (mean 4.55 ± 1.04 STD mA). The high intensity stimulus was determined to be under the motor threshold and clearly distinguishable from the low intensity stimulus (mean 7.23 ± 1.63 STD mA).

3.2.3 Experimental Procedure

The experimental paradigm consisted of a continuous stimulus presentation during one block lasting about 12 minutes and including 1153 events of stimulation, referred to as *trials* from hereon. Each participant received 8 blocks of stimulation in total. During stimulation, participants did not complete any kind of task but specifically were instructed to not pay attention to the stimuli. This was done to test the attention-independence of the sMMN. As a means of distraction, they watched episodes of *Shaun the Sheep* (Starzak and Sadler, 2007) without sound (2 episodes per block, 16 episodes in total). Block order and episode order were randomized independently, to insure that the measured EEG activity is uncorrelated with specifics of the episode, as well as possible artifacts arising from episode traits (e.g., facial muscle contractions due to smiling). We instructed participants to fixate a fixation cross placed in the middle of the screen to minimize eye movements. Between the blocks, participants could take a short break to rest from the stimulation and to avoid fatigue. Participants were not informed about the regularity of the sequences or the sequence-generation process.

Specifically, stimulus sequences of high and low amplitude electrical stimulation were set according to the Markov-chain roving-like paradigm specified in Section 2.2.2, with the distinction that TPs (referred to here as P_{high} and P_{low}) could differ between blocks (see Figure 3.1 for a graphical display of the hierarchical Markov-chain). Each stimulation sequence was sampled anew for each participant and each block.

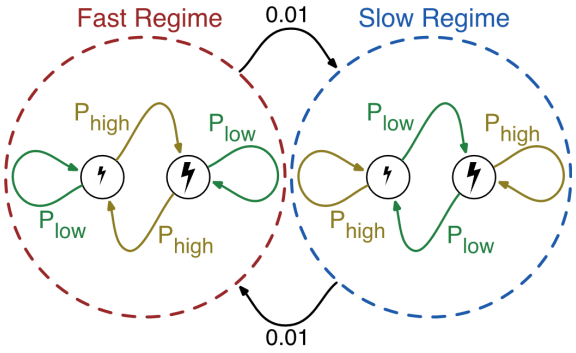

Figure 3.1: Markov Model Schema for the Markov-chain roving-like paradigm. For each observable state, transition probabilities are defined by the regime. In addition, at each state there is a probability of 0.01 for a regime change. In the fast regime, a state switch is more likely than a repetition, while in the slow regime, the opposite holds true. SP was always constant at 0.5 for both stimuli.

This paradigm can take on two different observable states in two different regimes, which are defined by the second and first level of the Markov-model, respectively. The observable states correspond to high- and low-amplitude stimulation in the MMN-paradigm, while the regime-states remain hidden and can only be inferred by the frequency of transitions.

Table 3.1: Transitional probabilities in the hierarchical Markov chain for conditions A-D

Conditions	Transition probabilities		Regime change
	P_{high}	P_{low}	
A (control condition)	0.495	0.495	0.01
B	0.62	0.37	0.01
C	0.745	0.245	0.01
D	0.87	0.12	0.01

Four experimental conditions of TP were distributed randomly over 8 blocks, such that each condition appeared in two of the 8 blocks. Figure 3.2a shows a schematic example procedure. Specifically, conditions differed in their TP matrix defined by P_{high} and P_{low}. In all but one

condition, the sequence switched between a fast and a slow regime. In the remaining one condition, the so-called control condition, there was a static transitional probability of 0.495 to either change the state or stay in the state (with a inconsequential 0.01 probability to switch regimes). Thus, the regimes did not differ in the control condition.

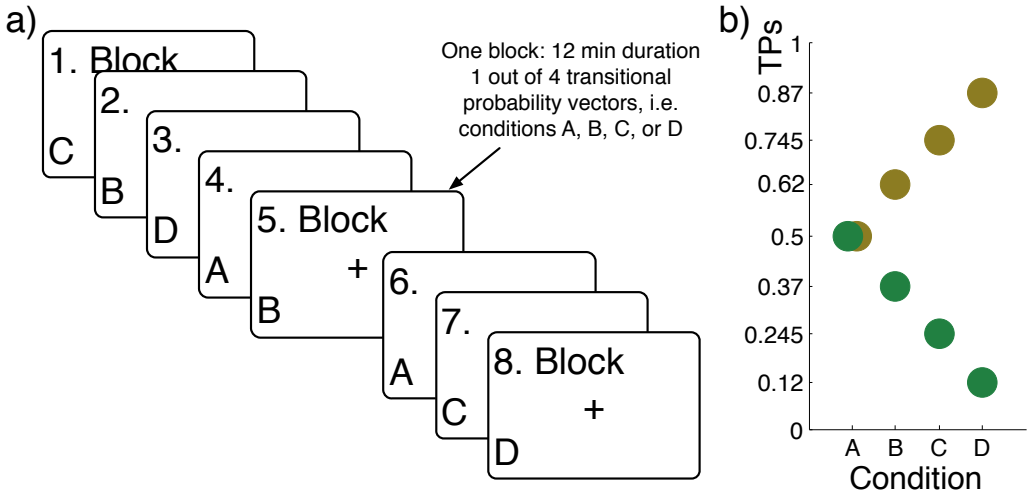

Figure 3.2: a) Schema for the block-structure of the experiment. Each subject completed 8 blocks, with a randomized order of TP conditions. Each condition occurred twice in the experiment. b) Transitional probabilities for conditions A, B, C, and D. High TPs for all conditions are shown in yellow, low TPs in green. Condition A is special in that TPs are equal with 0.5, resulting in a non-hierarchical zeroth-order Markov chain from which the sequence is sampled.

See Table 3.1 as well as Figure 3.2b for an overview for TPs in the four conditions. From condition A (the control condition) to condition D, differences between the two regimes become gradually more apparent.

In our experimental procedure, we asked the question if hidden states modulate the electrophysiological mismatch response. In the paradigm applied here, hidden states were operationalized by the different regimes, i.e., switching TPs. If the MMN should indeed be modulated by these first-order Markov-chain properties, modulation should be strongest for condition D, weak for condition B and non-existent for the control condition without regime differences. It is important to note that we chose the TPs to be independent of the preceding event, such that the overall probability for each possible event (high or low amplitude stimulus) was 0.5, just like in a classical roving paradigm. This insured that stimuli of both intensities could take on the role of either standard or deviant stimuli with equal likelihood. That way, we eliminated stimulation-intensity and stimulus-specific adaptation effects from the mismatch response.

3.2.4 EEG Recording and Preprocessing

3.2.4.1 General Preprocessing

EEG was recorded with a 64-channel active electrode system with electrodes placed according to the extended 10-20 system and at a sampling rate of 2048 Hz (ActiveTwo, Biosemi). All data preprocessing and analysis steps were performed with Matlab 2011 and SPM8 (Litvak et al., 2011). Before analyzing the EEG signals of interest, the following standard-preprocessing steps were performed: rereferencing to average reference, high-pass filtering with a 1 Hz cutoff, eye-blink correction using a topological confound approach (Berg and Scherg, 1994; Litvak et al., 2007), low-pass filtering with a 40 Hz cutoff, epoching using a peri-stimulus time interval of -50 to 650 ms, baseline correction using the pre-stimulus interval of -50 to 0 ms, and artifact rejection of all trials containing amplitudes larger than 60 μV. Subsequently, we inspected all data trials and marked bad channels that continuously showed artifact activity such as muscle-artifacts or remainders of line-noise. In datasets where we found such bad channels, we marked them in the raw data and repeated all automatic preprocessing steps, to insure that no bad-channel-activity can influence other channels through rereferencing. All further ERP analyses were done on the original 2048 Hz sampling rate data to avoid downsampling-artifacts. For the subsequent single-trial analysis, however, we downsampled the preprocessed data to 512 Hz for shorter computing times.

3.2.4.2 Stimulation-Artifact Removal

The electrical stimulation elicited a stronger than expected artifact influencing not only amplitudes at $t = 0$ but also up to several milliseconds after stimulation (see Figure 3.3, bright red line). Presumably, this was due to the active electrodes used to record the data. Because the stimulation current has a high peak at $t = 0$, the active electrodes measure a large voltage difference between head and amplifier, which the system tries to balance by applying current as well. This leads to an offset right after the stimulus, which affects the time course of almost the whole ERP.

In order to remove this artifact in an adequate spatio-temporal manner, we used the same topological confound approach as Litvak et al. (2007) proposed for EEG artifacts induced by transcranial magnetic stimulation (TMS) relying on the method by Berg and Scherg (1994). Put briefly, we create a source model consisting of artifact as well as brain topographies. For this, we compute the artifact topographies with a principal component analysis (PCA) decomposition of

the averaged artifact and assume a set of brain topographies as multiple dipoles modeling brain activity. Subsequently, we can compute a linear inverse operator to decompose the data into a linear combination of artifact and brain activities. Now, we can subtract the estimated artifact activities while leaving the modeled brain activity unchanged (see Litvak et al. (2007) for a detailed mathematical description of this approach, as well as SPM functions `spm_eeg_spatial_confounds` and `spm_eeg_correct_sensor_data` for application). The advantage of this approach is that no assumptions about independence of brain- and artifact-activities have to be made. In our experimental setup, high and low amplitude stimuli were administered by two different stimulator machines, and in many cases, stimulus artifacts for high and low stimuli were reversed in many cases. This is why we applied the artifact correction for each stimulus amplitude separately. We performed this artifact correction step after rereferenceing and high-pass filtering (1 Hz cutoff) and before eye-blink correction. Figure 3.3 shows a timecourse of an average over low-amplitude stimuli before and after stimulation-artifact removal (data from one subject). While the offset right after the stimulus is rectified with the artifact correction, it also has an effect on the amplitude several hundred milliseconds after the artifact.

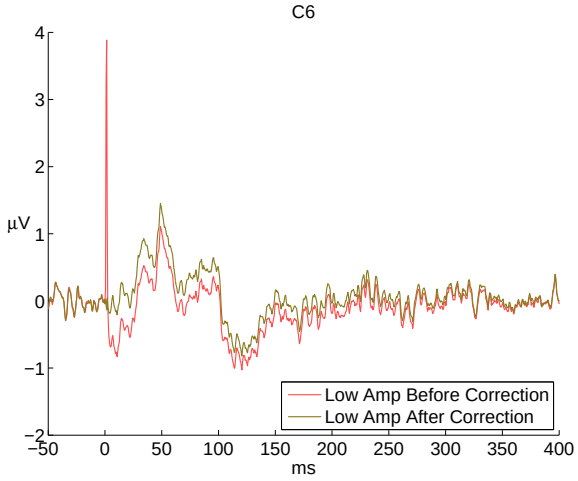

Figure 3.3: Single subject average over raw EEG signal to low amplitude stimuli, both before (bright red) and after (dark yellow) PCA artifact correction. In the uncorrected timecourse, the peak at $t = 0$ is elicited by the current of the stimulation, while the "undershoot" right after the stimulus was most likely induced by the active electrodes trying to even out the voltage difference between head and amplifier. By correcting with PCA, this undershoot vanished. Timecourses are shown right before and after PCA correction, meaning that the data are rereferenced and high-pass filtered (1 Hz cutoff) only.

3.2.4.3 ERP Averaging

Somatosensory evoked potentials (SEPs) of experimental factors of interest were computed by standard averaging. After preprocessing, we excluded EEG-data of nine participants from further analyses due to either bad data quality (artifacts in more than 30% of the trials, two subjects), no significant SEP for low intensity stimulation/no difference in SEP between high and low intensity (one subject) or both (six subjects). All further analyses were computed on the 11 participant datasets left. Channels marked as bad in at least one of the subjects were excluded from further analyses, remaining channels are shown in Figure 3.4. For scalp topographies, activities of bad channels were interpolated using the freely available *fieldtrip* software (Oostenveld et al., 2011). To avoid *double-dipping* (as described in Kriegeskorte et al., 2009), we pre-defined electrodes of interest to encompass a spatial array over contralateral somatosensory cortex as well as some frontal areas. This array was chosen based on somatosensory mismatch effects in the literature (Ostwald et al., 2012; Kekoni et al., 1997; Restuccia et al., 2007). Specifically, after left median nerve stimulation with two different intensities in a roving paradigm, Ostwald et al. (2012) found mismatch responses on electrodes FCz, C4, C6, and F4, while Restuccia et al. (2007) showed oddball effects at CP6, C4, FC2, and F4 using electrical stimulation of left thumb and fifth finger. Consequently, our array of electrodes of interest covers all of these right-centroparietal and frontal areas.

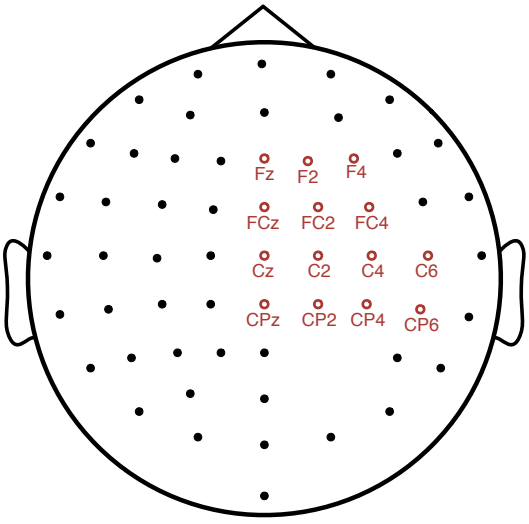

Figure 3.4: Array of electrodes included in group analyses after the exclusion of bad channels. Electrodes of interest are marked in red. Strongest artifacts and thus bad channels were observed around the DLR electrode of the Biosemi active electrode system, as evidenced by the lack of good channels in the right parietal area of the channel array.

3.2.5 ERP Analysis

For the ERP analysis, we computed two kinds of general linear model (GLM) on the 14 channels of interest (as displayed in Figure 3.4). In GLM (1), all trials that passed artifact exclusion were used to look at general effects of stimulation amplitude and regime during the whole poststimulus time window. We established GLM (2) to specifically look at mismatch effects, and thus, only trials defined as either deviant or standard were included in the model. Here, *deviants* were defined as the first stimulus differing in intensity after a train of at least two equal stimuli (cf. Figure 3.5a). The stimulus right before a deviant is the *standard*. This definition of comparably short minimum-length stimulus trains as well as a wide range of possible train length was chosen, so that a high number of trials of each experimental condition and regime would factor into the ERP of the MMN. For mismatch effects, responses within $100 - 250$ ms were of interest, based on findings in literature for auditory (Garrido et al., 2009) and somatosensory (Kekoni et al., 1997; Shinozaki et al., 1998) MMN.

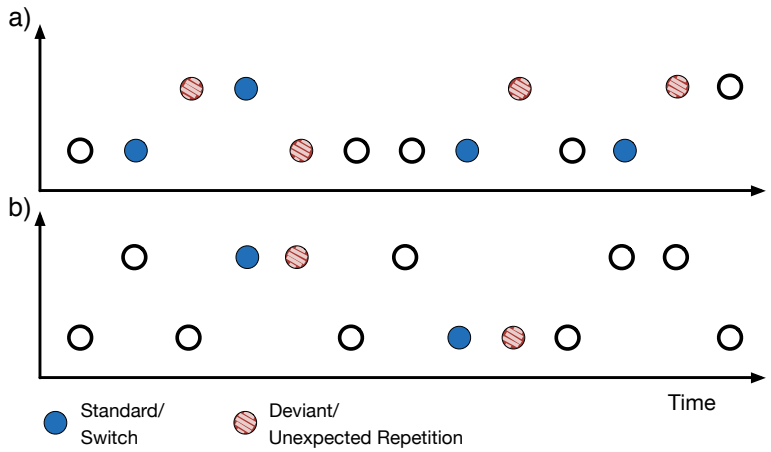

Figure 3.5: a) Schema of trial definition as *standard* or *deviant* as applied to our roving paradigm in GLM (2). Deviants are the first stimulus differing in intensity after a train of at least 2 equal-intensity stimuli. b) Trial definition of *switches* as stimulus after at least 2 alternations and *unexpected repetitions* as events of equal intensity right after a switch, used in GLM (3). In both panels, the y-axis represents the stimulus feature dimension along which two different stimuli vary (in our case, intensity of electrical stimuli).

GLM (3) was computed to explore whether stimulus repetitions after several alternations could elicit ERP effects similar to the MMN. For this purpose, stimuli were categorized as *unexpected repetitions* when the same intensity event occured after at least two alternations, and as *switches* for stimuli preceding the repetition.

