Universitätsklinikum Hamburg-Eppendorf

Institut für systemische Neurowissenschaften
Prof. Dr. med. Christian Büchel

Towards a unifying account of generalization in cognitive neuroscience

Dissertation
zur Erlangung des akademischen Grades Dr. rer. biol. hum.
an der medizinischen Fakultät der Universität Hamburg

vorgelegt von

Lukas Michael Neugebauer
aus Mainz am Rhein

Hamburg
2023

Angenommen von der
Medizinischen Fakultät der Universität Hamburg am 23.11.2023

Veröffentlicht mit Genehmigung der
Medizinischen Fakultät der Universität Hamburg.

Prüfungsausschuss, der/die Vorsitzende: Prof. Dr. Christian Büchel

Prüfungsausschuss, zweite/r Gutachter/in: Prof. Dr. Tobias Donner

Contents

List of Figures iii

List of Tables iv

List of Abbreviations vi

1 Introduction 1

1.1 Stimulus generalization . . . 2

1.1.1 Generalization gradients . . . 2

1.1.2 The role of perception . . . 3

1.1.3 Fear generalization . . . 5

1.1.4 Appetitive stimulus generalization . . . 8

1.1.5 Neural correlates of stimulus generalization . . . 9

1.2 Bayesian models of inductive reasoning . . . 13

1.2.1 Shepard's universal law . . . 14

1.2.2 Further Bayesian models . . . 15

1.2.3 The relationship with stimulus generalization . . . 18

1.3 Generalization in Reinforcement Learning . . . 19

1.3.1 Generalization as abstraction . . . 21

1.3.2 Dimensionality reduction . . . 21

1.3.3 Neural mechanisms . . . 23

1.4 An integrated view . . . 26

1.5 Contributions of this thesis . . . 28

2 An integrated Bayesian model of generalization 29

2.1 Limitations of previous models . . . 29

2.2 Conceptual description . . . 31

2.3 Mathematical formulation . . . 33

3 An improved method for the estimation of perceptual spaces 39

3.1 Measuring psychological spaces . . . 39

3.1.1 Multidimensional Scaling . . . 39

3.1.2 Maximum Likelihood Difference Scaling . . . 40

3.2 Hierarchical Mean Posterior Difference Scaling . . . 42

4 Empirical studies 45

4.1 General methods . . . 45

4.1.1 Stimulus Space . . . 45

4.1.2 Bayesian model fitting . . . 47

4.1.3 Model comparison . . . 48

4.2 Behavioral mechanisms of fear generalization . . . 49

4.2.1 Experimental design . . . 49

4.2.2 Model predictions . . . 54
4.2.3 Computational modeling . . . 58
4.2.4 Results . . . 63
4.2.5 Interim discussion . . . 66
4.3 Neural mechanisms of fear generalization . . . 67
4.3.1 Behavioral methods . . . 67
4.3.2 fMRI methods . . . 67
4.3.3 Behavioral results . . . 70
4.3.4 fMRI results . . . 73
4.3.5 Interim discussion . . . 83
4.4 Appetitive generalization in an online sample . . . 84
4.4.1 Methods . . . 84
4.4.2 Model predictions . . . 85
4.4.3 Results . . . 86
4.4.4 Interim discussion . . . 88

5 General discussion 91
5.1 Recapitulation of the research question . . . 91
5.2 Summary of theoretical and empirical results . . . 91
5.3 Implications for the field . . . 93
5.3.1 Generalization as a perceptual process . . . 93
5.3.2 Fear generalization as a distinct phenomenon . . . 94
5.3.3 A common mechanism of generalization . . . 94
5.3.4 Neural mechanisms of generalization . . . 95
5.4 Limitations . . . 96
5.5 Future outlook . . . 96

Bibliography 97

Abstract 112

Zusammenfassung 114

Acknowledgements 116

Curriculum Vitae 117

Eidesstattliche Erklärung 119

List of Figures

1.1 Canonical shapes of generalization gradients 3

2.1 An example for an associative map in two-dimensional space. 32
2.2 An example for dimensionality reduction. 33

3.1 Invariances of a two-dimensional MLDS model 42
3.2 Constraints to identify a two-dimensional MLDS model 43

4.1 Stimulus spaces for all three studies 46
4.2 Top level structure of the experimental design 50
4.3 An exemplary quadruplet trial. 51
4.4 Reduced stimulus space for quadruplet task. 51
4.5 Schematic depiction of a microblock. 53
4.6 Priors on λ 55
4.7 Priors on μ for aversive conditioning 56
4.8 Prior on ρ. 56
4.9 Model predictions for partial dimensionality reduction. 57
4.10 Model predictions for full dimensionality reduction. 58
4.11 Optimal perceptual space. 59
4.12 Group level perceptual spaces in behavioral study. 64
4.13 Mean ratings and posterior predictive checks for the behavioral study. 65
4.14 Posterior distributions of the model parameters for the behavioral study. . . . 66
4.15 Perceptual spaces in fMRI study. 71
4.16 Mean ratings and posterior predictive checks for the fMRI study. 72
4.17 Posterior distributions of the model parameters for the fMRI study. 74
4.18 Interpolation of behavioral models to the fMRI study. 75
4.19 GLM1 results. 76
4.20 GLM2: fMRI correlates of baseline shock expectation and emotion. 76
4.21 GLM2: fMRI correlates of the impact of proximity to the reinforced CS (CS+). 77
4.22 Depiction of model representational dissimilarity matrices (RDMs). 79
4.23 Static representations. 80
4.24 Dynamic representations. 81
4.25 Hierarchical structure of regression model. 82
4.26 Posterior distributions of regression coefficients. 83
4.27 Distribution of the number of successful oddball hits. 86
4.28 Perceptual space of the online study 87
4.29 Mean ratings and posterior predictive checks for the online study. 89
4.30 Posterior distributions of the model parameters for the online study. 90

List of Tables

4.1 Model comparison for perceptual spaces in behavioral study. 63
4.2 Model comparison for behavioral models in behavioral study. 64
4.3 Model comparison for perceptual spaces in fMRI study. 70
4.4 Model comparison for behavioral models in fMRI study. 73
4.5 Model comparison for behavioral models in online study. 88