The computed GLMs had the following form

$$y = X\beta + \epsilon, \tag{3.1}$$

with data matrix y of preprocessed single trials, stacked across subjects, design matrix X, estimated parameters β and an error term ϵ. The design matrix of all three models included an intercept and dummy variables for subjects to explain between-subject variance. For GLM (1), regressor variables were stimulation amplitude (1 for high vs. -1 for low stimulation) and regime (1 for fast vs. -1 for slow regime). GLM (2) had regressors for deviance (1 for standard vs. -1 for deviant), experimental condition of TPs ($1, 2, 3$, and 4, for A, B, C, and D, respectively, see section 3.2.3 and Table 3.1 for a description of TP conditions), and stimulation amplitude (1 for high vs. -1 for low stimulation). GLM (3) mimicked (2) with switches and unexpected repeats (1 and -1, respectively) instead of the deviance variable. It is important to note, that here, the experimental condition regressor assumes a parametric modulation of ERP effects, i.e., with larger TP difference a larger ERP effect is modeled. The GLMs were fitted using the Matlab function `lmfit` and included interaction terms. `lmfit` returns t- as well as p-values for the respective parameters β fitted in the model. In this experiment, interactions between deviance and stimulation amplitude, as well as deviance and TP condition were of special interest.

3.2.6 Single-Trial Analysis

For a single-trial analysis of the MMN component, Chapter 2 established computational models that can capture certain statistical regularities from the stimulation sequence following the form of a sequential Bayesian learner (SBL, c.f. Bishop, 2006). Using an SBL to test the BBH is a straightforward approach, since it puts its basic tenets into a mathematical form. However, there are many possible ways to explicitly formulate such models. Here, we consider a subset of the models, namely the Beta-Bernoulli (BB) model capturing SP and TP statistics, and their functions for predictive, Bayesian, and confidence-corrected surprise (PS, BS, and CS, respectively) as specified in Section 2.3.1.1.

These SBL models use conjugate distributions (i.e., the Beta distribution is a conjugate prior for Bernoulli-likelihood function) and in effect assume the hidden variable s to be static. They have the mathematical advantage of integrals to be analytically solvable.

We tested versions of the BB models without and with exponential forgetting applying $\tau = 0.14$

and $\tau = 0.035$ corresponding to a half-life of 5 and 10 trials, respectively. Because of their comparatively low performance, we did not pursue the analysis of models with exponential forgetting any further (c.f., Figure 3.6).

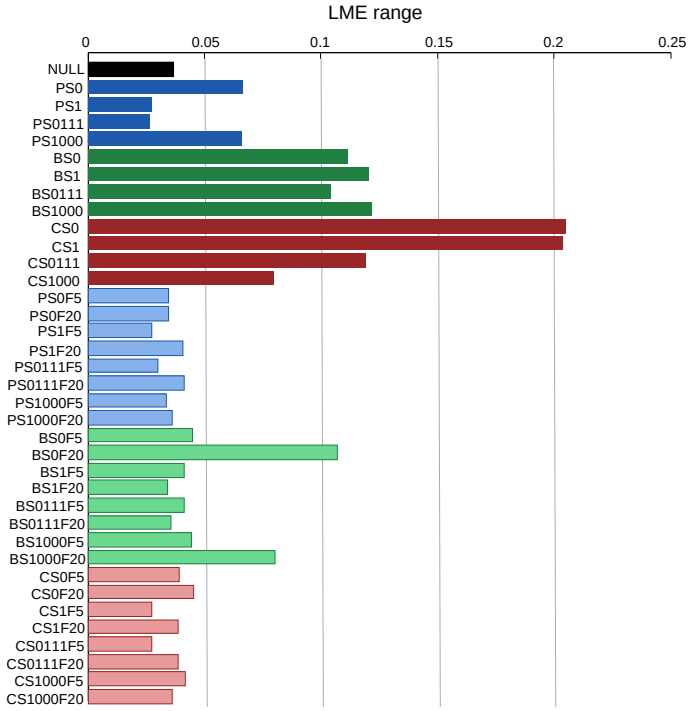

Figure 3.6: Log model evidence range of the group sum over electrodes of interest for static BB SP and TP (dark colors) as well as BB SP and TP with exponential forgetting parameters $\tau = 0.14$ (half-life of 5 trials, ending with F5) and $\tau = 0.035$ (half-life of 20 trials, ending with F20). For a detailed description of the implementation of exponential forgetting, the reader is referred to Section 2.3.1 of this thesis.

In addition to conjugate models, we also set up Gaussian Random Walk (GRW) models (c.f. Behrens et al., 2007), where the hidden variable is assumed to change with variance σ^2 of the Gaussian distribution (c.f., Section 2.3.2). To test in what range σ^2 could fit to our data, we tested models with $\sigma^2 = [2, 1, 0.2, 0.1, 0.05]$ and specify those with the endings H, M, L, LL, and LLL in the model names, respectively (see Table 3.2).

For each combination of the BB or Gaussian models and three surprise measures, we calculated surprise regressors according to the subject- and block-specific stimulation sequences. With each start of a new block, priors were set to $t = 0$, assuming no memory transfer between blocks but rather starting without assumptions into the next block (as in Mars et al., 2008). Because we assumed the effects to be very small, to maximize statistical power we concatenated the regressors for one subject over the whole course of the experiment in order to test all trials at once. However, concatenating only regressors of one SBL model would make it impossible to test the conceivable

case that the somatosensory system learns transitional probabilities only in environments where they actually vary (i.e., experimental conditions B,C, and D) and not when they stay the same. The opposite case (learning only constant TPs and not those that vary) can also be a reasonable mechanism. To gain insight into these possibilities, we established two kinds of mixture-regressors for all surprise functions, with BB SP (or GRW SP) underlying for condition A and BB TP (or GRW AP) for all others (ending with 0111), as well as the opposite (BB TP/GRW AP for condition A, BB SP/GRW SP for all others; ending with 1000).

As shown in Table 3.2, the conjugate combinations lead to a 4×3 model space consisting of the factors order (0th, 1st, or two mixtures), and form of surprise (predictive, Bayesian, or confidence-corrected surprise). GRW models were tested with 5 different variance-parameter specifications and thus led to a $20 \times 3 = 60$ model $\times$ surprise function combinations. Additionally, to explore the noise-level in our data, we tested a Null-model regressor containing of 1 for the first trial included in the analysis (not necessarily the first trial measured per se) and zeros for all other trials. Furthermore, a primitive stimulus-change model using a regressor with 0 if $o_t = o_{t-1}$ and 1 if $o_t \neq o_{t-1}$ was established and tested (as hypothesized by Näätänen, 1992; Schröger and Winkler, 1995). Thus, we analyzed 74 regressors in total.

All 74 regressors were tested using the `spm_PEB` (i.e., parametric empirical Bayes) function, which is also part of the SPM8 (and SPM12) software package (Litvak et al., 2011). This function serves as as a means to calculate Bayesian estimates of log model evidences for hierarchical GLMs of the form (e.g, with two levels, Dempster et al., 1981)

$$\begin{aligned} y &= X^{(1)}\theta^{(1)} + \epsilon^{(1)} \\ \theta^{(1)} &= X^{(2)}\theta^{(2)} + \epsilon^{(2)}. \end{aligned} \tag{3.2}$$

Here, $y \in \mathbb{R}^N$ is an $N \times 1$ data vector with N trials, $X^{(1)} \in \mathbb{R}^{N \times R}$ is the design matrix with R predictors, $\theta^{(1)} \in \mathbb{R}^R$ is the parameter vector and $\epsilon^{(1)} \in \mathbb{R}^N$ is the error vector on the first level. The second level (second row of (3.2)) gives the opportunity of modeling the first level parameters $\theta^{(1)}$ with another GLM. For our purposes, with the `spm_PEB` function we implemented non-hierarchical Bayes by using a two-level model and setting the second level design matrix to zeros (Dempster et al., 1981; Kolossa, 2016), leaving

$$\theta^{(1)} = \epsilon^{(2)}. \tag{3.3}$$

Table 3.2: Surprise regressors as a combination of SBL model and surprise function

SBL model	Surprise functions		
	PS(t)	BS(t)	CS(t)
BB SP	PS0	BS0	CS0
BB TP	PS1	BS1	CS1
BB Mixture 0111	PS0111	BS0111	CS0111
BB Mixture 1000	PS1000	BS1000	CS1000
GRW SP $\sigma^2 = 2$	PS0H	BS0H	CS0H
GRW SP $\sigma^2 = 1$	PS0M	BS0M	CS0M
GRW SP $\sigma^2 = 0.2$	PS0L	BS0L	CS0L
GRW SP $\sigma^2 = 0.1$	PS0LL	BS0LL	CS0LL
GRW SP $\sigma^2 = 0.05$	PS0LLL	BS0LLL	CS0LLL
GRW AP $\sigma^2 = 2$	PS1H	BS1H	CS1H
GRW AP $\sigma^2 = 1$	PS1M	BS1M	CS1M
GRW AP $\sigma^2 = 0.2$	PS1L	BS1L	CS1L
GRW AP $\sigma^2 = 0.1$	PS1LL	BS1LL	CS1LL
GRW AP $\sigma^2 = 0.05$	PS1LLL	BS1LLL	CS1LLL
GRW Mixture 0111 $\sigma^2 = 2$	PS0111H	BS0111H	CS0111H
GRW Mixture 1000 $\sigma^2 = 2$	PS1000H	BS1000H	CS1000H
GRW Mixture 0111 $\sigma^2 = 1$	PS0111M	BS0111M	CS0111M
GRW Mixture 1000 $\sigma^2 = 1$	PS1000M	BS1000M	CS1000M
GRW Mixture 0111 $\sigma^2 = 0.2$	PS0111L	BS0111L	CS0111L
GRW Mixture 1000 $\sigma^2 = 0.2$	PS1000L	BS1000L	CS1000L
GRW Mixture 0111 $\sigma^2 = 0.1$	PS0111LL	BS0111LL	CS0111LL
GRW Mixture 1000 $\sigma^2 = 0.1$	PS1000LL	BS1000LL	CS1000LL
GRW Mixture 0111 $\sigma^2 = 0.05$	PS0111LLL	BS0111LLL	CS0111LLL
GRW Mixture 1000 $\sigma^2 = 0.05$	PS1000LLL	BS1000LLL	CS1000LLL
Null model		Null	
Stimulus-change model		STMC	

This sets an unconstrained prior on the first-level parameters $\theta^{(1)}$ and thus allows for single level Bayesian inference (Friston et al., 2007). In our case, $X^{(1)}$ is a single column vector with just one predictor, i.e. surprise regressor per GLM. Subsequently, PEB estimates the conditional posterior probability densities of parameters θ, i.e. $p(\theta|y)$ using the expectation maximization (EM) algorithm (Dempster et al., 1977). Regarding the parameters as random variables instead of fixed values makes this approach distinct from classical parameter optimization (Kolossa, 2016). Roughly, EM is an iterative scheme alternating between an E-step and M-step. In PEB, the E-step calculates the conditional mean and covariance of the model parameters θ, while keeping error-parameters fixed. Conversely, in the M-step all model parameters are left unchanged while error-covariances are estimated (Friston, 2002). The algorithm repeats these steps until convergence and estimates the parameter densities as well as the variational free energy with an accuracy and complexity term that can be used for model selection. The variational free energy is the lower bound approximation of the log likelihood $\log p(y|m)$, which is usually not computable. For a more detailed explanation and matrix-notation of hierarchical GLM and PEB, the reader is referred to Kolossa (2016, pp. 19-28). For our analyses, we used this variational free energy as estimates of the log model evidence (LME).

Figure 3.7: Schematic deptiction of the independent variables in this study divided by experimental manipulation (top) and investigation in the analysis (bottom).

On the single-subject level we calculated for each electrode (64 in total) and each peri-stimulus time bin (333 in total, due to 512 Hz data sampling and a -50 to 600ms time window) a LME value for the respective regressor (74 in total) across trials. Next, we summed the LMEs over subjects to gain the group log evidence for each model, electrode, and timebin (following Garrido et al.,

2007). After exploring the group LMEs of all 74 regressors, we decided on only further analyzing the subset of *conjugate models* (12 regressors) for group effects. We then subtracted the LME of the Null-model from all remaining 12 models at each respective time bin for a relative LME (equal to the log-Bayes factor in Kolossa et al., 2015). To further investigate effects on the group-level, we used the `spm_BMS` function (again, part of SPM8 and SPM12; Stephan et al., 2009; Rigoux et al., 2014), which performs Bayesian model selection for group studies as a random effects analysis. This function consists of a variational Bayes method, where the model itself treated as a random variable. An important output of this function is the exceedance probability, which quantifies the probability of any given model being more likely than all other models in the probability set (an even more robust and conservative measure is the protected exceedance probability, see Rigoux et al., 2014). Also, the function returns the conditional expectations of model probabilities and a vector of model probabilities α (the estimated parameters of the Dirichlet distribution, to be interpreted as the number of model occurrences within the sample of subjects). While all three outputs can be used to rank models at the group level, the exceedance probabilities are the most useful measure to base the model selection on (Stephan et al., 2009). Finally, we compared model families (using a fixed-effects model with the SPM function `spm_compare_families`) of surprise functions (columns of Table 3.2) as well as SBL models (rows of Table 3.2), to see whether one surprise function or SBL model is better in explaining the data overall.

Figure 3.7 shows all independent variables considered in this study. To summarize, here we investigate the somatosensory MMN with a Markov-chain roving-like paradigm in order to examine the influence of transitional probabilities and the overall statistical environment on sMMN amplitude. Furthermore, in a single-trial analysis we looked at different SBL models and surprise functions to assess their contribution to the amplitude of the sMMN.

3.3 Results

In this study, we sought to replicate the MMN in the somatosensory domain, as well as investigate the assumed domain-, attention-, and paradigm-independence of the EEG potential. Our key contribution lies in the systematic comparison of several SBL models and surprise functions in a single-trial analysis in order to assess in what way Bayesian learning contributes to the MMN amplitude. Thus, our approach combines the standard ERP analysis with a more fine-grained approach to the MMN (roughly comparable to Lieder et al., 2013a; Ostwald et al., 2012). To our

knowledge, this is the first study with such a broad model and surprise function comparison for the MMN (see Kolossa et al., 2015, for a similar approach investigating the P300 component).

3.3.1 Event Related Potentials

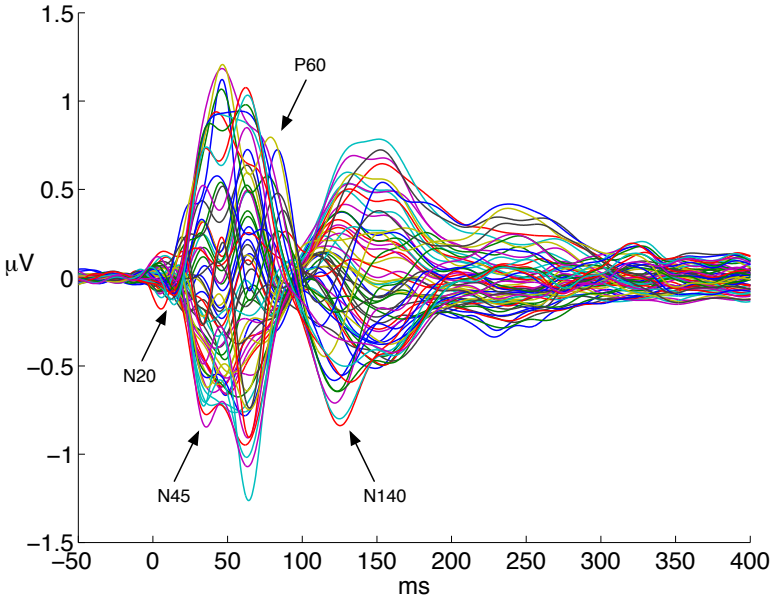

Figure 3.8: Grand Average over all all stimuli and all subjects. Each Graph represents one electrode. The respective components of a sensory evoked potentials (SEPs) are indicated with arrows.

First, we tested whether the median nerve stimulation was successful and lead to standard SEP components. ERP curves of all electrodes after preprocessing, averaged over all stimuli and participant data are shown in Figure 3.8. The main components of an SEP (namely, N20, N45, P60, and N140) could be replicated. While electrodes frontal to the central sulcus (i.e., more frontal that Cz, see Figure 3.4 for electrode array) usually measure a larger positive polarity change after about 40 ms, caudal electrodes show the opposite course (see Allison et al., 1991, for a thorough review of median nerve SEPs).

3.3.1.1 Stimulation-Amplitude Effects

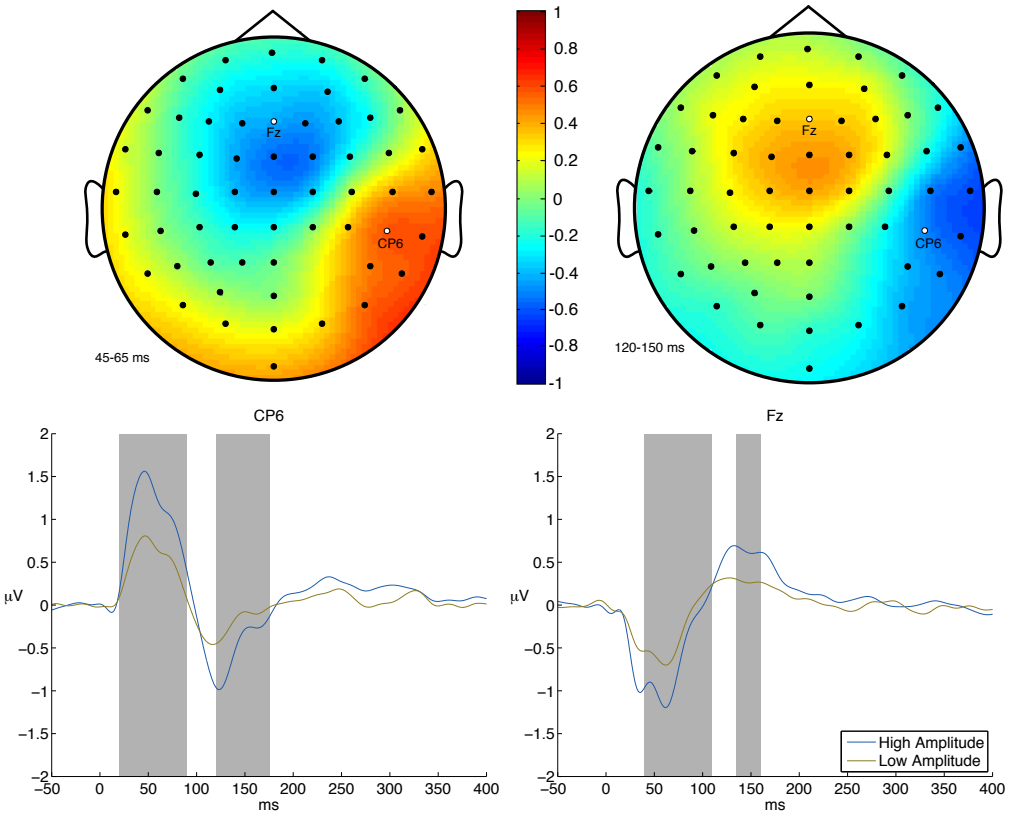

Figure 3.9: Topography and ERP-timecourse of *high* - *low* stimulation amplitude. Upper Panel: scalp-topographies of *high* - *low* contrast averaged over $45 - 65$ ms as well as $120 - 150$ ms post stimulus. Lower Panel: ERP-timecourses of electrodes CP6 and Fz averaged over high and low amplitude stimuli, respectively. Grey bars mark significant ($p < 0.05$, uncorrected) difference timewindows between high and low amplitude trials.

As a control analysis, we compared high to low amplitude stimulation in GLM (1). Higher stimulus amplitude should lead to higher SEP component amplitude in all subjects over contralateral somatosensory cortex and frontal areas. Main effects of stimulation-amplitude were present in all of the electrodes of interest. Figure 3.9 shows electrodes with peak effects. Significant differences in ERPs from stimulation amplitude were present as early as 20 ms and lasted up until 170 ms after stimulation. See Table 3.3 for significant time windows and peak effects of electrodes Fz and CP6. Thus, the median nerve stimulation can be deemed as successful.

Further, we asked whether the overall statistical environment in terms of our stimulation regimes (c.f. Figure 3.1) with fast and slow changing stimulus trains has an influence on the ERP. These regime differences were nonexistent in condition A and largest in condition D (see Figure 3.2b). No

Table 3.3: Main effects of stimulation amplitude high vs. low

Electrode	Early effect		Late effect	
	Duration	Peak	Duration	Peak
Fz	$38 - 110$ ms	52 ms, $p < 0.001$	$134 - 161$ ms	147 ms, $p < 0.001$
CP6	$20 - 90$ ms	48 ms, $p < 0.001$	$118 - 176$ ms	149 ms, $p < 0.001$

Note. Durations of effects are reported for all p-values < 0.05. All p-values are uncorrected for multiple comparisons.

main effects of regime (fast or slow) as well as experimental condition (none to high difference in transitional probabilities) could be found. Also, no significant interactions were revealed in GLM (1). Therefore, across all stimuli, TPs had no categorical effect on the ERP amplitude in terms of a higher *staying* or *switching* probability, and no "parametric" effect of larger regime differences (experimental conditions A-D).

3.3.1.2 Mismatch Effects

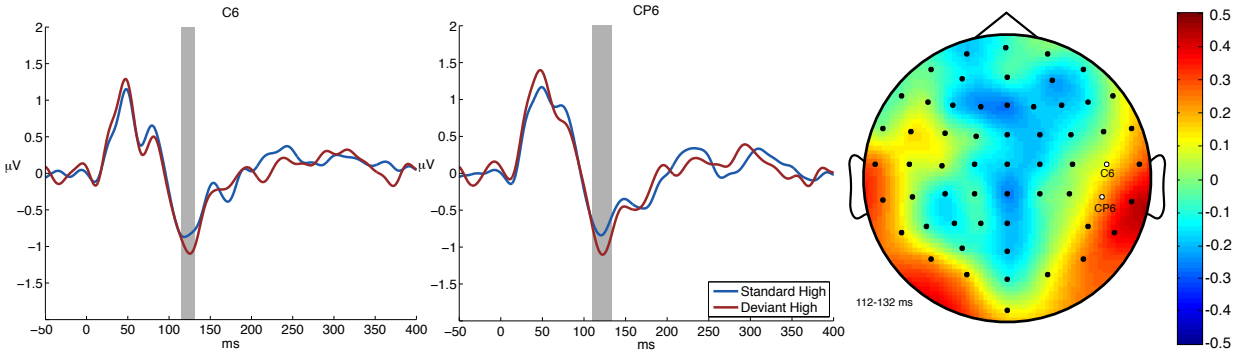

Figure 3.10: Early Mismatch effect. First two panels: Time-courses of electrodes showing significant differences ($p < 0.05$, uncorrected, indicated by gray bars) between high-amplitude standard and deviant stimuli. Third panel: Scalp topography of Standard - Deviant difference waveform.

We looked for a mismatch response (higher amplitude for deviant than for standard stimuli, negative as well as positive) with GLM (2). Here, we defined stimuli as *deviants* and *standards* according to a liberal definition of trains of at least two equal stimuli before a deviant (cf. Figure 3.5). This definition of comparably short minimum-length stimulus trains as well as a wide range of possible train length was chosen, so that a high number of trials of each experimental condition and regime

would factor into the ERP of the MMN. Still, in GLM (2), across all stimulus intensities (high and low amplitude of stimulation), no significant main effect for mismatch (standard vs. deviant) within the MMN time-window of $100 - 250$ ms could be identified. This was possibly due to the weak electrophysiological response to low amplitude stimuli overall (cf., Figure 3.11).

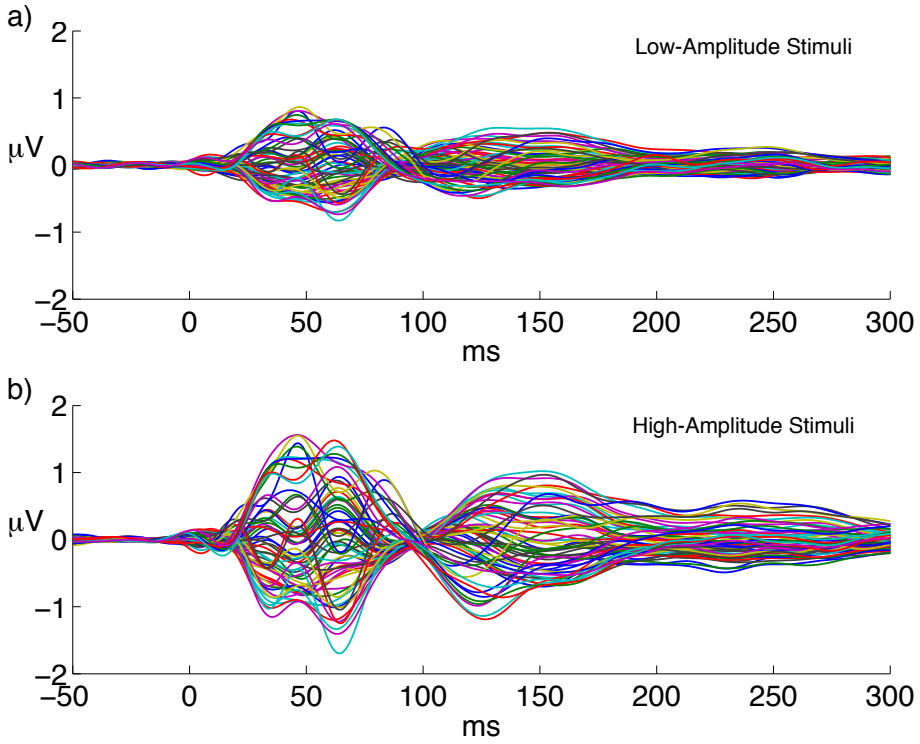

Figure 3.11: ERP-timecourse of low-amplitude stimulation (upper panel) and high-amplitude stimulation (lower panel) over all electrodes. With ERP-amplitude being below 1 μV at SEP-peaks, low-amplitude stimulation elicited comparably weak SEP-response.

There was, however, an interaction between TP condition and mismatch effect, and thus, a parametric influence of TP condition on MMN amplitude. An earlier interaction was found on electrode CP6 peaking at 142 ms ($p = 0.042$), while a later effect becomes apparent at FCz (peak at 228 ms, $p = 0.022$) and FC2 (peak at 225 ms, $p = 0.01$). Broadening the time-window of interest to include earlier *standard - deviant* effects, we found a larger P50 for deviants than standards on electrode C4, ranging from 50 to 66 ms with a peak at 59 ms post-stimulus ($p < 0.006$). This incidental finding is discussed in Section 3.4.

Table 3.4: Main effects of high-amplitude standard vs.deviant stimuli

Electrode	Early effect	
	Duration	Peak
C6	$111 - 132$ ms	125 ms, $p < 0.006$
CP6	$107 - 133$ ms	123 ms, $p < 0.004$
Electrode	Late effect	
	Duration	Peak
Fz	$171 - 180$ ms	176 ms, $p < 0.027$
F2	$173 - 179$ ms	176 ms, $p < 0.032$
FC2	$169 - 183$ ms	175 ms, $p < 0.012$

Note. Durations of effects are reported for all p-values < 0.05. All p-values are uncorrected for multiple comparisons.

When conducting GLM (2) with high-amplitude stimuli only, significant MMN-effects in two different time windows become apparent (see Table 3.4). The early MMN-effect is shown in Figure 3.10 and exhibited by electrodes C6 and CP6, while Figure 3.12 shows the more frontal later mismatch effect of electrodes Fz, F2 and FC2 around 175 ms post-stimulus.

To sum up, all mismatch-effects are very small and do not hold up to multiple-comparison correction. This is why all results presented are uncorrected for multiple comparisons. Possible reasons for small small mismatch-effects are discussed in Section 3.4. Nonetheless, our uncorrected mismatch effects replicate the spatiotemporal outlay found by Ostwald et al. (2012) as well as Spackman et al. (2010).

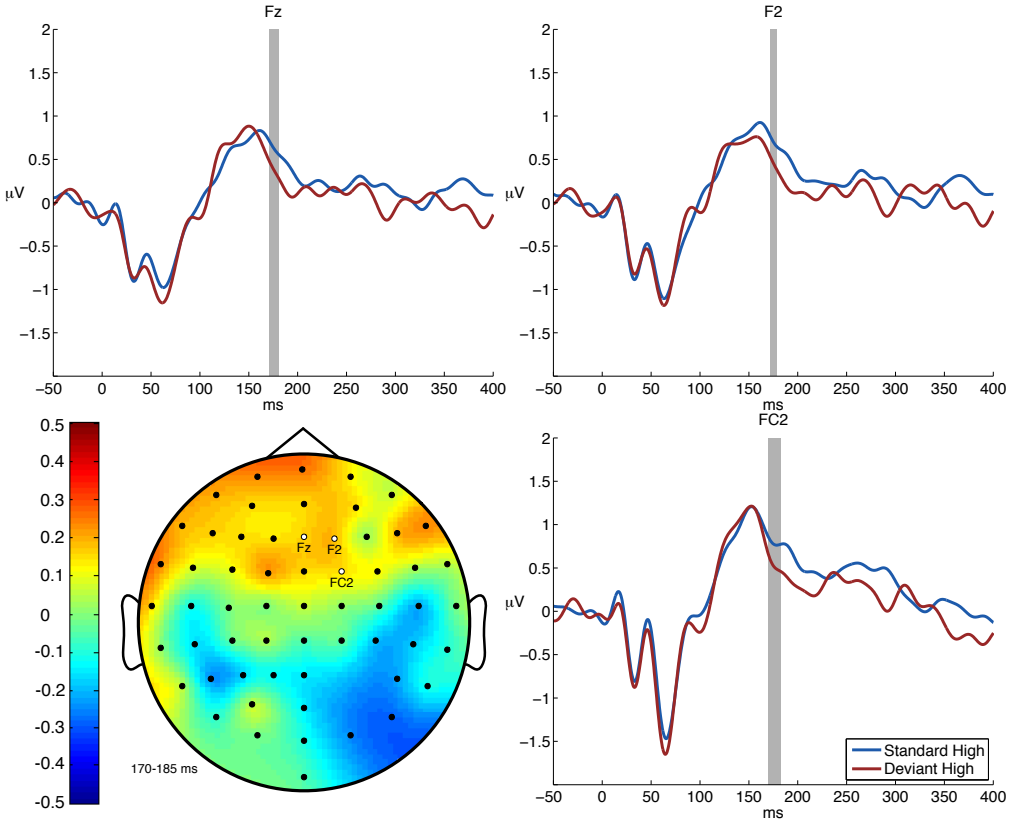

Figure 3.12: Late Mismatch effect. Upper two and lower right panel: Time-courses of electrodes showing significant differences ($p < 0.05$, uncorrected, indicated by gray bars) between high amplitude standard and deviant stimuli. Lower left panel: Scalp topography of Standard - Deviant difference waveform.

The explorative analysis of GLM (3) did not yield significant effects for unexpected repetitions. Specifically, no "reverse-MMN-effect" after repetitions, analogous to an MMN after a stimulus-change, could be observed in our data.

3.3.2 Single-Trial Analysis

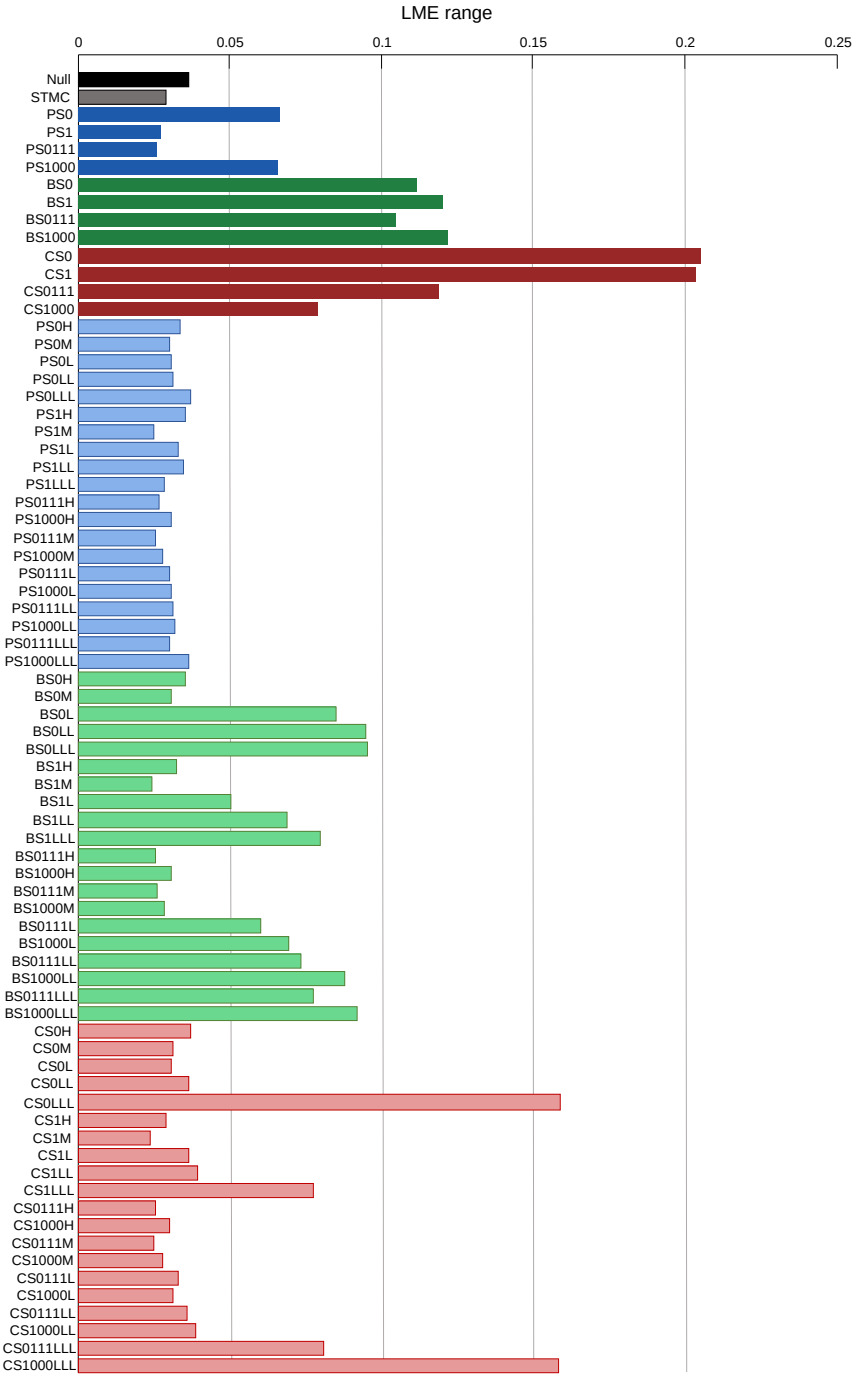

Figure 3.13: Log model evidence range of the group sum for all SBL models and surprise functions over electrodes of interest (difference between minimal and maximal LME). Predictive surprise models are shown in blue, Bayesian surprise in green and confidence-corrected surprise in red. Bars in dark colors refer to conjugate, lighter ones to Gaussian random walk models. Note that the stimulus-change-model (STMC, second bar) has a lower range than the Null-model.