List of Abbreviations

2-AFC	two alternative forced choice
ACC	anterior cingulate cortex
AI	artificial intelligence
AIC	Akaike information criterion
aIC	anterior insula
ANOVA	analysis of variance
BIC	Bayesian information criterion
BLA	basolateral amygdala
BOLD	blood oxygen level dependent
CA	cornu ammonis
CDF	cumulative density function
CdN	caudate nucleus
CE	central nucleus of the amygdala
CR	conditioned response
CS	conditioned stimulus
CS+	reinforced CS
CS-	non-reinforced CS
dACC	dorsal anterior cingulate cortex
DG	dentate gyrus
dlPFC	dorsolateral prefrontal cortex
DMN	default mode network
DNN	deep neural network
EC	enthorhinal cortex
EEG	electroencephalography
ELPD	expected log predictive density
EPI	echo-planar imaging
FEF	frontal eye field
fMRI	functional magnetic resonance imaging
FPN	frontoparietal attention network
FWE	family-wise error
GAD	generalized anxiety disorder
GLM	general linear model
GPI	generalized policy iteration
HDI	highest density interval
HMC	Hamiltonian Monte Carlo
hMPDS	hierarchical mean posterior difference scaling
HPC	hippocampus
IPL	inferior parietal lobule
IPS	intraparietal sulcus
ITC	inferior temporal cortex
ITI	inter-trial interval

JND	just noticeable difference
LA	lateral nucleus of the amygdala
LOO-CV	leave-one-out crossvalidation
LSS	least squares separate
LTP	long-term potentiation
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MDP	Markov decision process
MDS	multidimensional scaling
MFG	middle frontal gyrus
MLDS	maximum likelihood difference scaling
MNI	Montreal Neurological Institute
MPRAGE	magnetization-prepared rapid acquisition gradient echo
MRP	Markov reward process
MTG	middle temporal gyrus
MVPA	multivariate pattern analysis
NMDA	N-Methyl-D-Aspartat
NMDS	non-metric multidimensional scaling
NS	neutral stimulus
OFC	orbitofrontal cortex
PCA	principal component analysis
PCC	posterior cingulate cortex
PCU	precuneus
PET	positron emission tomography
PO	parietal operculum
PPC	posterior parietal cortex
PSIS-LOO	Pareto smoothed importance sampling leave-one-out crossvalidation
PTSD	post-traumatic stress disorder
rACC	rostral anterior cingulate cortex
RDM	representational dissimilarity matrix
RL	reinforcement learning
RSA	representational similarity analysis
S1	primary somatosensory cortex
S2	secondary somatosensory cortex
SAD	social anxiety disorder
SMA	supplementary motor area
SN	salience network
SVC	small volume correction
UCR	unconditioned response
UCS	unconditioned stimulus
VB	Variational Bayes
vmPFC	ventromedial prefrontal cortex

1 Introduction

> *Because any object or situation experienced by an individual is unlikely to recur in exactly the same form and context, psychology's first general law should, I suggest, be a law of generalization.*
>
> Roger Shepard, 1987, p. 1317

As pointed out eloquently by Shepard (1987), in the real world almost nothing ever happens twice in the exact same way. Whether or not this implies that a *law of generalization* should be the first law of psychology, at the very least this observation emphasizes the paramount importance of the ability to transfer knowledge between situations and stimuli. Beyond the general importance of this cognitive ability, it also happens to be one that humans excel in, which becomes especially apparent in contrast to artificial intelligence (AI). While AI agents[1] have reached superhuman levels in certain domains (e.g. the game of Go (Silver et al., 2017)) and despite recent advances in the field (Lake et al., 2015; Y. Wang et al., 2020), their ability to generalize and to learn from few examples stays far behind human abilities (Witty et al., 2021).

It is therefore not surprising that considerable effort in cognitive science, psychology and cognitive neuroscience has been put into the study of generalization and transfer learning. Yet, our understanding remains sparse and as of today a coherent theory that accounts for generalization in different contexts is missing. One likely reason lies in the fact that different subdisciplines that have historically been fairly separated have studied different applications of generalization in associative learning (Dymond et al., 2015; Ghirlanda & Enquist, 2003), reinforcement learning (RL, Lehnert et al., 2020; Niv, 2019) and inductive reasoning (Shepard, 1987; Tenenbaum & Griffiths, 2001a) in isolation. This separation has led to contradictory explanations of behavioral and neural effects and few efforts have been made to unify those (but see J. C. Lee, Lovibond, Hayes, & Navarro, 2019; Norbury et al., 2018).

In this thesis, I will argue for an integrated view and propose an underlying mechanism that is common to all these phenomena. To support this claim I will provide an opinionated review of the relevant literature, theoretical considerations and empirical results from three studies. In addition I will describe an improved method to estimate perceptual spaces, the development of which I deemed necessary during my PhD.

[1] Throughout this thesis I will use the terms agents and organism interchangeably when referring to learning entities. This is to emphasize that many of the concepts I discuss apply to artificial agents as well as animals and humans.

1.1 Stimulus generalization

Stimulus generalization (Ghirlanda & Enquist, 2003) describes a form of generalization in associative learning or – to be more precise – in classical conditioning (Pavlov, 1927). Classical conditioning is a mechanism by which organisms associate a previously neutral stimulus (NS) with an outcome (unconditioned stimulus (UCS)) that leads to a certain unconditioned response (UCR). An example for an UCS would be a painful electric shock that most likely leads to a fast motor response (the UCR) like withdrawing the affected body part. From a pattern of repeated observations in which a NS (like a picture) is followed by the UCS, organisms infer an association and start responding in a certain way that prepares them for the UCS when confronted with the former NS. When this happens, the formerly neutral stimulus has become a conditioned stimulus (CS) and the response is called a conditioned response (CR).

If, after having learned the association between the CS and the UCS, an organism shows a reaction that mirrors the CR in response to a stimulus that is similar to, but not identical with the CS, we speak of stimulus generalization (Ghirlanda & Enquist, 2003). Likely the first description of this phenomenon was given by Pavlov (1927) who noticed that his famously conditioned dog would show a CR to sounds that were similar to the bell that it had been conditioned with. In the next decades, the first researchers conducted empirical studies (Spence, 1937). Theoretical debates, e.g. about the role of perception in generalization (Lashley & Wade, 1946) initiated a more formal scientific treatment.

1.1.1 Generalization gradients

Already in this early phase, Spence (1937) introduced *generalization gradients* as a means to quantify stimulus generalization, a technique that until now is the most common way to approach stimulus generalization empirically (Webler et al., 2021). Generalization gradients are closely linked to the typical study designs, that are a defining feature of research on stimulus generalization (Dymond et al., 2015). In these designs, subjects first learn the association between a CS and a UCS in a *conditioning phase*. Typically, this is contrasted by another stimulus which is never reinforced. To distinguish those stimuli we call them the reinforced CS (CS+) and the non-reinforced CS (CS-) respectively. In the following *generalization phase*, subjects are presented with stimuli that are parametrically altered along a physical dimension (e.g. wavelength or size) and their expectation of an outcome given those stimuli is measured via explicit ratings (Dunsmoor et al., 2009), psychophysiological measures[2] (Greenberg et al., 2013a; Onat & Büchel, 2015) or a startle response (Lissek et al., 2008). In the case of non-human animals like pigeons, the strength of generalization can be measured by stereotypic behavior like *pecking* (Guttman & Kalish, 1956; Soto & Wasserman, 2010). The measured quantity can then be visualized and plotted against the perceptual continuum that the *generalization stimuli* differed on and the emerging pat-

[2] Psychophysiological measures include, but are not limited to, skin conductance and pupil dilation responses.

tern is called the generalization gradient (Figure 1.1). A typical finding is that the CR is strongest to the CS+ and decreases monotonically with increasing perceptual dissimilarity (Ghirlanda & Enquist, 2003). The shape of generalization gradients is usually described as

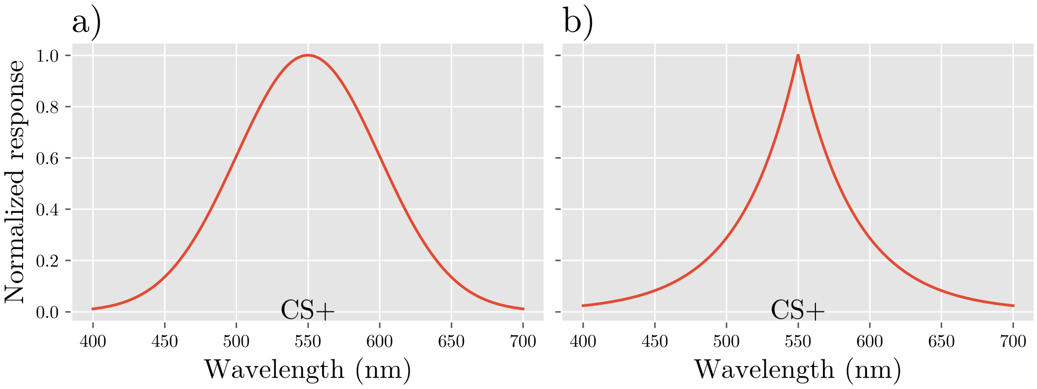


Figure 1.1: Canonical shapes of generalization gradients. Strength of responding along a perceptual continuum of wavelength after conditioning with a stimulus at $550nm$. The shape of gradients is typically assumed to be either a) Gaussian or b) exponential.

exponential (Shepard, 1987; Spence, 1937) or Gaussian (Ghirlanda & Enquist, 2003; Onat & Büchel, 2015). This discrepancy can mostly be attributed to aspects of study designs like the number of learning examples (Shepard, 1987) or the perceptual distance between stimuli (Ghirlanda & Enquist, 2003).

1.1.2 The role of perception

> *Indeed, an animal would be ill served by the assumption that just because it can detect a difference between the present and a previous situation, what it learned about that previous situation has no bearing on the present one.*
>
> Roger Shepard, 1987, p.1319

The obvious correlation between perceptual dissimilarity, which implies discriminability, and the strength of generalization has led researchers to propose that generalization is merely a perceptual by-product, i.e. a failure to discriminate novel stimuli from the CS+ (Guttman & Kalish, 1956; Struyf et al., 2015). As an argument for this mechanism, Guttman and Kalish (1956) showed that pigeons show a steeper generalization gradient in regions of a perceptual continuum in which they are known to have better discrimination abilities. Despite descriptions of phenomena that are not in line with a purely perceptual mechanisms, like *peak shifts* (Baddeley et al., 2007), i.e. stronger responding to stimuli other than the CS+ or intensity generalization (Ghirlanda & Enquist, 2003) and regardless of vehement criticism of this proposal (Shepard, 1987), this view is still being defended today by some researchers. Struyf et al. (2015) suggested that a Gaussian generalization gradient could emerge from averaging responses over trials in which a novel stimulus is either correctly classified as novel or confused for the CS+. Assuming Gaussian perceptual noise and that subjects respond only to a stimulus they assume to be the CS+, this mechanism would indeed lead to Gaussian gradients. To support this claim, the group reported that generalization gradients are flat when only considering the subset of trials in which stimuli were categorized as the CS+ (Struyf et al., 2017) and that misidentifying a stimulus as the CS+ leads to higher generalization (Zaman et al., 2019; Zaman et al., 2021). Contradicting the idea of a perceptual mechanism, Tuominen et al. (2019) carefully accounted for perceptual accuracy and found behavioral generalization gradients that were wider that what would be expected based on perceptual confusion alone[3]. While they do report that gradients in skin conductance do not extend beyond perceptual thresholds, this result is based on a non-significant *p-value* and reflects the very common statistical misunderstanding that a non-significant result provides evidence for the *null hypothesis* (Wasserstein & Lazar, 2016).

Since generalization is a very adaptive process, one could argue that this mechanism lacks face validity. If for instance an early human narrowly escaped a tiger's attack and thus formed an association between the stimulus *tiger* and the outcome *nearly dying*, they would be well advised to avoid lions as well, although it seems unlikely to confuse them with tigers. In general, it does not appear plausible that such an important mechanism is a by-product of perceptual imperfection. The ability to generalize is so fundamental to survival that Shepard (1987) proposed it as the *first law* of psychology. It seems that the proposal of a purely perceptual process is largely based on two aspects: Specifics of the employed study designs and a fundamental confusion *between the map and the territory* (Korzybski, 1933).

With respect to the first point, many studies (e.g. Tuominen et al., 2019; Zaman et al., 2019) employ very long generalization phases, i.e. many repetitions of the generalization stimuli. Implicitly, the learning process is treated as if it were constrained to the conditioning phase. However, this is not the case. In the generalization phase subjects are presented with an abundance of trials in which they can learn that generalization stimuli are not followed by an UCS and if one averages response over the whole phase, this average at least partly consists of trials in which there is no uncertainty left about the predictive value of stimuli. In addition, studies that claim to provide evidence for a perceptual mechanism often ask subjects to identify whether a stimulus is the CS+ or not before providing an outcome expectation rating (Struyf et al., 2017; Zaman et al., 2019). This design feature could lead to a situation in which subjects feel obliged to respond as they would for the CS+ instead of integrating their response over the perceptual uncertainty. Essentially, identifying a stimulus and passive viewing are two different tasks and can lead to different outcomes.