In addition to usual mismatch effects detectable in averaged ERPs, we asked which SBL models and surprise functions could contribute to these effects. For our PEB single-trial analysis we set up 74 regressors (c.f. Table 3.2) as combinations of SBL models and surprise functions (as well as a Null-model and a stimulus-change regressor). First, to look at LMEs on a broad scale across all 74 regressors, we calculated the LME range as the difference between the maximal and minimal LME per regressor within electrodes of interest. The results are visualized in Figure 3.13, showing a clear superiority of conjugate models (dark blue/green/red) as opposed to GRW models. Confidence-corrected surprise shows the largest range in conjugate as well as Gaussian SBL models. Moreover, the stimulus-change model did not perform better than the Null-model, excluding it from further consideration. Among GRW models, those with the lowest variance parameter (0.05) show largest LME ranges.

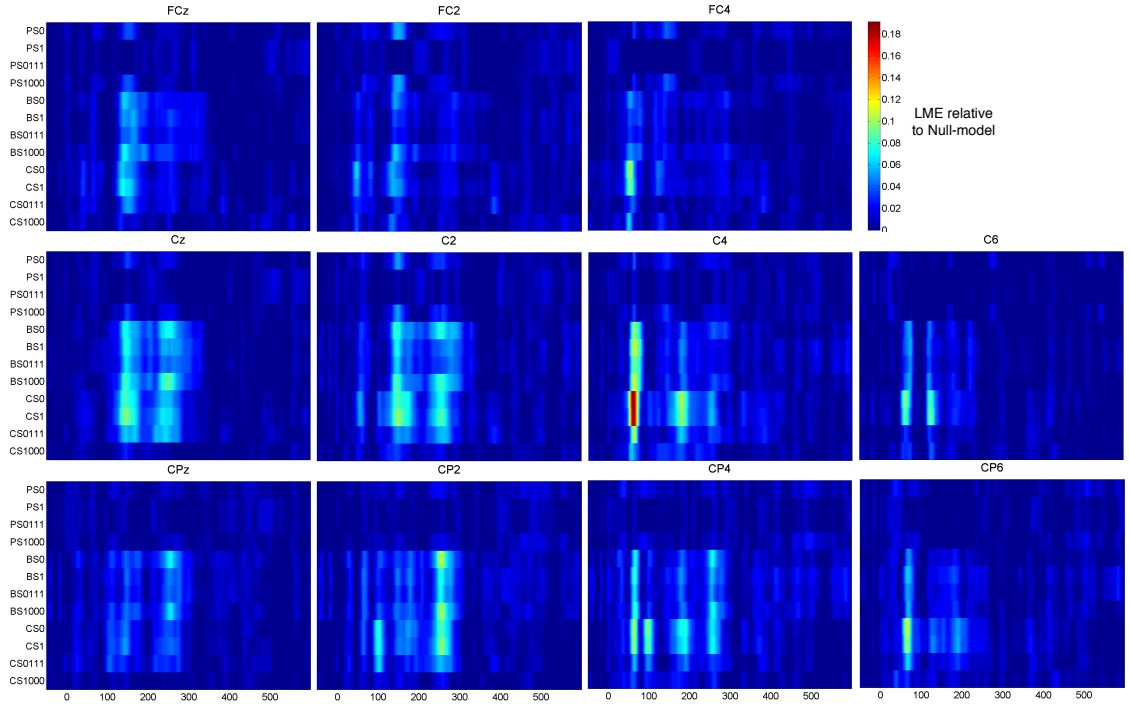

Figure 3.14: Log model evidence values relative to the null-model summed up over subjects for all regressors (y-axis), electrodes of interest and across the whole trial (x-axis). LME values are color-coded with lowest values in blue and highest values in red.

Summed up LMEs for the remaining 12 regressors as a result of the `spm_PEB` function and relative to the Null-model are shown in Figures 3.14 for electrodes of interest (without the most frontal electrodes FZ, F2 and F4, since their relative LMEs were much lower). Highest relative LME values and thus the best surprise regressors can be observed on electrode C4 around 60 ms

with CS0, closely followed by CS1. While a 60 ms peak post stimulus is very early for a regressor to be picking up MMN-like characteristics, other electrodes have LME peaks more in the MMN time window, while the overall LME peak is lower for them (e.g., Cz for regressor CS0 at 144 ms). On C4, the 60 ms peak is followed by a weaker one for CS0 and CS1 at 179 ms. In general, surprise regressors for zero- or first-order SBL models alone gain higher LMEs than mixture models. All relative LMEs were very low, with values ranging to 0.19 (comparable studies using PEB on EEG-data for MMN like Ostwald et al., 2012, have found relative LMEs from $200 - 300$). Thus, even though the pattern looks realistic and not noisy on the descriptive level, possible effects are extremely small if present at all. To test this, we performed Bayesian model selection on the presented relative LME values for each regressor, timebin and electrode of interest.

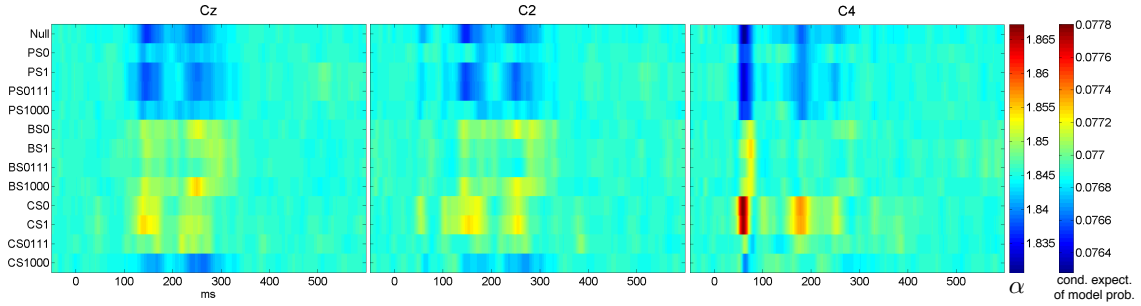

Figure 3.15: BMS results as obtained by the `spm_BMS` function on relative LME values for electrodes Cz, C2, and C4. Color-coded results can be read as either α values to be interpreted as model-occurrences in the sample of subjects, or as conditional expectation of model probabilities (with corresponding color-bars on the right).

Results of the Bayesian model selection for strongest LME-electrodes Cz, C2, and C4 are shown in Figure 3.15. While the α values estimating the number of occurences of a model in a sample of participants are reflecting the basic picture of the LME values, they are rather low and show little differences (ranging from 1.83 to 1.87). Thus, there is no model that can represent data with a similar spatio-temporal pattern in at least two subjects. Similarly, the conditional expectations of model probabilities have a quite low maximum and narrow range (from 0.0763 to 0.0778), while showing the same pattern as the α values. The expectation under the posterior mirrors the pattern of α-values, and similarly, shows a very narrow range, indicating that the data are inconclusive. The exceedance probabilities underscore this picture with failing to provide higher probabilities for peak differences. Instead, they show an overall quite noisy pattern which can be observed in Figure 3.16. Thus, exceedance probabilities do not reveal a clear guide to model selection, meaning

that our results from the single-trial analysis do not directly support the preference of any SBL model and surprise function over another.

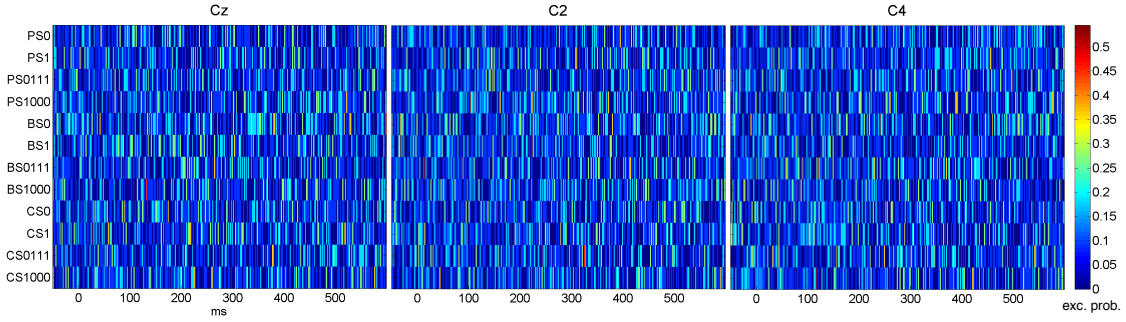

Figure 3.16: Exceedance probabilities for electrodes Cz, C2, and C4 as a function of different surprise regressors and peri-stimulus time bin.

Our single trial analysis relates two a two-factorial model space, namely the kind of SBL and the surprise function applied to distributions of the SBL. Thus, we performed model family comparison with the `spm_compare_families` function for the different surprise functions (Bayesian surprise, predictive surprise and confidence-corrected surprise) on the one hand, and the SBL-models (BB SP, BB TP, and the two mixture models) on the other hand. In terms of the posterior probability for surprise functions, confidence-corrected surprise performs best across electrodes of interest, while Bayesian surprise shows a more sustained activity at electrodes Cz and C2. In the SBL model family comparison, posterior probabilities are lower overall, and not consistently able to distinguish between families. Consequently, while confidence-corrected surprise, and partly Bayesian surprise seem to support the trial-by-trial based theory of the MMN. Nevertheless, we cannot state that either form of surprise or SBL model explain MMN amplitudes specifically in our roving-like paradigm without the participants' attention to the stimuli.

To sum up, in our ERP analyses we could show that we successfully elicited SEPs with median nerve stimulation and replicated an sMMN with our roving paradigm in the absence of attention for high-amplitude stimuli which was located at C6 and CP6 around 125 ms and at Fz,F2 and FC2 around 176 ms post stimulus. However, we did not find regime differences in the ERP in general or in the MMN specifically. Also, we did not find and interaction of regime differences $\times$ TP condition, meaning that even with larger regime differences, those did not have an effect on the SEPs.

When categorizing events into standards and deviants, we did not find a general deviance effect across high- and low-amplitude stimuli. Nevertheless, when analyzing only high amplitude stimuli,

we found an early mismatch effect after 110-130 ms over electrodes C6 and CP6, as well as a later effect after 170-180 ms over Fz, F2, and FC2. No MMN effects were found for unexpected repetitions.

In the single-trial analysis, GRW models in general did not perform as well as BB SP and BB TP. Overall, relative LME values were extremely low and peaked for $CS(o_t)$ of the BB SP on C4 at 60 ms, followed by a much lower peak 144ms on Cz. These descriptive results could not be backed-up statistically by Bayesian model comparison and model family comparison.

3.4 Discussion

The present study using a novel Markov-chain roving-like somatosensory stimulation paradigm sought at not only replicating a somatosensory MMN for the first time in a roving paradigm without having participants attend to the stimuli, but also to further investigate how the MMN responds to changing transitional probabilities in a sequence and what Bayesian learning mechanism might underlie its occurrence. Here, we have shown evidence for the existence of a somatosensory MMN in a roving paradigm and without attention. More complex effects of transitional probabilities on the MMN could not be found (e.g., an MMN when stimulus repeat was unexpected, as could be hypothesized in fast regime periods). We show first indications for confidence-corrected surprise of a BB SBL weakly modulating very early responses around 60 ms, but also later more broad amplitudes at 144 ms post stimulus.

The present study presents first evidence for an attention-free sMMN component in a roving mismatch paradigm. Previously, several studies (Kekoni et al., 1997; Shinozaki et al., 1998; Restuccia et al., 2007; Spackman et al., 2010) showed sMMN effects with oddball paradigms in the absence of attention to the stimuli, while Ostwald et al. (2012) showed widespread mismatch effects in a roving paradigm while participants counted the number of stimulus-amplitude changes. Thus, here we provide a limited instance for the MMN being modality- and the sMMN being paradigm-independent.

Earlier than the usual MMN time windows, we found a small deviance-effect of heightened P50 for deviant compared to standard stimuli on electrode C4. Other studies have found similarly early deviance-effects with somatosensory oddball- (Desmedt et al., 1984; Götz et al., 2011; Mima et al., 1998) or roving-detection tasks (Ostwald et al., 2012), while the present study shows this effect in a task free roving stimulation. This finding is supporting the idea of an early memory mechanism

in somatosensory cortex (Mima et al., 1998).

Mismatch effects belonging to the MMN time frame were only observed in high-amplitude stimuli and notably small in size (c.f., no family-wise error correction for significance statements) compared to studies with either oddball paradigms or a task set related to the stimuli. The auditory MMN increases with attention as well (Auksztulewicz and Friston, 2015), while appearing to be largely paradigm-independent (Phillips et al., 2016). A likely explanation for our sMMN being limited to high-amplitude stimuli comes from previous studies: Spackman et al. (2010) measured mismatch-responses from electrocorticography in an oddball paradigm with stimuli of different durations. Similar to our high-amplitude-only sMMN effect, their long-duration standard-to-deviant comparisons had much larger effect sizes than those for shorter stimulus durations, with a similar spatiotemporal layout of MMN effects (earlier response in S1 and later in frontal regions such as mediofrontal gyrus).

Possible reasons for an attenuated MMN in the somatosensory compared to the auditory domain could lie in a weaker thalamo-cortical coupling for unattended tactile stimuli (Campo et al., 2018). Another important aspect for MMN effect-size is the difference in stimulus features. For the auditory domain, Schröger (1996) finds larger MMN effects for vaster stimulus-differences. Perhaps median nerve stimulation in two different stimulation amplitudes below motor threshold does not offer a spectrum broad enough for a reliable automatic stimulus-change detection.

In summary, our findings put restraints on the notion of a modality- and paradigm-independent MMN. While we do find mismatch effects in our paradigm, they are notably weaker than those in other modalities (auditory, visual) and paradigms (oddball).

Contrary to our hypotheses, we did not find an effect of a slow- or fast-changing regime on SEPs in any of our analyses, and no interaction of regime $\times$ deviance or regime $\times$ TP-condition. In addition, there was no MMN to unexpected repetitions during fast regimes in any of the TP-conditions. In the auditory domain, numerous studies have shown such an MMN-effect for unexpected repetitions (e.g., Nordby et al., 1988; Alain et al., 1994; Horváth et al., 2001).

Furthermore, these null-findings seem at odds with results from the local-global paradigm (Bekinschtein et al., 2009), where the expectation of a stimulus-change is violated, eliciting a later, more frontal negativity together with a posterior-parietal positivity. Originally, we assumed fast regimes to build up an expectation of stimulus change that would be stronger for larger TP-differences (i.e., weak in condition B and strongest in condition D). This could have lead to the

explored unexpected-repetition-MMN, similar to the global-rule violation in Bekinschtein et al. (2009).

While attentional effects might have contributed to the global MMN-effect of Bekinschtein et al. (2009) shown predominantly in a condition where participants counted global rule violations, other studies specifically deflected participants' attention away from the stimuli. However, the unexpected-repetition MMN-effect was found to be larger with larger tone differences and longer ISI (Alain et al., 1994), and most studies have a 10% or smaller rate of unexpected stimuli, as well as a rule-initialization phase in the beginning without any rule violations. Thus, if such automatic effects exist in the somatosensory domain, they might be found with increased stimulus-intensity differences, a longer ISI and a (deterministic) phase of rule establishment at the beginning of a sequence.

Our ERP-analysis did reveal a parametric effect of TP-condition on sMMN-size, being largest in TP-condition D and smallest in A. In the auditory domain, MMN-effects have shown to increase with the number of standard-repetitions before a deviant (Näätänen, 1992; Haenschel et al., 2005). In our experiment, slow regimes most likely contain longer stimulus trains with larger TP differences, since they were generated with higher probabilities for stimulus repetitions. Consequently, stimulus-train length is most likely the factor to have mediated this parametric effect of increased MMN-size, which has not been shown before in the somatosensory domain.