To expand on the second point, it feels instructive to remember the purpose of empirical research, namely the discovery of mechanisms and the use of empirical data to constrain the theory development process. The idea of confusing the map with the territory goes back to the philosopher Korzybski (1933) and describes a logical fallacy in which one treats a model of a *thing* as the thing itself. When I apply this fallacy to the present context, the *terri-*

[3] Note that this study assumed constant discriminability, an assumption that is challenged by other studies. This research is summarized in the next paragraph.

tory is the actual generalization process, while the *map* are theories and empirical results. Confusing the map with the territory thus refers to the attempt to come up with a mechanism that is able to explain empirical data without sufficient consideration of the validity of this mechanism with respect to the actual underlying process. While it is possible that a purely perceptual mechanism is able to explain most or even all results in a very narrowly defined field, even just outside this field it is insufficient to explain the data. For instance, in intensity generalization[4], the typical finding is a monotonically increasing generalization gradient. This implies stronger responding to some generalization stimuli than to the CS+, even if those are clearly distinguishable (Ghirlanda, 2002). Having Occam's razor in mind, it seems unlikely that there are two completely independent mechanisms behind these very related phenomena and thus a purely perceptual account is not sufficient as an overarching theory of generalization.

Aversive conditioning and perceptual tunings. While perception is likely not the whole story, discriminability is an important consideration in the context of generalization. First of all, it is a definite lower bound. If a stimulus or situation is objectively different from a known stimulus or situation, but one fails to discriminate it from the known one, one would respond as if they were the same[5]. In addition to that, some studies have revealed interesting perceptual consequences of aversive conditioning.

Based on the assumption that due to a *better safe than sorry* strategy, generalization around stimuli with a negative association should be wider, a retuning of perceptual thresholds has been proposed as one implementing mechanism of this strategy (Laufer et al., 2016; Schechtman et al., 2010). In line with this suggestion, it has been shown that the just noticeable difference (JND), a measure of perceptual tuning, increased around negatively reinforced stimuli (Resnik et al., 2011), but not around neutral or positively reinforced stimuli (Laufer & Paz, 2012). While all these studies conceptualize generalization as a failure to discriminate, which clearly emphasizes the role of perception, they do compellingly suggest a role of perceptual retuning in generalization. These findings imply a necessity to account for perceptual accuracy and changes thereof in studies on and theories of generalization.

1.1.3 Fear generalization

In the current millennium, most studies on stimulus generalization have been concerned with a special case, namely *fear generalization* (Dymond et al., 2015). Fear generalization[6] describes the tendency to react to unknown stimuli in a similar way as to stimuli that have been associated with an aversive outcome like a painful electric shock (Dymond et al., 2015). Research on fear generalization has produced a large body of behavioral (Dymond et al., 2015) and neural (Webler et al., 2021) work, in parts due to its relevance as a potential

[4] Intensity generalization describes stimulus generalization along a continuum that changes quantitatively, e.g. notes of the same frequency but with changing amplitudes.

[5] If this was all there is to generalization, it would be the inverse equivalent of discrimination and therefore redundant.

[6] Or maybe more accurately *threat generalization* (LeDoux, 2014).

clinical marker for anxiety disorders (Berg et al., 2020; Greenberg et al., 2013b; Lissek, 2012) like the generalized anxiety disorder (GAD) and post-traumatic stress disorder (PTSD).

In general, research on fear generalization follows the study designs of earlier work on stimulus generalization (Ghirlanda & Enquist, 2003). Typically, subjects are presented with all stimuli (i.e. CS+, potentially CS- and generalization stimuli) before a conditioning procedure in order to probe baseline responding. In a following conditioning phase, one stimulus (the CS+) is probabilistically paired with an aversive outcome (e.g. an electric shock, Onat & Büchel, 2015) while another stimulus (the CS-) is never paired with the outcome and thus acts as a safety signal. Lastly, in the generalization phase, all stimuli are presented again and some measure of outcome expectation or arousal is recorded[7]. One interesting aspect of this line of research is that most studies are conducted as neuroimaging studies, using functional magnetic resonance imaging (fMRI), which allows for an investigation of neural aspects and has generated a vast amount of data and converging evidence on brain areas that are involved (Webler et al., 2021).

Fear generalization studies overwhelmingly report behavioral generalization gradients as they would be expected from the stimulus generalization literature (Dymond et al., 2015; Kausche, Zerbes, Kampermann, Müller, et al., 2021; Lissek et al., 2008; Onat & Büchel, 2015). An interesting exception is one of the first behavioral studies on fear generalization (Dunsmoor et al., 2009) and a replication of that study that also includes fMRI data (Dunsmoor et al., 2011). These studies used stimuli that depicted a human face and differed along a perceptual continuum of fearfulness from neutral to maximally fearful expressions. In contrast to the typically symmetric gradients, the authors found monotonically increasing gradients (Ghirlanda & Enquist, 2003) that are more typical for generalization along intensity dimensions (like loudness, Ghirlanda, 2002). This was even true when the most fearful face served as the CS-, which is in contrast to the literature on intensity generalization. The authors interpret this as a mixture of associative learning processes and selective sensitization for stimuli that are *a priori* linked to fear (Dunsmoor et al., 2009). A more formal explanation of these findings might be given by a Bayesian approach that emphasizes the importance of prior information (Tenenbaum & Griffiths, 2001a). This approach would give a mathematical account for how the conditioning experience is integrated into prior assumptions to generate the observed gradients.

Factors influencing fear generalization. Most of the research on fear generalization is not primarily concerned with a behavioral generative model of generalization, i.e. an explanation of *why* generalization gradients look the way they do, but rather with the comparison of healthy controls and anxiety patients and the quest for a clinical marker for anxiety disorders. However, some studies have explored factors that have an influence on fear generalization. One important factor that seems to influence the width of generalization is intensity of the UCS: Dunsmoor et al. (2017) report that subjects that received a stronger UCS generalized more widely. This is in line with the idea of a *better safe than sorry* strategy

[7] See subsection 1.1.1.

(Laufer et al., 2016; Schechtman et al., 2010). Another important factor is prior knowledge, independently of how it is administered (Ahmed & Lovibond, 2015a, 2015b; Vervliet et al., 2010). Vervliet et al. (2010) used stimuli that differed on two dimensions: Shape and color. They instructed subjects that only one of these dimensions were predictive of the outcome and found stronger generalization to stimuli that shared the relevant feature with the CS+ than to stimuli that shared the irrelevant feature. Ahmed and Lovibond (2015b) expanded on this finding by replacing instructions with a phase in which subjects learned that either color or shape predict an outcome. Their findings mirror those of Vervliet et al. (2010) in that subjects generalized to stimuli that shared the relevant feature. Finally, Ahmed and Lovibond (2015a) could replicate the findings of Vervliet et al. (2010) even when instructions of feature relevance were given after the conditioning phase. These findings on the role of prior knowledge are very interesting because they fit in nicely with two aspects that have been emphasized in other research on generalization. First of all, the relevance of prior knowledge is in line with the Bayesian interpretation of the results in Dunsmoor et al. (2009) and Dunsmoor et al. (2011) and with Bayesian models of inductive reasoning that I will describe later (see section 1.2, Shepard, 1987; Tenenbaum & Griffiths, 2001a). In addition, the use of multidimensional stimuli and the induced different relevance of stimulus dimensions provides a link to the concept of generalization via dimensionality reduction in reinforcement learning (RL, see section 1.3, Niv, 2019).

Conceptual fear generalization. Another line of research that raises substantial questions about the idea of a purely perceptual mechanism of generalization especially in humans is generalization on a conceptual level (Dunsmoor & Murphy, 2015). Dunsmoor et al. (2012) used pictures of tools and animals as the CS+ and CS- category (counterbalanced between subjects) and found generalization to other members of the same category despite perceptual differences between e.g. a hammer and a saw. Vervoort et al. (2014) expanded on this finding and first induced arbitrary categories of visual stimuli in a learning task. Using these arbitrary categories they found that associative learning as well as extinction generalized to other members of the category. Boyle et al. (2016) showed that when words were associated with an aversive outcome, these associations generalized to synonyms of these words. Lastly, Morey et al. (2020) replicated the results of generalization within categories of animals and tools (Dunsmoor et al., 2012) while simultaneously recording fMRI data[8].

Avoidance generalization. While most of the research on fear generalization in humans measures purely associative learning via explicit expectations or psychophysiological signals, some studies have been concerned with the translation into behavior by measuring avoidance behavior (Lommen et al., 2010; Norbury et al., 2018; van Meurs et al., 2014). Importantly, this approach differs from RL, because the learning phase is still purely associative[9]. Still,

[8] summarized in subsection 1.1.5.

[9] While associative learning is passive and only concerned with the predictive value of stimuli, RL is active and describes how an agent learns the value of different actions in different states *from* taking these actions and observing the outcomes. For a more detailed description of RL, see section 1.3.

the study of avoidance generalization adds ecological validity because it questions how these associations translate into a binary choice between approach and avoidance.

The first study on avoidance generalization that I am aware of was actually primarily concerned with the effect of neuroticism (Lommen et al., 2010). This is not the scope of this thesis, but we can still learn from the results that avoidance gradients follow the structure of gradients in purely associative generalization: Subjects tended to avoid stimuli that were more similar to the CS+ more strongly just as they rated stimuli that were more similar to be more likely to be associated with an outcome.

A few years later, van Meurs et al. (2014) used a *farmer's game* in which subjects were instructed to maximize their *harvest*. In an associative conditioning phase, subjects observed a farmer going the *short* route to the harvesting field in the context of different visual stimuli. One of these stimuli predicted a shock for both the farmer and the subjects themselves. Afterwards, they went through an associative and then instrumental generalization phase. In the associative generalization phase, they rated the shock expectation for different generalization stimuli without any behavioral choices. In contrast to that, they had the option to take either the known, short route or a previously unobserved long route when prompted with a stimulus in the instrumental generalization phase. The short route led to a safe harvest but came at the risk of receiving a shock. Differing from this, the long route never led to a shock, but came at the cost of potentially losing the harvest. The comparison of associative and instrumental generalization gradients revealed the same pattern as in the study by Lommen et al. (2010): Subjects showed the highest shock expectation and avoidance for the CS+ and a gradual decline with perceptual dissimilarity. In addition, these two measures were strongly correlated between subjects.

Norbury et al. (2018) conducted another study in order to bridge the gap to instrumental learning while controlling for perceptual accuracy. In line with previous studies they found a correlation between associative and instrumental generalization (van Meurs et al., 2014), wider generalization around negatively than neutrally reinforced stimuli (Laufer & Paz, 2012; Resnik et al., 2011) and independent contributions of perceptual and value-based processes (Zaman et al., 2021).

Taken together, these findings emphasize the behavioral relevance of associative learning and provide a bridge between purely associative generalization and generalization in RL. This suggests that generalization is not purely a passive process, but it translates into decisions and thereby allows us to adaptively interact with our environment.

1.1.4 Appetitive stimulus generalization

While the focus of generalization research in recent years has been on fear generalization, a few studies have investigated generalization in the context of appetitive conditioning. Andreatta and Pauli (2019) reported an experiment that closely followed the structure of those in fear generalization research. They adapted the paradigm of Lissek et al. (2008) to appetitive generalization by using an appetitive UCS instead of an electric shock. The UCS consisted of either a salty or a sweet snack while the decision for either was made by subjects individually. The authors found generalization gradients that closely followed the results of fear generalization studies (Lissek et al., 2008). Due to the conceptual similarity, this study is well suited to compare the results of fear and appetitive generalization.

Other studies that can be considered as being concerned with appetitive generalization are not as easily comparable to fear generalization since they use a less direct operationalization of an appetitive outcome. FeldmanHall et al. (2018) suggested that the mechanism that is employed in stimulus generalization is also relevant in social interactions, specifically when deciding whom to trust. To this purpose, they conducted a study in which subjects played a game against opponents that were indicated by a picture of a face and displayed different levels of trustworthiness. In the generalization phase, subjects had to chose with whom to play out of a set of new potential players. Unbeknownst to the subjects, those *new* faces were morphs of the players that they had already played against. In line with other research on stimulus generalization in general (Ghirlanda & Enquist, 2003) and fear generalization more specifically (Dymond et al., 2015), FeldmanHall et al. (2018) found that subjects' tendency to play with new players increased with perceptual similarity to the trustworthy players in a typical generalization gradient. Kampermann et al. (2021) investigated the generalization of a conditioned placebo expectation. To this purpose they used tonic heat pain stimuli and conditioned subjects to expect a pain reduction in the context of a specific *physician*, a computer-generated face, by lowering the temperature while the face was displayed. In contrast, another physician (i.e. another face) predicted less pain relief. When probing the *placebo response* – a lowering in pain ratings – to faces that differed from both previous faces, they found the Gaussian pattern that is typical for generalization gradients. Subjects showed the strongest placebo response to the *effective physician* and this effect decreased with perceptual dissimilarity.

These findings and their overlap with those from fear generalization research question the (at least) implicit assumption that fear generalization is fundamentally different from generalization in other contexts.