In our single-trial analysis, we provide a first wide-spread test of applying different SBL forms as well as predictive, Bayesian, and confidence-corrected surprise to time courses of EEG data. While Kolossa et al. (2015) compared predictive and Bayesian surprise components within the P300 complex in higher-cognitive statistical inference, to our knowledge, different surprise forms of Bayesian learning have not been compared for the MMN in any domain yet.

For the auditory domain, Lieder et al. (2013b) model the MMN as precision-weighted prediction error from a hierarchical dynamic model with generalized Bayesian filtering, while their empirical model comparison revealed model adjustment theories of the free-energy principle to be best at explaining MMN amplitude in a single-trial analysis (2013a). Although some of their compared models appear to be similar to our predictive surprise formulations (c.f. models *Novelty 1* and *2*), no comparison of different surprise functions and Bayesian learning mechanisms was conducted. Here, we provide a novel account of such a broad single-trial analysis in an MMN-paradigm.

However, our results have to be interpreted with caution. While, on a descriptive level, we

find confidence-corrected surprise of the BB SP as well as BB TP to modulate electrophysiological responses around 60 ms post stimulus at electrode C4, and later weaker modulations (regressor CS0 on electrode Cz at 144 ms, C4 at 179 ms), Bayesian model selection did not yield any model to more likely underlie our EEG data than any other. This is most likely due to the very low relative LME values derived from the PEB analysis approach (tenths in our data compared to hundreds in Ostwald et al., 2012). One key difference in analyses is that in employing PEB, Ostwald et al. (2012) used input vector lengths of 500 trials, while we had about 9000 trials per analysis per subject. This was necessary due to high noise levels in the data, but might have impeded PEB analysis on the other end, with predictions over thousands of trials being to specific to return higher LME values.

Although it is conceivable that our single-trial analysis results show a null-effect, with descriptive peaks perhaps being meaningless, it is in fact interesting that the peak LME effect on electrode C4 for confidence-corrected surprise is mirrored in the ERP analysis with a deviance-effect in high- and low-amplitude stimuli. In the absence of attention, one can speculate that an earlier surprise system takes over, suppressing later stimulus- and thus surprise-processing. Contrasting the fact that we have highest relative LME values for static BB SP and BB TP models, Ostwald et al. (2012) showed Bayesian Surprise in a BB SP model with implemented exponential forgetting to be superior. There, participants counted stimulus changes, and thus top-down attention might have lead to involvemend of a memory-trace for past stimuli with said exponential decay. Such effects were absent in our study, making (early) static Bayseian learning the more likely processing principle. In addition, we replicate the finding by Ostwald et al. of excluding a very simple stimulus-change-detection system to be responsible for sMMN, as the stimulus-change regressor did not yield higher LME values than the Null-model.

In summary, however, our single-trial analysis does not lead to conclusive results about the employed Bayesian learning mechanism, while our data weakly point into the direction of confidence-corrected surprise of BB SP and BB TP models. Since the measure of confidence-corrected surprise has only been published very recently, there have not yet been other applications to automatic perceptual stimulus processing to compare our results with. More research with a broader variety of stimulation paradigms (i.e., for example patterned local-global paradigms, or our Markov-chain roving-like paradigm in the auditory modality) could provide more insight into incidences of confidence-corrected suprise in neural signals.

Generally, the MMN is an electrophysiological effect of great benefit for clinical use, showing differences in schizophrenia patients vs. healthy controls (Umbricht and Krljes, 2005) and possibility to predict recovery from coma (e.g., Kane et al., 1993), which has been extended with the local-global paradigm (Bekinschtein et al., 2009). A better understanding of sMMN could extend these clinical benefits to patients impaired hearing or vision, as well as provide useful in understanding the functional extent of cerebellar damage (Restuccia et al., 2007).

Nonetheless, differences to the auditory system have to be noted. Our study provides a strong account for the disparities between auditory and somatosensory automatic information processing and rule abstraction. The sMMN appears to not be paradigm-independent (as is claimed in the auditory domain, see Phillips et al., 2016), which is evidenced by our very small effect size in an attention-free roving paradigm while oddball-findings have been quite consistent (Kekoni et al., 1997; Restuccia et al., 2007; Spackman et al., 2007, 2010). In addition, the sMMN does not seem to pick up slightly more complex rule violations such as unexpected repetitions. All in all, changes in TPs did not influence SEP-amplitudes, while we did find larger TP-differences to be related to larger sMMN. Possibly an attenuated feed-forward processing already at the level of thalamocortical connections during passive tactile perception (Campo et al., 2018) is the underlying cause of a weaker somatosensory system for automatic rule extraction.

In our Bayesian model-based single-trial analysis, we show descriptive results for a BB Bayesian learner and confidence-corrected surprise modulating electrophysiological responses over somatosensory cortex, while our experimental data and setup have shown unfit to make claims on a Bayesian model for MMN. Future research is needed to disentangle attentional and automatic processes in somatosensory perceptual Bayesian learning and identify the extent of a modality-independent system of the human brain for perceptual learning.

In conclusion, while our data support the existence of an sMMN in roving paradigms, we call the presumed modality-independence of the MMN into question. The somatosensory analogue to the auditory MMN - strictly defined as preattentive (c.f. Näätänen, 1992) - shows far more susceptible to changes in paradigm (being only robustly found with oddball setups) and no sensitivity to more complex rule violations such as unexpected repetitions. As a proof-of-concept, we tested possible Bayesian learner models and surprise functions for underlying mechanisms. Their relevance for different MMN-manifestations remains to to be shown by future research in more detail.

Chapter 4

Discussion

In this thesis, I formulated a broad computational model space of surprise during sequential Bayesian learning in Chapter 2 and consequently empirically tested a subset of these models on EEG data from a somatosensory mismatch negativity (sMMN) experiment designed to examine a possible complex system of surprise in Chapter 3. The EEG data revealed a small sMMN, while neither providing evidence for more complex MMN-effects known from other modalities, nor showing support for a computational surprise measure contributing to trial-by-trial changes in EEG amplitude. Possible reasons for small effects and null-findings are discussed in detail in Section 3.4.

The present Chapter discusses the overall significance of computational models in the research field of neural surprise, with an emphasis on the somatosensory modality and the MMN. Furthermore, it relates the work presented in Chapters 2 and 3 to the existing scientific literature, while considering our theoretical and empirical results and open research questions.

4.1 Furthering Computational Modeling of Bayesian Learning

Computational models are the backbone of current ways to understand the brain. Most influential theories and frameworks in cognitive neuroscience have their roots in computational principles, such as the free-energy principle (FEP, Friston, 2005). Marr sees the computational model as defining the overall goal of neural processing or any information-processing machine's activity

(1982). In that way, a good computational model can carry relevance in itself, by making the goal of an information processor explicit. Through this explication it has an additional value to, for example, deep neural networks (such as the kind proposed by Heeger, 2017), that mainly function on the algorithmic and implementational level. Critically, Marr states that the computational theory is of crucial importance at the perceptual level, since he assumes there is more to be gained from understanding the computational problems that perception has to solve (which are, as is commonly agreed upon, most aptly to be cast in Bayesian terms; c.f., Knill and Pouget, 2004) than from inspecting the "hardware" of perception (Marr, 1982, p. 27). Thus, from the angle of a sound computational theory, an algorithmic formulation and implementational mechanism will follow.

The computation of a surprise quantity that brain responses exhibit might further Bayesian learner models to a more algorithmic characterization, in that a given area in the brain is hypothesized to compute or approximate this form of surprise. In this thesis, I used computational modeling of Bayesian learning and surprise as a principled way to reevaluate the traditional EEG effect of the MMN. On the same note, I tested the generalizability of Bayesian learning in the MMN by investigating the somatosensory perceptual domain. To do so, in Chapter 2 I structured the model space of Bayesian learners for 2-item sequences into (i), sequence feature (the considered hidden state, i.e., SP, AP, or TP), (ii), distribution family, (iii), volatility or implemented forgetting, and (iv), surprise function. In the following, I first discuss findings from Chapter 2 in relation to Bayesian modeling by Meyniel et al. (2016) as well as by Faraji et al. (2018) in light of my findings and theoretical model structure. Further, I view how model properties (i)-(iv) relate to each other and regard other factors possibly influencing a Bayesian model.

Work from the literature closest to Chapter 2 is most likely the recent outline of SBL models by Meyniel et al. (2016). In summary, their work makes an excellent case for the careful construction and investigation of an SBL in modeling human perception: Their study features a "local transition probability model", which is local in the sense that it mainly considers the most recent observations. The resulting model is equivalent to our Beta-Bernoulli transitional probability (BB TP) model with exponential forgetting (which Meyniel et al. call "leaky integration"). In the article, Meyniel et al. comprehensively compare their TP model to models of item- and alternation probability as well as a dynamic-state model family (very similar to the Gaussian random walk models in 2.3.2) with a parameter for assumed volatility.

On the theoretical side, a very convincing case for the learning of TPs is the asymmetrical perception of randomness, meaning that sequences with *more* than 50% alternations are usually perceived as "more random" than truly unbiased sequences. Meyniel et al. explain how a Bayesian TP learner model can account for this asymmetry: While for repeating stimuli only one contingency is learned, alternating sequences with the same amount of stimuli and the same alternation-bias, i.e., non-randomness, distribute the evidence onto two contingencies. Consider the following exapmple: In a sequence $o_{1:7}$ of 1111111 we only learn $p(o_t|o_{t-1} = 1))$, while alternations allow for learning of both contingencies (0101010 allows for $p(o_t|o_{t-1} = 1)$ and $p(o_t|o_{t-1} = 0)$. In both sequences, the observer is lead to predict $o_8 = 1$, however, this prediction is a lot stronger for the first sequence than for the second one. This is because the first is based on 6 transitions of $o_t = 1|o_{t-1} = 1$, while the second sequence provides only 3 transitions of $o_t = 1|o_{t-1} = 0$. Thus, the Bayesian learning of TPs may lead to the described asymmetrical perception of randomness by yielding a lower-entropy distribution $p(o_t = 1|o_{t-1})$ for the 100% repetition sequence, while the entropy for $p(o_t = 1|o_{t-1} = 0)$ in the 100% alternation sequence remains much higher.

Moreover, Meyniel et al. test their TP model on P300 data and show that it is in all cases an equally good or better explanation, and if equally good, a more sparse one. Thus, they convincingly show that humans are very much attuned to learning TPs in a Bayesian way.

While Meyniel et al. pointed out the suitability of the TP model to the MMN as well, they did not present any validation of the model with MMN data. Here, one might wonder whether an application was tried unsuccessfully, which would be particularly interesting for the subject of this thesis and will be further discussed in Section 4.2. In addition, the authors restricted themselves to the surprise formulation of PS, neglecting the possibility that BS could have been a more suitable form of surprise in some data. In this thesis, I show and apply SBL models with three different surprise functions, with the aim to further complete the picture of Bayesian learning in two-item sequences.

The recent article by Faraji et al. (2018) greatly influenced the work in this thesis on modeling surprise responses. Their notion of a confidence-corrected surprise rests on the intuition that an *unlikely* event does not necessarily constitute an *unexpected* event, and that an event which is with all its specific details very unlikely (e.g., seeing a grey cat with white spots crossing from left to right at 7:34 a.m.) should not be surprising if no particular event in that range has been expected in the first place (and hence, no commitment to an expectation was present, c.f., Schmidhuber,

2002). Thus, they correct for the confidence in a model by calculating the KLD between the current prior and the posterior under a naïve prior (c.f., Equation (1.8)). However, in the BB case (see 2.3.1.1), this leads to zero surprise (and thus a full correction for lacking model-confidence) in $t = 2$, and not, as it ideally should be, in $t = 1$ (which is the trial with the least model confidence in a prior because the prior is flat).

Faraji et al. also formulate an iterative updating scheme (the SMiLe rule, p. 46), which approximates Bayesian learning, but does not require solving complicated integrals (in that way, it is an explanation of Marr's algorithmic level). In addition, it realizes the important feature of dynamically increasing the uncertainty of a hypothesis in light of highly surprising data. This aspect is missing in non-hierarchical SBL models, where introducing exponential forgetting can only function in a static manner that does not depend on the specific amount of surprise that was recently experienced.

In Section 2.3, I applied Faraji et al.'s definition of CS to exact (i.e. BB) and approximate (i.e., Gaussian) SBL models. In the inspection of example surprise regressors built from those models and the surprise function, I note that CS in a BB model can lead to counter-intuitive courses of surprise regressors, especially with implemented exponential forgetting. Here, CS increases throughout the whole sequence because of its additive model-commitment $C(o_t)$ component that increases with each observation, and the fact that the model will never make an accurate prediction because every possible observation ($o_t = 1$ or 0) will diverge from the model. If strong empirical evidence is found which supports this measure of surprise in a BB-model, this will be a very convincing case for CS because it is a course that no other surprise measure predicts. Chapter 3 shows the first application of CS with EEG-data, and although the results are not significant, CS of a Bayes-optimal learner descriptively shows highest LME values in the sMMN study, compared to PS and BS from Bayes optimal and Gaussian SBL models.

The analysis of specific surprise regressors in Section 2.4 revealed several more noteworthy aspects of the various definitions of SBL and surprise function that have, to my knowledge, not been previously addressed in the literature. For instance, in surprise regressors of Gaussian SBLs with high variance parameters, i.e. assumptions of high volatility in the environment, surprise quantities of PS, BS, and CS converge toward each other. Thus, under the assumption that the world changes constantly, the specific surprise function will not have much of an impact on the estimated amount of surprise.

Furthermore, by the prediction-error formulation in Equation (2.20) a relationship between Gaussian and BB models can be drawn. With this relationship, it becomes apparent, that while GRW models employ a static variance parameter that governs the impact of new observations (sometimes called "learning rate"), static BB models implicitly weigh the current observation inversely to the number of events observed so far (i.e., to t). Accordingly, stimuli presented earlier in a sequence have a much stronger impact on the belief distribution than later ones.

However, it should be noted that there are competing SBL models that have not been applied in Chapters 2 and 3, e.g. hierarchical Gaussian filter (HGF) models by Mathys et al. (2011; 2014). While it is not yet clear how the brain would invert such a complex hierarchical structure for inference (Faraji et al., 2018, p. 35), these models have been fairly successful in explaining MMN amplitude variations (Lieder et al., 2013a; Stefanics et al., 2018). However, applications of HGF to the MMN generally use the second level of the hierarchy and do not make use of the implementation of possible inter-individual differences in response behavior (because there are no responses in MMN-paradigms). In that way, they differ from the GRW models in 2.3.2 only in the method employed for solving integrals (i.e., variational Bayes vs. numerical integration).

Finally, it is important to point out that the effectiveness of any Bayesian model in reducing surprise and making accurate assumptions about the world depends greatly on the kind of environment it is subjected to. That is why future work should apply the models presented in 2.3 to sequences with different statistical generative models than our hierarchical Markov-chain (see 2.2.2) and analyze their robustness in light of different environmental conditions.

4.2 Bayesian Learning in Electrophysiological Potentials

In this Section, I will review the rationale and results from the experiment in Chapter 3 as they relate to a more global theoretical background and the current state of research. Specifically, I discuss the impact of two important factors: attention and perceptual modality, and their interacting effects on the results.

In Chapter 3, I studied the MMN in the somatosensory domain. The overall motivation for the experiment was twofold: (i), obtaining evidence for the modality-independence of the MMN itself, and (ii), test if the assumption that the MMN is created through surprise during Bayesian learning, is generalizable to the somatosensory domain.

Whereas previous studies have shown the existence of the sMMN in an oddball-paradigm that

featured uneven item-frequencies (usually $10 - 20$ % oddball stimuli, e.g., Kekoni et al., 1997; Restuccia et al., 2007; Spackman et al., 2010), I set out to test the MMN in a roving paradigm of even item-frequency. Roving paradigms decorrelate possible adaptation-effects from the role of the stimulus as standard or deviant, and are thus crucial for establishing a truly expectation-based MMN (Heilbron and Chait, 2017). So far, in somatosensory roving paradigms, participants have always been asked to attend to the stimuli and detect or count the changes in stimulus feature in a sequence (Ostwald et al., 2012; Allen et al., 2016). This cannot yield a pure MMN because of possible attentional confounds within the ERP curve (Näätänen, 1992). Since the MMN is also assumed to be paradigm independent (Phillips et al., 2016), a *true* sMMN should also be measurable in a roving paradigm without an explicit call for attention to the stimuli.

Bayesian learning has so far been shown for the auditory and visual MMN (Lieder et al., 2013a; Stefanics et al., 2018), as well as for a somatosensory MMN-like response, in which attentional effects cannot be ruled out (Ostwald et al., 2012). Here, evidence for Bayesian learning in an attention-free sMMN would provide strong evidence for a modality-independent Bayesian mechanism in the brain that is invoked at fairly low perceptual levels and is independent of top-down attentional influence.