1.1.5 Neural correlates of stimulus generalization

Neuroimaging research on stimulus generalization typically uses fMRI to measure brain activity non-invasively. Since this is only possible since the discovery of the blood oxygen level dependent (BOLD) signal by Ogawa et al. (1990), this line of research follows the general trend of research on generalization in cognitive neuroscience and is primarily concerned with fear generalization. While this means that there is very limited data on appetitive generalization, an astonishingly large fraction of fear generalization studies have used fMRI, resulting in a large body of literature. Despite some typical inconsistencies in the naming of brain areas that stem from a mix between anatomical (e.g. middle frontal gyrus (MFG)) and loosely defined functional labels (e.g. dorsolateral prefrontal cortex (dlPFC)), those results show lots of convergence with respect to brain areas that are involved (Webler et al., 2021).

Neural generalization gradients. The approach of most of empirical neuroimaging work on fear generalization has been a translation of behavioral gradients to the neural domain. The focus has been on identifying areas that show generalization gradients which follow behavior. In this context, a positively tuned area shows the strongest response to the CS+ and a decline of this activity for stimuli with increasing perceptual dissimilarity to the CS+. Negatively tuned areas show the opposite pattern with strongest deactivations for the CS+. Notably, this approach is purely correlational and as such cannot distinguish between the process of generalization itself and the results of generalization (such as enabling behavioral output). Thus, it does not provide mechanistic insight into the neural implementation of generalization.

Dunsmoor et al. (2011) conducted the first study that used this approach. They found positive gradients in the anterior insula (aIC), thalamus and caudate nucleus (CdN) and negative gradients in the rostral anterior cingulate cortex (rACC)[10]. Due to the focus on fear and the assumed crucial (LeDoux, 2003), but increasingly questioned (Fullana et al., 2016; Radua & Fullana, 2022) role of the amygdala in the context of threat processing, they also investigated amygdala connectivity per stimulus and found increased connectivity with visual areas for the CS+ and the closest generalization stimulus.

Shortly after that, two studies largely replicated these findings (Greenberg et al., 2013a, 2013b). In addition, Greenberg et al. (2013a) reported positive tunings in the dorsal anterior cingulate cortex (dACC) and negative gradients in the primary somatosensory cortex (S1) and Greenberg et al. (2013b) found positive gradients in the supplementary motor area (SMA).

Lissek et al. (2014) extended the list with positive tunings in the MFG and the inferior parietal lobule (IPL) and negative tunings in the hippocampus (HPC) and the precuneus (PCU). Taken together those areas comprise the list of areas that were typically found to show positive and negative tunings in following studies on fear generalization (e.g. Berg et al., 2020; Kaczkurkin et al., 2016; Kausche, Zerbes, Kampermann, Büchel, & Schwabe, 2021; Kausche, Zerbes, Kampermann, Müller, et al., 2021; Lange et al., 2017; Onat & Büchel, 2015; Tuominen et al., 2019; Webler et al., 2021).

Interestingly, similar results have been observed in studies that do not follow the typical fear generalization design, namely in conceptual (Morey et al., 2020) and cue-context fear generalization (de Voogd et al., 2020) as well as appetitive generalization (FeldmanHall et al., 2018). These results suggest a more general neural model of generalization that is not limited to fear.

The network view. In recent years, an interesting pattern in reported brain activations has been observed: Virtually all of the brain areas that show tuned generalization gradients are part of one of the three major brain networks[11] (Menon, 2011; Raichle, 2015; Yeo et al., 2011):

[10] Note that this activation likely maps onto the ventromedial prefrontal cortex (vmPFC) in other studies.

[11] Note that the naming of those networks is very inconsistent (Uddin et al., 2019). I am going to use the names that I deem most established in the relevant literature (e.g. Berg et al., 2020; Niv, 2019).

1. The default mode network (DMN) with important nodes in the vmPFC, HPC, posterior cingulate cortex (PCC), PCU and middle temporal gyrus (MTG).
2. The salience network (SN) with hubs in the aIC and dACC.
3. The frontoparietal attention network (FPN), primarily comprised of the MFG and the intraparietal sulcus (IPS) although other areas are sometimes considered part of the network as well (e.g. CdN, Uddin et al., 2019).

Tuominen et al. (2019) were the first to explicitly mention this network structure of neural results in fear generalization and linked deactivations in both the HPC and the PCC to the DMN. Berg et al. (2020) extended on this and understood positive tunings in the MFG and IPS as part of the FPN[12] and the SN. Importantly, areas in the DMN consistently show a negative tuning while the FPN and the SN are tuned positively, a pattern that is considered established by now (Webler et al., 2021).

The correlational structure of the bulk of neuroimaging results in conjunction with our understanding of how these networks interact (Goulden et al., 2014) allow for a possible interpretation of almost all neuroimaging results in fear generalization: The established idea of the interplay between those networks posits that the DMN is involved in self-referential processes and famously more active in the absence of a task that requires attention (Raichle, 2015). In direct contrast, the FPN is linked to attention and planning and more active during cognitively demanding tasks (Ptak et al., 2017). Lastly, the SN is thought to be involved in the detection and processing of salient stimuli and to modulate the activity of other relevant networks in response. In particular, activity in the SN is thought to initiate the switch from a brain-wide default state to a task state by inhibiting the DMN and activating the FPN (Goulden et al., 2014). One possible interpretation of the observed pattern of neural generalization gradients is therefore purely based on the results of learning. Stimuli that are more similar to the CS+ are more salient. This would lead to stronger activation of the SN which then upregulates activity in the FPN and downregulates the DMN, proportional to the salience of stimuli, in order to enable a *fight-or-flight* response. This interpretation is quite pessimistic as it implies that most neuroimaging results tell us nothing about the process of generalization itself and merely reflect arousal as a result of the learning and inference process. However, as explained later, I do not think that we need to be *that* pessimistic.

Mechanistic insights. A few studies went beyond the purely correlational approach and tried to investigate the actual neural mechanism of generalization, typically using multivariate approaches like representational similarity analysis (RSA, Kriegeskorte et al., 2008) or model-based fMRI (Gläscher & O'Doherty, 2010).

Onat and Büchel (2015) reported a tuning in the inferior temporal cortex (ITC) that didn't follow the usual Gaussian or exponential shape, but was characterized as a cosine curve with stronger responding to stimuli that were shown during the conditioning phase (i.e. CS+ and CS-). Since these are the stimuli that subjects knew about the most after

[12] Although they call it the *central executive network*.

conditioning, the authors interpreted this tuning as reflecting uncertainty. In addition, they reported a *hyper-sharp* tuning in the aIC, i.e. a steeper gradient than in behavior. Using multivariate pattern analysis (MVPA) they also found the representations of stimuli in the aIC to be increasingly similar to the representation of the UCS, depending on the perceptual dissimilarity to the CS+. These results contradict the idea of a purely perceptual process and question the arousal-based network view (see above).

Norbury et al. (2018) used RL models to distinguish between perceptual processes and value learning and found activity in the aIC and the striatum that reflected value-based processing above and beyond what could be explained by perceptual generalization.

Lastly, de Voogd et al. (2020) investigated fear generalization within and between contexts. Using a decoding algorithm (Diedrichsen & Kriegeskorte, 2017), they found the strength of context representation in the HPC to be linked to the strength of generalization. Importantly, this finding emphasizes the role of the HPC with respect to cognitive maps (Bottini & Doeller, 2020) and opens up a possible route towards a more general model of generalization.

A hippocampal model of fear generalization. The most prominent model of fear generalization was proposed by Lissek (2012), reiterated by Lissek et al. (2014) and recently extended by Webler et al. (2021). Due to two major assumptions, namely that fear generalization is different from other forms of generalization and that it is fundamentally based on perception, this model is based on fear learning in the amygdala (LeDoux, 2003) and pattern completion and separation in the HPC (McHugh et al., 2007; Rolls, 2013; Yassa & Stark, 2011). The rough outline is as follows:

An aversive UCS leads to activation of the lateral nucleus of the amygdala (LA) via the thalamus. The LA then activates the central nucleus of the amygdala (CE), which projects to areas of a fear network that are instrumental in preparing an appropriate response (e.g. aIC). A CS before conditioning does not directly activate the LA, but leads to a release of glutamate. This is happening via the *quick and dirty* or *low road* that projects directly from the thalamus to the LA without involving sensory areas like the visual cortex in the case of visual stimuli. If this CS is followed by the UCS, the activation of the LA in the context of released glutamate leads to long-term potentiation (LTP) that is mediated via N-Methyl-D-Aspartat (NMDA) receptors. Following this neural learning process, the CS can now activate the LA (and downstream the CE and aIC) in the absence of an UCS. If a stimulus that is similar to the CS is presented after conditioning, the thalamus will forward this information to both visual areas via the *high road* and to the amygdala via the *low road.* The route via the amygdala will already activate fear areas given enough overlap with the CS. The visual cortex will forward the information to the HPC, which performs pattern separation or completion, depending on the perceptual overlap to the CS (i.e. the representational similarity). In the case of pattern completion, it will activate fear areas including the amygdala and the aIC. In the case of pattern separation, it will activate the vmPFC, which in turns inhibits fear areas that were already activated by the amygdala. The latest iteration of the model (Webler et al., 2021) adds another effect of the *low route*, namely a pre-activation of the cornu ammonis (CA) subfields CA1 and CA3, which is supposed to bias the HPC towards pattern completion. In addition, they partly acknowledge the network view by stating that fear areas including those of the SN activate the FPN to recruit attentional resources.

At a first glance, this model is intriguing because it combines the animal literature on fear learning with the human literature on fear generalization and one aspect of the role of the HPC. Another strength is that it makes specific predictions for the direction of generalization gradients in different subfields of the HPC. This is because pattern completion and separation are thought to be performed in different areas, with the CA3 being involved in pattern completion while the dentate gyrus (DG) seems to perform pattern separation (Rolls, 2013). However, the model has multiple empirical and conceptual shortcomings. Starting empirically, Huggins et al. (2021) performed a segmentation of hippocampal subfields and the nuclei of the amygdala and investigated their fear tunings separately. They found a negative tuning in CA1, CA3 and DG, which directly contradicts the predictions of the model. In addition, they did not observe any tuning in the LA or CE and found a negative tuning in the basolateral amygdala (BLA). In general the role of the amygdala in fear generalization is unclear due to inconsistent results (Dunsmoor et al., 2011; Kaczkurkin et al., 2016; Onat & Büchel, 2015) and even the idea of the amygdala as the *fear center* is increasingly controversial (Fullana et al., 2016; Radua & Fullana, 2022). In addition, since the model is based on perceptual similarity, it cannot explain differential generalization along different dimensions, if discriminability is matched (Ahmed & Lovibond, 2015a, 2015b; Vervliet et al., 2010).

Conceptually, the model has two implicit assumptions about fear generalization, namely that it is fundamentally a perceptual process and that fear generalization employs a different neural architecture than other forms of generalization and is thus distinct. Those assumptions can be derived from the focus on pattern separation and completion in the hippocampus and on the amygdala respectively. As I have argued in previous sections, both of these assumptions are questionable.

1.2 Bayesian models of inductive reasoning

> *The „universality, invariance, and elegance“ of Shepard's exponential law [...] are in themselves impressive, but perhaps ultimately of less significance than the spirit of rational analysis that he has pioneered as a general avenue for the discovery of perceptual-cognitive universals.*
>
> Tenenbaum and Griffiths (2001a), p. 639

Another line of research in cognitive science that started with the seminal paper by Shepard (1987) characterizes generalization as inductive reasoning. Although the examples in that paper can be understood as a stimulus generalization, the approach is completely different because it is firmly rooted in *rational analysis*[13] (J. R. Anderson, 1990). In contrast to mechanistic explanations that describe the structure of phenomena, rational analysis is concerned with their purpose. It considers the environment that an agent interacts with, the goals that it is trying to achieve and some plausible constraints about what the agent can do. Following these considerations, rational analysis tries to come up with an optimal solution given the constraints (J. R. Anderson, 1990). This discrepancy explains why the explanations of stimulus generalization and inductive reasoning diverge drastically — it does not seem very reasonable to assume that generalization is based on perceptual confusion when the goal is to adaptively behave in an ever changing world. As a consequence, although almost all of empirical papers on stimulus generalization cite Shepard's paper, the two approaches have developed completely independently until very recent attempts to unify them (J. C. Lee, Lovibond, Hayes, & Navarro, 2019).