The results of the experiment from Chapter 3 can be briefly summarized as follows: An ERP analysis yielded a small sMMN effect that was only significant without family-wise error correction, but tested in a spatio-temporal region-of-interest (ROI, see Figure 3.4 for electrodes of interest) using previous findings and in agreement with those. In addition, the single-trial analysis did not identify any significant SBL- and surprise-model to underlie the MMN, while CS from a BB model for item frequency and TP showed strongest explanatory value in the spatial ROI. Furthermore, a simple change-detection model (regressor was 0 when $o_t = o_{t-1}$ and 1 when $o_t \neq o_{t-1}$) did not explain the data at all on a descriptive level (en par with the null-model), as was the case in other applications of Bayesian learner models to sequential perception (Ostwald et al., 2012; Lieder et al., 2013a; Stefanics et al., 2018). For detailed results, see Section 3.3.

While the paradigm from Chapter 3 with positive results would have been powerful in emphasizing a supposed modality independence of the MMN, it is not particularly well-suited for identifying the reasons for negative results, since several factors have been varied simultaneously in the experimental design compared to previous studies. Thus, in the following, I briefly speculate on possible interactions of factors that lead to a small sMMN-effect. Subsequently, I move on to

the main theme of this work by applying those factors to the descriptive but not significant support for CS in BB Bayesian learning in the sMMN.

A preliminary note should be made before discussing the findings: Because the different factors influencing the experimental results are most likely highly interconnected, here, the same themes, such as attention, will be addressed repeatedly. More specifically, it is plausible that the small sMMN effect and the null-finding in the single-trial analysis hinge on each other, and thus, that a stronger ERP effect would have entailed a better basis to find a well fitting SBL model (presuming that the model space included a model that parallels the computations that the brain carries out). As with two sides of the same coin, the two kinds of results might consequently be influenced by the same factors.

4.2.1 Factors Influencing ERP Results

There are three factors that presumably worked together and lead to a small sMMN-effect in our roving paradigm: perceptual modality, attention, and type of paradigm. Here, I focus on each of these three factors while referring to interactions between them.

While the auditory system comprises the best-studied perceptual modality in MMN-research, less is known about the visual MMN (Stefanics et al., 2014), with only a handful of studies providing evidence for a somatosensory analogue. Both auditory and visual MMN studies can replicate the preattentive effect in a roving paradigm (Garrido et al., 2008; Czigler and Pató, 2009), however, the only EEG-study investigating sequential perception with somatosensory stimuli imposed an alternation-counting task on the subjects (Ostwald et al., 2012) and thus cannot investigate effects that are unaffected by top-down attention. A roving paradigm requires abstracting rare events, whereas the rarity in the oddball stimuli is bound to the respective stimulus features, a roving paradigm yields rare events that consist of stimulus-changes, where stimuli of any feature can take on the role of deviant (c.f. Figure 2.1b). Since the visual domain features only few roving paradigm studies (Stefanics et al., 2014, p. 5), with one study finding no visual MMN at all in a roving paradigm (Sulykos et al., 2013), it is plausible to assume that the auditory system is especially attuned to changes in sequential input and able to abstract statistical properties from stimulus features at a preattentive stage. Possibly, the fact that information in the auditory domain is alway conveyed over time (extracting pitch information relies on *frequencies* of auditory waveforms that extend in time), lead to the preferential processing of sequential information in the

hierarchy of the auditory system. Barascud et al. (2016) reason that locomotion sounds arising from living organisms in the environment are often regular and repetitive. Thus, it could be highly adaptive for an organism to evolve to be attuned to such regularities in auditory input and their violations in order to detect and locate other living organisms.

Another modality-specific information processing factor could consist of a task-set dependency in thalamocortical feedforward sweep in the somatosensory system. Campo et al. (2018) have shown the attenuation of somatosensory processing without a task-set in non-human primates. Single-cell activity in thalamus and SI of corresponding receptive fields during a vibrotactile detection task yielded significantly stronger feedforward connections from thalamus to SI than trials of passive perception. Perhaps this reduction in thalamocortical feedforward connections during the absence of a stimulus-dependent task is particular to the somatosensory system. As such it could have contributed to the small sMMN in the roving paradigm without top-down attention. Stronger mismatch responses to oddball stimulation without attention could, on the other hand, be influenced by stimulus-specific adaptation.

Furthermore, the particular paradigm used in Chapter 3 might have been unable to uncover the full range of sequential information processing in the somatosensory system. This paradigm had several specificities designed to detect other probabilistic information-processing features that are found in the auditory system. Specifically, the paradigm alternated with a probability of 0.01 between a slow and fast regime. This means that its TP structure changed between that of a standard roving paradigm with rare transitions between stimuli (slow) to that of rare repetitions (fast), purportedly creating a low expectation for repetitions (see Figure 3.5b). In the auditory domain, unexpected repetitions elicit an MMN-response (e.g., Horváth et al., 2001). However, other MMN studies with similarly abstract rules employ an initial "grounding" phase in which no rules are violated, to build up a certain expectation (Mullens et al., 2016; Wacongne et al., 2011). Wacongne et al. (2012) even use such a grounding phase to train their neural network for the MMN. Thus, the fact that our sequences were always probabilistically defined and lacked the initial establishment of an exception-free rule (e.g., 50 trials of stimulus-alternations as a rule for the fast regime) could have hindered the creation of expectations in the fast regime. Notably, the necessity for a grounding phase for fast regimes, but not for the standard slow-regime roving paradigm can be explained theoretically following the observations by Meyniel et al. (2016), which I review below when discussing Bayesian models.

In summary, my findings regarding an sMMN constrain a supposed modality-independent mechanism for the preattentive MMN, as the sMMN seems significantly impacted by paradigm specificities. Nonetheless, it is still possible that the MMN is elicited by a common underlying principle, yet, such a principle should also be able to explain differences between the modalities in their various forms. One such way could be to classify MMN-eliciting phenomena along levels of abstraction and identify the level of cortical hierarchy that a non-attended stimulus in either perceptual domain can reach, linked to the computations that specific area is fit to carry out. Further research is needed to truly cast the MMN in a framework of modality-independence.

4.2.2 Factors Influencing Single-Trial Analysis Results

Findings on the level of computational modeling in Chapter 3 are linked to those of the small sMMN. The sMMN is not well described by sequential Bayesian learner (SBL) models in our data. Notably, I am aware of only two studies which show a version of Bayesian learning in a trial-by-trial fashion in the MMN: Lieder et al. (2013a) in the auditory and Stefanics et al. (2018) in the visual domain. In both, versions of hierarchical Gaussian filter models (Mathys et al., 2011, 2014) were employed to represent MMN amplitude variations per trial.

In the earlier study by Lieder et al. (2013a), several possible computational models of the MMN were applied to EEG data from an auditory roving paradigm with tones in 7 different pitches. Using Bayesian model comparison, Lieder et al. could rule out change-detection and adaptation as computational mechanisms behind the MMN. However, the study could not clarify whether prediction-error or model-adjustment models are more likely on the basis of their data. Furthermore, all Bayesian models tested were Gaussian, and apart from PS (in "novelty detection" models), no other surprise function was applied. Moreover, whereas the same authors published a neurocomputational model of the auditory MMN waveform (Lieder et al., 2013b), I am not aware of an application or test of their waveform model or of the winning models of the model-comparison study (Lieder et al., 2013a). Future MMN research should validate those models using more complex sequences in order to gage the robustness of the hypothesis. Thus, even in the well-studied auditory domain, a computational model for the MMN accounting for all the different effects, and that is also a good candidate on the implementational level, is still lacking (Heilbron and Chait, 2017).

In the visual domain, the recent study by Stefanics et al. (2018) investigated the visual MMN

using hierarchical Gaussian models of SBL and their resulting precision-weighted prediction errors (Mathys et al., 2011, 2014). While the Gaussian models again fared better than change detection, they did not test any other model or surprise function. Hence, the visual domain also needs further scrutiny in establishing a complete computational model for the MMN.

The study presented in Chapter 3 discouraged top-down attention, whereas Ostwald et al. (2012) found BS in a roving paradigm during a task that demanded attention towards the stimuli. Therefore the lack of top-down attention most likely plays a significant role in the null-effect of my single-trial analysis.

It should be noted that the subject of attention itself is a highly controversial one. After more than a century since the seminal work by James (1890), there is still no consensus among researchers regarding its definition, and mechanisms to describe it remain rather heterogeneous (Whiteley and Sahani, 2012). To give attention research a more principled structure, theories within a Bayesian framework have been proposed (Whiteley and Sahani, 2012; Dayan and Zemel, 1999). One possible Bayesian explanation is that unattended stimuli are likely processed with a *static* Bayesian learner, treating them as noise to be filtered out and keeping the amount of surprise awarded to them at a minimum. The strength of such a Bayesian filtering mechanism might vary among perceptual domains. For example, the brain could regard recent visual input as inherently more relevant, even without top-down attention, and consequently employ a mechanism that either maintains some uncertainty (as in the SMiLe-rule by Faraji et al., 2018), or constant exponential down-weighting of events far in the past (as modeled in Ostwald et al., 2012; Meyniel et al., 2015).

Considering the attentional effects, one might wonder whether the MMN is not simply not the most responsive electrophysiological effect for measuring surprise from Bayesian learning. Several studies investigating SBL with a task involving sequential stimulation found considerable effects on the P300, such as during perceptual detection (Meyniel et al., 2016, reanalyzing data from Squires et al., 1976), counting stimulus alternations (Ostwald et al., 2012), and probabilistic inference about hidden states (Kolossa et al., 2015). Particularly the study by Ostwald et al. indicates that Bayesian learning is employed by the somatosensory system, but possibly not without directed attention.

Considering the descriptive findings from the computational single-trial analysis in 3.3.2, the best results are obtained for CS as a model of surprise during the stimulus sequence. A possible explanation for this connects back to the fact that no grounding phases establishing a rule were

presented, leading to considerable uncertainty about future events in all of the presented sequences (even though, one could point out that condition D was the most predictable, condition A the least predictable in terms of TPs). Because CS takes into account the entropy of a prior, leading to less surprise with a less-specific prediction, CS seems to be a good candidate for neural responses to the unstable, unlawful sequences. For MMN-effects that are supposedly elicited by surprise in Bayesian learning, a model committing to a prediction that is being violated is necessary. Perhaps such a committed model could not be established during the rather volatile sequences presented in Chapter 3, which also resulted in comparatively small MMN-effects. Meyniel et al. (2016) hint at another relevant factor regarding the commitment within a model about TPs: Longer streaks of the same stimulus (as they occur in standard roving paradigms, like in the slow regime in Chapter 3) lead to stronger models about TPs than more frequent alternations. The many alternations invoked in the fast regimes could thus have lead to high model-entropy (i.e., low commitment) about TPs and demanded a neural commitment-correction as featured in CS.

Another descriptive result from the single-trial analysis revealed the superiority of static SBL models over those with exponential down-weighting of past events. This finding is at odds with the studies by Ostwald et al. and Meyniel et al., where models discounting events from the distant past yielded highest explanatory power for electrophysiological responses. Furthermore, similar to forgetting- or volatility-parameters, the updating algorithm by Faraji et al. even insures the remainder of a small model uncertainty after longer static periods, thus keeping open the possibility for a quick change in belief after a highly surprising input. This is a necessary mechanism for an adaptive organism in a dynamically changing world. However, those studies feature experimental paradigms where participants were engaged in a task that required top-down attention to the stimuli (e.g., Ostwald et al., 2012). In contrast to exponentially down-weighted Bayesian updating, the "descriptively winning" SBL models without forgetting from Chapter 3 associate longer observation sequences with lesser impact on formed beliefs - as expressed in Equation (2.20). Hence, the question remains, whether the preattentive MMN relies on a model-updating mechanism using a memory trace from past events (which are temporally discounted), or whether model-updating occurs on-line with no memory necessary, but with a reduced possibility of adjusting to a changing environment.

To speculate about reasons for this apparent dichotomy, one can again invoke the concept of attention and its boosting property that might be crucial in the somatosensory domain to form a

memory trace and adjust models to recent input regularities (as in Ostwald et al., 2012). While the auditory system might provide the possibility for a more complex preattentive information-processing on the basis of a memory trace, similar processes in the somatosensory system could require top-down attention toward the stimuli.

In summary, many open questions regarding computational Bayesian modeling of electrophysiological surprise responses remain. As long as the results from Lieder et al. and Stefanics et al. have not been replicated and compared to competing Bayesian SBL models and surprise functions, the controversy regarding underlying mechanisms of the MMN remain unsolved. Simpler sMMN-studies of roving paradigms will also be indispensable for gaining insight into its nature in comparison to the auditory and visual perceptual domain.

4.3 From Computation to Implementation

In this section, I address the theoretical foundations of the work presented in this thesis in light of its limitations, possible implementations and criticism while consulting Marr's levels of analysis (1982).

The work in this thesis focuses strongly on the computational aspects of sequence perception, with some digressions to algorithmic considerations. However, the implementational level as a description of how representations and algorithms are realized physically should not be completely disregarded here as well. In particular, an algorithmic theory of information processing that is impossible to be realized given the structures in the brain will likely need revisions in algorithm. Nonetheless, these revisions need not change the computational goal that was defined for the algorithm, if indeed, it proves to be fitting to computations that must be carried out by the brain. Hence, when criticizing a model, it is crucial to determine the level of analysis on which it is being judged (Marr, 1982, p. 27).

Applied to criticism of the BBH, this means that pointing to intractable integrals in order to discredit the theory of the human brain employing Bayes' rule is mislead, because the BBH merely specifies the computations to be carried out, and not the algorithm that the brain should use to do so. In fact, algorithmic theories for the integration of probability distributions in the brain propose approximations instead of an exact model inversion, like variational Bayes in the FEP (Friston and Stephan, 2007) or Monte-Carlo sampling (Sanborn and Chater, 2016).

Adding to this criticism, Bowers and Davis (2012a) denounce the BBH because they believe

that it implies near-optimality as well as unfalsifiability (because of the possibility to tweak priors and distribution families to be used). Yet, the mere employment of Bayes' rule does not imply optimality at all. Rather, Bayes' optimality refers to the way that a new piece of information is integrated into a current belief distribution. What is optimal, is not necessarily the resulting behavior, but only the way that new input is combined with previous beliefs about the hidden cause of this input. Hence, if a prior belief is far away from reality, despite new input being integrated using Bayes, the resulting beliefs and behavior might be even farther from the truth. Moreover, while the BBH as a framework does not have to meet a falsifiability criterion, every concrete hypothesis derived from the framework as a Bayesian input-output relationship can be proven wrong, if the computational output pattern does not match the respective behavior or brain responses.

Bowers and Davis are right in pointing out that Bayesian hypotheses about the brain entail priors and that there are possibly more free parameters to be specified (such as the ones determined in Chapter 2). However, this aspect is not necessarily an argument against Bayesian computations per se, but more questioning the rigorousness of the researchers testing those hypotheses. To address this concern, one can make use of a standard practice in machine-learning; using separate data-sets for training and testing computational models (c.f., James et al., 2006, p. 26). When priors and other free parameters are fit to one dataset, they can be verified on independent data to check the generalizability of a specific model. Thus, if parameters were only tweaked to fit very special instances, the model will most likely fail when tested on a separate data-set. This would allow generalizable predictions to arise from multiple Bayesian models concerning many levels of cognitive processes, and it can be warranted to assume a common algorithmic process behind them.

As a whole, I do not find the criticism brought against the BBH concerning the framework's usefulness very concerning. Nonetheless, scrutiny in identifying Bayesian models of the brain activity is strongly advisable. Specifically, simply casting cognitive processes into a Bayesian framework might not be meaningful in itself, only in specifying properties of a Bayesian model does it become informative and testable. This is why I think that testing several Bayesian models against each other can be a beneficial approach for identifying an adequate computational model. Then, deriving an algorithmic explanation of how the computations can possibly be carried out is the next step towards an explanation of the mechanism underlying the brain's information processing.

Computational theories describe input-output relationships of information processing systems (Marr, 1982). Yet, Bayesian computations typically involve probability distributions, which are, if at all represented as such in the brain, not directly accessible to the researcher. In the case of sequential Bayesian learning, this is where surprise functions and neural surprise responses come into play. To render hypotheses about SBL models testable, I made assumptions that are not entirely computational in essence: Namely, I proposed that while there is no access to a belief distribution, measurable surprise responses somehow relate to the belief distribution and the current sensory input. Hypotheses about a relationship between the magnitude of a surprise response, input, and belief model are expressed in the different surprise functions. To show that there can be considerable differences in a surprise trajectory depending on the concrete model, I specified them for a wide array of SBL models in Chapter 2. In application cases, again, quantifications of surprise are to be tested against each other to assess their respective fit to the data. Accordingly, hypotheses with a stronger algorithmic emphasis could then be based on the best computational surprise function. In my view, in light of overwhelming psychophysical evidence (e.g., reviewed in Knill and Pouget, 2004) there is no question *if* the brain employs Bayesian inference, but more *when* and *what kind* of Bayesian inference. When those questions are rigorously answered on a computational level, corresponding algorithms to carry out Bayesian computations in the brain can be developed and identified much more easily.

Drawing on the brain's hierarchical architecture (Felleman and Van Essen, 1991), it is plausible to assume that belief distributions and surprise responses can be found at various levels of abstractions in the brain. Hierarchical processing as a way to carry out Bayesian inference with respect to cognition on those different levels relates to the algorithmic level of analysis (Marr, 1982). Applied to cognitive processes, there are countless studies identifying a hierarchy of surprise or prediction error in neural responses (e.g., Dürschmid et al., 2016; Iglesias et al., 2013; Chennu et al., 2013; Wacongne et al., 2011; Phillips et al., 2015).

Algorithmic theories of how the brain could employ Bayesian inference on a hierarchical level have been proposed by George and Hawkins (2009) as well as Heeger (2017). Especially the work by Heeger shows that while a theory can be fitting on the computational level, the assumed algorithms to reach the algorithmic goal can vary greatly. For instance, while Heeger's theory encompasses the computations of both Bayesian inference and predictions, and thus agreeing with the BBH and PC on the computational level, it differs with algorithmic aspects of PC in how predictions are

passed through the neuronal hierarchy.

In this thesis, with measuring preattentive MMN components I concentrate on a rather low level of cognitive processes within the cortical hierarchy. Making use of the hierarchical structure in the stimulation paradigm (3.2.3), I made an attempt to detect a form of hierarchical structure within the sMMN that differs in levels of abstraction and timescale. However, this attempt was unsuccessful for the sMMN (possible reasons are discussed above; 3.4 and 4.2). Nonetheless, I think that the investigation of hierarchical information processing on such a low perceptual level is absolutely valuable. On the one hand, it can serve as an apt way to test Bayesian theories of attention (such as Dayan and Zemel, 1999; Whiteley and Sahani, 2012). On the other hand, it might lead to - as was my attempt for this thesis - a specification of the conditions relating to a sequence, under which certain structures in that sequence might be relayed to higher cognitive processing, i.e., grasp bottom-up attention, independent of perceptual modality. In that way, understanding the processing of low-level input is fundamental to research on higher cognitive processes.

However, it is entirely possible that the MMN-effect does not specifically relate to surprise or prediction error from Bayesian inference processes, but constitutes a mixture of other mechanisms and is, especially in somatosensation, impacted by adaptation. In electrophysiology, investigating processes under top-down attention that elicit a P300 potential might be more fruitful. Especially in the somatosensory domain, such studies have yielded valuable insights into Bayesian information processing (e.g., Ostwald et al., 2012, Herding et al., in prep.). In addition, data from electrocorticography with a higher spatial resolution and diminished susceptibility to artifacts, can produce more reliable results on hierarchical processes (c.f., Sedley et al., 2016).

With this in mind, questions about the possible implementation of Bayesian inference processes on any hierarchical level of the cortex have yet to receive satisfactory answers. The implementational level is purportedly the one of Marr (1982)'s three levels that my work contributes the least to. Nonetheless, many researchers have proposed possibilities in which Bayesian inference could be implemented in the brain. On a grand scale, the most influential camps are perhaps the ones of probabilistic population codes and Monte-Carlo sampling on the one hand (Ma et al., 2006; Griffiths et al., 2012; Sanborn and Chater, 2016), and predictive coding with canonical microcircuits and variational Bayes on the other (Bastos et al., 2012; Friston, 2010). Future research will show which form of implementation is more accurate, and if they can at all be combined into one

approach. However, as Pouget et al. (2013) points out in his optimistic review on probabilistic approaches to understanding the brain, the main question for future research will not be whether probability distributions are represented in the brain, but mainly to what extent the brain makes use of them.

In summary, my work contributes to existing research on sequential Bayesian updating in the brain mainly on the computational level, by investigating the mathematical concepts and their concrete realizations for surprise regressors (Chapter 2). In Chapter 3, the underlying assumption on the algorithmic level was that the cortex emits a surprise response relating to a belief distribution in the form of the preattentive MMN. As my evidence for this assumption from the somatosensory system is scarce, and replications and model comparisons from the visual and auditory domains are lacking, it is possible that the MMN is in fact unfit such questions. More concrete specifications of MMN generation as well as the applications of competing models are necessary to resolve this controversy.

4.4 Conclusion and Outlook

The thesis at hand investigates concrete applications of the BBH to sequential information processing (i) by examining a large model space of Bayesian computations as well as three different ways in which to compute surprise from the models (Chapter 2), and (ii), by applying these models to a neural surprise response, the somatosensory MMN, in order to provide evidence for the independence of neural surprise from perceptual domains (Chapter 3). While the latter attempt did not support a common principle of sequential information processing behind the MMN, the former approach yielded strong indications for the importance of testing SBL models against each other, as different model assumptions about probability distributions and surprise functions can result in vastly different assumptions about brain activity. This finding underlines the coarseness and flexibility of the BBH: While it is intriguing to assume one common principle behind all information processing in the brain, and has inspired neuroscience considerably (Friston, 2012), the real work starts when specifying operationalized hypotheses from the BBH.

Concerning the MMN, I conclude that its underlying processes are most likely not entirely independent of perceptual domain. Together with findings from the literature, my results hint at a possibility that the auditory domain is favored in preattentive sequential processing, followed by the visual, while the somatosensory system might need top-down attention in order to strengthen

the thalamo-cortical feedforward connections (Campo et al., 2018) and detect statistical properties in sequences.

Future research in the somatosensory domain could make use of the signal-enhancing effect of attention and conduct a similar stimulation paradigm with different allocations of attention. Specifically, participants could be encouraged to either count regime changes to invoke a hierarchically higher processing level, or to count stimulus repetitions in a fast regime or alternations in a slow regime, for the detection of more simple, local sequence features. By including a top-down attentional factor, P300 potentials become more measurable and could provide further insight into appropriate computational SBL schemes.

Apart from attentional boosting, another possibility to test more complex processing in the somatosensory domain could be to make use of a grounding phase of stimulus regularity in the beginning of a sequence. A longer phase of stable alternations could lead to the build-up of stronger predictions and thus to possibly more complex MMN-effects.

References

Aitchison, L. and Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227.

Alain, C., Woods, D. L., and Ogawa, K. H. (1994). Brain indices of automatic pattern processing. *NeuroReport*, 6(1):140–144.

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., and Roepstorff, A. (2016). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *NeuroImage*, 127:34–43.

Allison, T., McCarthy, G., Wood, C. C., and Jones, S. J. (1991). Potentials evoked in human and monkey cerebral cortex by stimulation of the median nerve. A review of scalp and intracranial recordings. *Brain*, 114:2465–2503.

Ashby, W. R. (1947). Dynamics of the cerebral cortex automatic development of equilibrium in self-organizing systems. *Psychometrika*, 12(2):135–140.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.

Auksztulewicz, R. and Friston, K. J. (2015). Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. *Cerebral Cortex*, pages 1–11.

Auksztulewicz, R., Spitzer, B., and Blankenburg, F. (2012). Recurrent neural processing and somatosensory awareness. *The Journal of Neuroscience*, 32(3):799–805.

Baldi, P. and Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., and Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, 113(5):E616–E625.

Barber, D. (2011). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA.

Barto, A. G., Mirolli, M., and Baldassarre, G. (2013). Novelty or Surprise? *Frontiers in Psychology*, 4(Dec):1–15.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bayes, T., Price, R., and Canton, J. (1763). An essay towards solving a problem in the doctrine of chances. *C. Davis, Printer to the Royal Society of London London, U. K.*

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–21.

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677.

Bensmaia, S. J., Denchev, P. V., Dammann, J. F., Craig, J. C., and Hsiao, S. S. (2008). The Representation of Stimulus Orientation in the Early Stages of Somatosensory Processing. *Journal of Neuroscience*, 28(3):776–786.

Berg, P. and Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and clinical neurophysiology*, 90(3):229–41.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Blankenburg, F., Ruben, J., Meyer, R., Schwiemann, J., and Villringer, A. (2003). Evidence for a rostral-to-caudal somatotopic organization in human primary somatosensory cortex with mirror-reversal in areas 3b and 1. *Cerebral Cortex*, 13(9):987–993.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components - A tutorial. *NeuroImage*, 56(2):814–825.

Bowers, J. S. and Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414.

Bowers, J. S. and Davis, C. J. (2012b). Is that what Bayesians believe? reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, 138(3):423–6.

Burton, H. and Sinclair, R. J. (2000). Attending to and Remembering Tactile Stimuli. *Journal of Clinical Neurophysiology*, 17(6):575–591.

Calvillo, D. P. and Gomes, D. M. (2011). Surprise influences hindsight-foresight differences in temporal judgments of animated automobile accidents. *Psychonomic Bulletin and Review*, 18(2):385–391.

Campo, A., Vázques, Y., Álvarez, M., Zainos, A., Deco, G., and Romo, R. (2018). Single-neuron interactions between the somatosensory thalamo-cortical circuits during perception. *bioRxiv*.

Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62.

Chennu, S., Noreika, V., Gueorguiev, D., Blenkmann, A., Kochen, S., Ibáñez, A., Owen, A. M., and Bekinschtein, T. A. (2013). Expectation and attention in hierarchical auditory prediction. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(27):11194–205.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences*, 36(3):181–204.

Cornella, M., Leung, S., Grimm, S., and Escera, C. (2012). Detection of simple and pattern regularity violations occurs at different levels of the auditory hierarchy. *PLoS ONE*, 7(8).

Cowan, N., Winkler, I., Teder, W., and Näätänen, R. (1993). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *Journal of experimental psychology: Learning, Memory, and Cognition*, 19(4):909–921.

Czigler, I. (2007). Visual Mismatch Negativity. *Journal of Psychophysiology*, 21(3):224–230.

Czigler, I. and Pató, L. (2009). Unnoticed regularity violation elicits change-related brain activity. *Biological psychology*, 80(3):339–347.

Czigler, I., Weisz, J., and Winkler, I. (2006). ERPs and deviance detection: Visual mismatch negativity to repeated visual stimuli. *Neuroscience Letters*, 401(1-2):178–182.

Daw, N. D. (2013). *Advanced Reinforcement Learning*. Elsevier Inc., 2nd edition.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz Machine. *Neural Computation*, 7(5):889–904.

Dayan, P. and Zemel, R. S. R. (1999). Statistical models and sensory attention. In *9th International Conference on Artificial Neural Networks: ICANN '99*, pages 1–6.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm By. *Journal of the Royal Statistical Society*, 39(1):1–38.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of American Statistical Association*, 76(76):341–353.

Denève, S. and Machens, C. K. (2016). Efficient codes and balanced networks. *Nature Neuroscience*, 19(3):375–382.

Desmedt, J. E., Bourguet, M., Huy, N. T., and Delacuvellerie, M. (1984). The P40 and P100 Processing Positivities That Precede P300 Closure in Serial Somatosensory Decision Tasks. *Annals of the New York Academy of Sciences*, 425(1 Brain and Inf):188–193.

Doya, K. (2007). Bayesian Brain: Probabilistic Approaches to Neural Coding.

Duncan-Johnson, C. C. and Donchin, E. (1977). On Quantifying Surprise: The Variation of Event-Related Potentials With Subjective Probability.

Dürschmid, S., Edwards, E., Reichert, C., Dewar, C., Hinrichs, H., Heinze, H.-J., Kirsch, H. E., Dalal, S. S., Deouell, L. Y., and Knight, R. T. (2016). Hierarchy of prediction errors for auditory events in human temporal and frontal cortex. *Proceedings of the National Academy of Sciences*, pages 1–6.

Eickhoff, S. B., Grefkes, C., Zilles, K., and Fink, G. R. (2007). The somatotopic organization of cytoarchitectonic areas on the human parietal operculum. *Cerebral Cortex*, 17(8):1800–1811.

Eickhoff, S. B., Weiss, P. H., Amunts, K., Fink, G. R., and Zilles, K. (2006). Identifying human parieto-insular vestibular cortex using fMRI and cytoarchitectonic mapping. *Human Brain Mapping*, 27(7):611–621.

Ekman, P. and Davidson, R. J., editors (1994). *The nature of emotion: Fundamental questions.* Series in affective science. Oxford University Press, New York, NY, US.

Faraji, M., Preuschoff, K., and Gerstner, W. (2018). Balancing New against Old Information: The Role of Puzzlement Surprise in Learning. *Neural Computation*, 30:34–83.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130.

Friston, K. J. (2002). Beyond Phrenology: What Can Neuroimaging Tell Us About Distributed Circuitry? *Annual Review of Neuroscience*, 25(1):221–250.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1456):815–36.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature reviews. Neuroscience*, 11(2):127–38.

Friston, K. J. (2012). The history of the future of the Bayesian brain. *NeuroImage*, 62(2):1230–1233.

Friston, K. J., Fortier, M., and Friedman, D. A. (2018). Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bulletin*, 2:17–43.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. D. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234.

Friston, K. J. and Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3):417–458.

Ganguli, D. and Simoncelli, E. P. (2014). Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Computation*, 26(10):2103–2134.

Gardner, E. P. and Kandel, E. R. (2000). Touch. In Kandel, E. R., Schwartz, J. H., and Jessell, T. M., editors, *Principles of Neural Science*, chapter 23, pages 451–471. McGraw-Hill, New York.

Garrido, M. I., Friston, K. J., Kiebel, S. J., Stephan, K. E., Baldeweg, T., and Kilner, J. M. (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. *NeuroImage*, 42(2):936–44.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., and Friston, K. J. (2007). Dynamic causal modelling of evoked potentials: a reproducibility study. *NeuroImage*, 36(3):571–80.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, 120(3):453–463.

George, D. and Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*, 5(10):e1000532.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Götz, T., Huonker, R., Miltner, W. H., Witte, O. W., Dettner, K., and Weiss, T. (2011). Task requirements change signal strength of the primary somatosensory M50: Oddball vs. one-back tasks. *Psychophysiology*, 48(4):569–77.

Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, 21(4):263–268.

Haenschel, C., Vernon, D. J., Dwivedi, P., Grunzelier, J. H., and Baldeweg, T. (2005). Event-Related Brain Potential Correlates of Human Auditory Sensory Memory-Trace Formation. *Journal of Neuroscience*, 25(45):10494–10501.

Harrison, L. M., Bestmann, S., Rosa, M. J., Penny, W. D., and Green, G. G. R. (2011). Time scales of representation in the human brain: weighing past information to predict future events. *Frontiers in human neuroscience*, 5(April):37.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8):1773–1782.

Heilbron, M. and Chait, M. (2017). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*.

Helmholtz, H. L. v. (1891). Versuch das psychophysische Gesetz auf die Farbenunterschiede trichromatischer Augen anzuwenden. *Z. Psychol. Physiol. Sinnesorg.*, 2:1–30.

Hendry, S. and Hsiao, S. (2008). Somatosensory System. In Squire, L. R., Berg, D., Bloom, F., and Lac, S., editors, *Fundamental Neuroscience*, pages 581 – 608. Elsevier Inc., 3rd edition.

Herding, J., Ludwig, S., Spitzer, B., and Blankenburg, F. (0). Centro-parietal EEG potentials in perceptual decision making: from subjective evidence to confidence.

Horváth, J., Czigler, I., Sussman, E., and Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Cognitive Brain Research*, 12(1):131–144.

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., and Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, 80(2):519–530.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–306.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2006). *An Introduction to Statistical learning*, volume 102. Springer US.

James, W. (1890). Attention. In *The Principles of Psychology*, pages 402–458. Dover Publications, XI.

Kaas, J. H. (1993). The functional organization of somatosensory cortex in primates. *Annals of Anatomy*, 175(6):509–518.

Kane, N. M., Curry, S. H., Bulter, S. R., and Cummins, B. H. (1993). Electrophysiological indicator of awakening. *Journal Of Clinical Pharmacology*, 341:688–689.

Kekoni, J., Hämäläinen, H., Saarinen, M., Gröhn, J., Reinikainen, K., Lehtokoski, A., and Näätänen, R. (1997). Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. *Biological Psychology*, 46(2):125–142.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

Kogo, N. and Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, 9(September):1–4.

Kolossa, A. (2016). *Computational Modeling of Neural Activities for Statistical Inference*. Springer Switzerland.

Kolossa, A., Kopp, B., and Fingscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *NeuroImage*, 106:222–237.

Kopp, B. (2007). The P300 Component of the Event- Related Brain Potential and Bayes' Theorem. *Cognitive Sciences*, 2(2):113–125.

Kording, K. P. (2014). Bayesian statistics: relevant for the brain? *Current Opinion in Neurobiology*, 25C:130–133.

Kording, K. P., Tenenbaum, J. B., and Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature neuroscience*, 10(6):779–86.

Krauel, K., Schott, P., Sojka, B., Pause, B. M., and Ferstl, R. (1999). Is There a Mismatch Negativity Analogue in the Olfactory Event-Related Potential? *Journal of Psychophysiology*, 13(1):49–55.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: from passive to active learning. *Learning & Behavior*, 36(3):210–226.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in cognitive sciences*, 10(11):494–501.

Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–9.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013a). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computational Biology*, 9(2):1–16.

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., and Friston, K. J. (2013b). A neurocomputational model of the mismatch negativity. *PLoS Computational Biology*, 9(11):1–14.

Litvak, V., Komssi, S., Scherg, M., Hoechstetter, K., Classen, J., Zaaroor, M., Pratt, H., and Kahkonen, S. (2007). Artifact correction and source analysis of early electroencephalographic responses evoked by transcranial magnetic stimulation over primary motor cortex. *NeuroImage*, 37(1):56–70.

Litvak, V., Mattout, J., Kiebel, S. J., Phillips, C., Henson, R. N., Kilner, J. M., Barnes, G. R., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W. D., and Friston, K. J. (2011). EEG and MEG data analysis in SPM8. *Computational Intelligence and Neuroscience*, 2011:1–32.

Liu, C., Lian, Z., and Han, J. (2006). How bayesians debug. In *IEEE International Conference on Data Mining*, pages 382–393.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–8.

Mach, E. (1902). *Die Analyse der Empfindungen und das Verhältniss des Physischen zum Psychischen.* Gustav Fischer, Jena.

Maguire, P., Moser, P., Maguire, R., and Keane, M. T. (2018). Seeing Patterns in Randomness: A Computational Model of Surprise. *Topics in Cognitive Science*, pages 1–16.

Marr, D. (1982). The Philosophy and the Approach. In *A Computational Investigation into the Human Representation and Processing of Visual Information*, chapter 1, pages 7–38. MIT Press, Cambridge, MA.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., and Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, 28(47):12539–45.

Mars, R. B., Shea, N. J., Kolling, N., and Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly journal of experimental psychology (2006)*, 65(2):252–67.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5(May):1–20.

Mathys, C., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, 8(November):1–24.

Meyniel, F., Maheu, M., and Dehaene, S. (2016). Human Inferences about Sequences: A Minimal Transition Probability Model. *PLoS Computational Biology*, 12(12):1–26.

Meyniel, F., Sigman, M., and Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, 88(1):78–92.

Mill, R., Coath, M., Wennekers, T., and Denham, S. L. (2011). A neurocomputational model of stimulus-specific adaptation to oddball and markov sequences. *PLoS Computational Biology*, 7(8).

Mima, T., Nagamine, T., Nakamura, K., and Shibasaki, H. (1998). Attention modulates both primary and second somatosensory cortical activities in humans: A magnetoencephalographic study. *Journal of Neurophysiology*, 80(4):2215–2221.

Mittag, M., Takegata, R., and Winkler, I. (2016). Transitional Probabilities Are Prioritized over Stimulus/Pattern Probabilities in Auditory Deviance Detection: Memory Basis for Predictive Sound Processing. *Journal of Neuroscience*, 36(37):9572–9579.

Mullens, D., Winkler, I., Damaso, K., Heathcote, A., Whitson, L. R., Provost, A., and Todd, J. (2016). Biased relevance filtering in the auditory system: A test of confidence-weighted first-impressions. *Biological Psychology*, 115:101–111.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, 66:241–251.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., and Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *PNAS*, 99(23):15164–9.

Näätänen, R. (1992). *Attention and brain function*. Psychology Press.

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, 42(4):313–329.

Näätänen, R., Kujala, T., and Winkler, I. (2011). Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. *Psychophysiology*, 48(1):4–22.

Nierhaus, T., Forschack, N., Piper, S. K., Holtze, S., Krause, T., Taskin, B., Long, X., Stelzer, J., Margulies, D. S., Steinbrink, J., and Villringer, A. (2015). Imperceptible Somatosensory Stimulation Alters Sensorimotor Background Rhythm and Connectivity. *Journal of Neuroscience*, 35(15):5917–5925.

Nordby, H., Roth, W. T., and Pfefferbaum, A. (1988). Event-Related Potentials to Breaks in Sequences of Alternating Pitches or Interstimulus Intervals. *Psychophysiology*, 25(3):262–268.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images.

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

O'Reilly, J. X., Schuffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., and Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 110(38):E3660–E3669.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., and Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *NeuroImage*, 62(1):177–88.

Phillips, H. N., Blenkmann, A., Hughes, L. E., Bekinschtein, T. A., and Rowe, J. B. (2015). Hierarchical Organization of Frontotemporal Networks for the Prediction of Stimuli across Multiple Dimensions. *Journal of Neuroscience*, 35(25):9255–9264.

Phillips, H. N., Blenkmann, A., Hughes, L. E., Kochen, S., Bekinschtein, T. A., and Rowe, J. B.

(2016). Convergent evidence for hierarchical prediction networks from human electrocorticography and magnetoencephalography. *Cortex*, 82:192–205.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–48.

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–8.

Raïffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Graduate School of Business Administration, Harvard University, Boston, MA.

Ranganath, C. and Rainer, G. (2003). Cognitive neuroscience: Neural mechanisms for detecting and remembering novel events. *Nature Reviews Neuroscience*, 4(3):193–202.

Rao, R. P. (2004). Hierarchical Bayesian Inference in Networks of Spiking Neurons. *Advances in neural information processing*, pages 1113–1120.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's Razor. *Neural Information Processing Systems*, pages 294–300.

Restuccia, D., Della Marca, G., Valeriani, M., Leggio, M. G., and Molinari, M. (2007). Cerebellar damage impairs detection of somatosensory input changes. A somatosensory mismatch-negativity study. *Brain*, 130:276–87.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, 84:971–985.

Sabri, M., Radnovich, A. J., Li, T. Q., and Kareken, D. A. (2005). Neural correlates of olfactory change detection. *NeuroImage*, 25(3):969–974.

Sanborn, A. N. and Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, 20(12):883–893.

Schmidhuber, J. (2002). Exploring the predictable. In Ghosh, A. and Tsutsui, S., editors, *Advances in Evolutionary Computing*, page 579–612. Springer Switzerland, New York.

Schröger, E. (1996). A Neural Mechanism for Involuntary Attention Shifts to Changes in Auditory Stimulation. *Journal of Cognitive Neuroscience*, 8(6):527–539.

Schröger, E. and Winkler, I. (1995). Presentation rate and magnitude of stimulus deviance effects on human pre-attentive change detection. *Neuroscience Letters*, 193:185–188.

Schubert, R., Haufe, S., Blankenburg, F., Villringer, A., and Curio, G. (2009). Now you'll feel it, now you won't: EEG rhythms predict the effectiveness of perceptual masking. *Journal of Cognitive Neuroscience*, 21(12):2407–2419.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., Howard, M. A., and Griffiths, T. D. (2016). Neural signatures of perceptual inference. *eLife*, 5:1–13.

Seer, C., Lange, F., Boos, M., Dengler, R., and Kopp, B. (2016). Prior probabilities modulate cortical surprise responses: A study of event-related potentials. *Brain and Cognition*, 106:78–89.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shinozaki, N., Yabe, H., Sutoh, T., Hiruma, T., and Kaneko, S. (1998). Somatosensory automatic responses to deviant stimuli. *Cognitive Brain Research*, 7(2):165–171.

Sohoglu, E. and Chait, M. (2016). Detecting and representing predictable structure during auditory scene analysis. *eLife*, 5(Se):1–17.

Spackman, L. A., Boyd, S. G., and Towell, A. (2007). Effects of stimulus frequency and duration on somatosensory discrimination responses. *Experimental brain research*, 177(1):21–30.

Spackman, L. A., Towell, A., and Boyd, S. G. (2010). Somatosensory discrimination: an intracranial event-related potential study of children with refractory epilepsy. *Brain research*, 1310:68–76.

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, 48(12):1391–1408.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112:92–97.

Squires, K. C., Wickens, C., Squires, N. K., and Donchin, E. (1976). The Effect of Stimulus Sequence on the Waveform of the Cortical Event-Related Potential. *Science*, 193(6):92–94.

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive Coding: A Fresh View of Inhibition in the Retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 216(1205):427–459.

Starzak, R. and Sadler, C. (2007). Season 1. In Goleszowski, R., editor, *Shaun the Sheep*. Aardman Animations, Bristol.

Stefanics, G., Heinzle, J., Attila Horváth, A., and Enno Stephan, K. (2018). Visual mismatch and predictive coding: A computational single-trial ERP study. *The Journal of Neuroscience*, pages 3365–17.

Stefanics, G., Kremlácek, J., and Czigler, I. (2014). Visual mismatch negativity: a predictive coding view. *Frontiers in Human Neuroscience*, 8(September):1–19.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–17.

Strelioff, C. C., Crutchfield, J. P., and Hübler, A. W. (2007). Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(1).

Sulykos, I., Kecskés-Kovács, K., and Czigler, I. (2013). Mismatch negativity does not show evidence of memory reactivation in the visual modality. *Journal of Psychophysiology*, 27(1):1–6.

Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, 150(3700):1187–1188.

Tales, A., Newton, P., Troscianko, T., and Butler, S. (1999). Mismatch negativity in the visual modality. *Neuroreport*, 10(16):3363–7.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, 5(11):1226–35.

Todorovic, A., van Ede, F., Maris, E., and de Lange, F. P. (2011). Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *Journal of Neuroscience*, 31(25):9118–9123.

Tribus, M. (1961). Information Theory as the Basis for Thermostatics and Thermodynamics. *Journal of Applied Mechanics*, 28(1):1.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Umbricht, D. and Krljes, S. (2005). Mismatch negativity in schizophrenia: A meta-analysis. *Schizophrenia Research*, 76(1):1–23.

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences*, 111(40):14332–14341.

Vilares, I. and Körding, K. P. (2011). Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, 1224(1):22–39.

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience*, 32(11):3665–3678.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T. A., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, 108(51):20754–9.

Wallenstein, G. V., Eichenbaum, H., and Hasselmo, M. E. (1998). The hippocampus as an associator of discontiguous events. *Trends in Neurosciences*, 21(8):317–323.

Wessel, J. R., Danielmeier, C., Morton, J. B., and Ullsperger, M. (2012). Surprise and Error: Common Neuronal Architecture for the Processing of Errors and Novelty. *Journal of Neuroscience*, 32(22):7528–7537.

Whiteley, L. and Sahani, M. (2012). Attention in a Bayesian Framework. *Frontiers in Human Neuroscience*, 6(June):1–21.

Winkler, I. and Czigler, I. (2012). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *International Journal of Psychophysiology*, 83(2):132–143.

Zénon, A., Solopchuk, O., and Pezzulo, G. (2017). An information-theoretic perspective on the costs of cognition. *bioRxiv*, (Oct):1–54.

Supplement

Freie Universität Berlin, FB Erziehungswissenschaft und Psychologie
Habelschwerdter Allee 45, 14195 Berlin

FB Erziehungswissensch
Psychologie
Neurocomputation and Neuroimaging Unit

Prof. Dr. Felix Blankenburg
Habelschwerdter Allee 45
14195 Berlin

Telefon +49 30 838-55738
E-Mail felix.blankenburg@fu-berli
Internet www.fu-berlin.de

30.04.2015

Probandeninformation zur Teilnahme an der wissenschaftlichen Untersuchung:

Welche Rolle spielt 'Predictive Coding' in der somatosensorischen Perzeption?

Sehr geehrte Probandin, sehr geehrter Proband,

die wissenschaftliche Studie wird an der Freien Universität Berlin im Fachbereich Erziehungswissenschaften und Psychologie unter der Leitung von Prof. Dr. Felix Blankenburg durchgeführt. Die an der Durchführung weiterhin beteiligten Wissenschaftler sind Dr. Dirk Ostwald und Dipl.-Psych. Kathrin Tertel.

Ziel der Studie ist es, die perzeptuelle Verarbeitung von Berührungsreizen beim Menschen besser zu verstehen. Dabei soll die Verarbeitung von komplexen Reizabfolgen studiert werden. Dazu werden Ihnen taktile Reize in Form von elektrischer Stimulation dargeboten.

Im Rahmen der Studie findet eine elektroenzephalographische (EEG) Untersuchung statt. Dazu werden Ihnen 64-Elektroden mit Hilfe einer Kappe am Kopf befestigt. Die taktilen Stimuli werden Ihnen am Handgelenk dargeboten. Vor der Untersuchung haben Sie die Möglichkeit, sich mit der Stimulation vertraut zu machen.

Freie Universität Berlin

Die gesamte Dauer der Untersuchung beträgt etwa knapp 3 Stunden. Diese setzen sich zusammen aus

- ca. 45 Minuten Vorbereitung (Informationen über den Versuch, Schwellenbestimmung der elektrischen Stimulation, Anlegen des EEG)
- ca. 110 Minuten Versuchsdurchführung (acht Stimulationsblöcke à 12 Minuten, zwischen den Blöcken können Pausen eingelegt werden)
- ca. 20 Minuten Nachbereitung (EEG entfernen, Haare waschen).

Die EEG-Aufzeichnung ist eine nichtinvasive Technik und mit keinen bekannten Nebenwirkungen für die Probanden verbunden. Durch die Anbringung der Elektroden kann es zu geringen Hautirritationen an der Kopfhaut kommen, welche als unangenehm empfunden werden. Dies wird durch die Verwendung von sog. Aktivelektroden minimiert.

Die Teilnahme an der Studie erfolgt freiwillig. Sie können jederzeit von der wissenschaftlichen Untersuchung ohne Angabe von Gründen zurücktreten. Abbruch der Studienteilnahme erfolgt auch durch den Widerruf der Einwilligung sowie durch einen Widerspruch gegen die Weiterverarbeitung der Daten. Weiterhin wird die Studie bei Unwohlsein abgebrochen. Sie stehen als Proband im unmittelbaren Kontakt mit dem Untersucher. Somit können Sie jederzeit Ihre Entscheidung einem der Studienleiter mitteilen.

Durch Ihre Unterschrift auf der Einwilligungserklärung erklären Sie sich damit einverstanden, dass der Studienleiter und seine Mitarbeiter Ihre personenbezogenen Daten zum Zweck der o.g. Studie erheben und verarbeiten dürfen. Personenbezogene Daten sind z.B. Ihr Geburtsdatum, Ihr Geschlecht oder andere persönliche Daten, die während Ihrer Teilnahme an der Studie erhoben werden. Der Studienleiter wird Ihre personenbezogenen Daten für Zwecke der Verwaltung und Durchführung der Studie sowie für Zwecke der Forschung und statistischen Auswertung verwenden. Er versieht die Studiendaten mit einer Codenummer (Pseudonymisierung der Daten). Auf den Codeschlüssel, der es erlaubt, die studienbezogenen Daten mit Ihnen in Verbindung zu bringen, haben nur der Studienleiter und seine Mitarbeiter Zugriff. Die vorhandenen Daten werden für die Zeit von 10 Jahren gespeichert und danach vernichtet. Sie haben das Recht auf Auskunft über alle beim Studienleiter vorhandenen personenbezogenen Daten über Sie. Sie haben auch Anrecht auf Korrektur eventueller Ungenauigkeiten in Ihren personenbezogenen Daten. In diesen Fällen wenden Sie sich bitte an den Studienleiter. Adresse und Telefonnummer finden Sie am Ende dieses Formblatts. Bitte beachten Sie, dass die Ergebnisse der Studie in der medizinischen Fachliteratur

veröffentlicht werden können, wobei Ihre Identität jedoch anonym bleibt. Sie können jederzeit der Weiterverarbeitung Ihrer im Rahmen der o.g. Studie erhobenen Daten widersprechen und ihre Löschung bzw. Vernichtung verlangen.

Als Aufwandsentschädigung zur Studienteilnahme erhalten Sie 10 € pro Stunde.

Fragen über alle Angelegenheiten, welche die Studie betreffen, insbesondere auch über Risiken können jederzeit an die Studienleiter gerichtet werden: **Kathrin Tertel (030 838 57967) oder Prof. Dr. Felix Blankenburg (Tel.: 030 838 55738), Habelschwerdter Allee 45, 14195 Berlin**.

Ich habe die schriftliche Information zur Durchführung der Studie gelesen, bin zusätzlich mündlich über die Studie aufgeklärt worden und habe keine weiteren unbeantworteten Fragen.

Berlin, den _____

_____

Unterschrift des Teilnehmers

Curriculum Vitae

For data protection, the curriculum vitae is not included in the online version.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

- dass ich die vorliegende Arbeit eigenständig und ohne unerlaubte Hilfe verfasst habe,
- dass Ideen und Gedanken aus Arbeiten anderer entsprechend gekenn-zeichnet wurden,
- dass ich mich nicht bereits anderweititg um einen Doktorgrad beworben habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze
- dass ich die zugrundeliegende Promotionsordnung vom 08.08.2016 anerkenne.

Berlin, den 23.08.2018

Kathrin Tertel