1.2.1 Shepard's universal law

In an attempt to form psychology into a first-class quantitative science, Shepard (1987) had the ambitious idea to derive a universal law that transcends the boundaries of species and would be the psychological equivalent to the law of gravity. Because a changing context is an underlying problem for all cognitive abilities, he suggested that generalization would be a good first cognitive universal. This priority is in stark contrast to the mechanistic explanations in the previous sections, in which an organism first learns and then generalizes (making it a second thought) and to previous approaches in stimulus generalization in which generalization depends on similarity (Guttman & Kalish, 1956) which emphasizes similarity as the primary concept. Shepard (1987) and others (Tenenbaum & Griffiths, 2001a, 2001b) have argued that similarity is an under-defined concept while generalization as an objectively measurable quantity is not and turned the relationship around by estimating similarity from generalization instead of the other way around. This approach necessarily rejects the role of objective physical similarity in generalization and replaces it with similarity in a *psychological space* that can be different for e.g. different species or individuals. This is a necessary precondition for an *universal law* and accounts for phenomena in which physical and perceived similarity do not match, such as tone frequencies[14]. Using non-metric multidimensional scaling (NMDS)[15], Shepard (1987) found that generalization gradients across stimuli and species consistently follow an exponential decay with increasing dissimilarity in the appropriate psychological space[16]. Motivated by this finding, Shepard developed his exponential law.

The basic idea of Shepard's law is that stimuli can be understood as points in psychological space. In this context a psychological space is a conceptual metric space in which stimuli

[13] Although Shepard's paper predates the definition of rational analysis, it is usually considered an example of this approach.

[14] As an example, two tones that are apart by an octave are perceived as more similar to each other than two tones that are apart by a fifth even though an octave is a larger distance in frequency space.

[15] see subsection 3.1.1.

[16] Note that the circularity of first estimating the psychological space from generalization data and then describing generalization in this space is broken by imposing the constraint of a metric psychological space.

are embedded. The distance between two stimuli in psychological space is the distance between their corresponding points and is inversely proportional to the perceived similarity between the stimuli. The task of the agent is to infer *natural kinds* of stimuli that lead to the same outcome. In the context of psychological space, natural kinds map onto *consequential regions*, i.e. areas in psychological space (Shepard, 1987). More intuitively, a natural kind is a subset of stimuli that lead to the same outcome and are clustered in a certain region of psychological space. This region is called a consequential region. Although later extensions go beyond this assumption (Tenenbaum & Griffiths, 2001a), Shepard's original law is only concerned with a special case of learning: Given that an organism observes that a single stimulus x leads to a certain outcome (i.e. is part of a certain natural kind), how likely is it that another stimulus y is part of the same natural kind. Mathematically speaking, an organism is trying to compute the posterior probability $g(y)$ that the relevant consequential region includes the novel stimulus given that it overlaps the first stimulus. Assuming a fixed size s, a uniform prior over the location of the consequential region and some constraints on their shape, this probability can be computed as the volume of the overlap $\boldsymbol{m}(x, y)$ between two regions, that are centered around x and y respectively, relative to the volume of a whole consequential region $\boldsymbol{m}(x)$:

$$g(y|s) = \frac{\boldsymbol{m}(x, y)}{\boldsymbol{m}(x)} \tag{1.1}$$

Because the size of the region is not known in practice, this uncertainty is accounted for by integrating Equation 1.1 over all possible sizes:

$$g(y) = \int_{s=0}^{\infty} p(s) \frac{\boldsymbol{m}(x, y)}{\boldsymbol{m}(x)} ds \tag{1.2}$$

Note that this equation is depending on a choice for the prior $p(s)$. However, using simulations, Shepard (1987) showed that the shape of the resulting gradients is remarkably indifferent to the prior $p(s)$ and that an approximately exponential shape emerged for a wide range of choices.

Even though this model is only applicable to cases in which an organism observes a single consequential observation, it proved to be very influential and motivated extensions that account for multiple observations (Tenenbaum & Griffiths, 2001a), sampling assumptions (Navarro et al., 2012), multiple latent causes (Soto et al., 2014) and negative evidence (J. C. Lee, Lovibond, Hayes, & Navarro, 2019; Voorspoels et al., 2015). I will discuss these extensions in the following section.

1.2.2 Further Bayesian models

As a reminder, Bayes' theorem is a result from probability theory that provides a mathematical way to integrate prior knowledge with new information (Blitzstein & Hwang, 2015). The basic formula is given by

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}, \tag{1.3}$$

which simplifies a bit when solving the integral in the denominator:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{1.4}$$

And because $p(y)$ is a constant, namely the normalizing constant, it can be dropped to yield the unnormalized posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{1.5}$$

The purpose of Bayes' theorem is to update the belief state over θ given new evidence y. In this context, $p(\theta)$ refers to the prior belief state, which is being updated using the likelihood of the observed data $p(y|\theta)$ to arrive at the posterior belief state $p(\theta|y)$. Bayes' theorem is being used extensively in statistics (Blitzstein & Hwang, 2015; Gelman, 2014), but more importantly for the present application, Bayesian inference has become a popular model of behavioral and neural functioning because organisms in the real world constantly use new information and prior knowledge (e.g. common sense) to interact with their environment (Chater et al., 2006; Darlington et al., 2018; Gershman, 2015; Ma et al., 2006).

Probably the most important extension to Shepard's law was given by Tenenbaum and Griffiths (2001a). While the original formulation is Bayesian in nature, Tenenbaum and Griffiths (2001a) rephrased the problem in the context of rational analysis (which was formally developed *after* 1987) and gave a full Bayesian treatment of the problem. In particular, they expressed the problem of generalization while emphasizing three main points that map onto aspects of Bayesian inference. First, the prior knowledge of organisms that maps onto the prior probability distribution. Second, the inference about new stimuli given the current belief state, which is closely linked to the prior (eq. Equation 1.6) or posterior predictive distribution (Equation 1.7, Gelman, 2014), depending on whether any new information has been given or not. In contrast to the prior and the posterior distribution, which are distributions over parameters, those are distributions over possible novel outcomes $\tilde{y}$ while the parameters are being integrated over:

$$p(\tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)d\theta. \tag{1.6}$$

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \tag{1.7}$$

Lastly, the mechanism by which new information is included refers to both the likelihood function and Bayes' theorem in general. Given this mapping, they derived that Bayesian inference is an optimal solution to the posed problem.

Just like Shepard (1987), they assumed that there is a set of possible consequential regions $h \in \mathcal{H}$ and some prior belief state $p(h)$ over this set, i.e. which region is how likely a priori. However, they relaxed the constraints that Shepard (1987) imposed on the shape of the consequential regions. As a consequence, the posterior can be computed over arbitrary hypotheses sets $\mathcal{H}$ and those can be specific to different contexts. For instance, multiples of 10 seem more similar than e.g. the set $\{34, 35, 36\}$ in some contexts, but not in others — a fact than can be expressed in differences in $\mathcal{H}$ and $p(h)$. Agents in the updated model are assumed to use Bayes' theorem to update their belief state about the relative plausibility of different consequential regions:

$$p(h|y) = \frac{p(y|h)p(h)}{\sum_{h \in \mathcal{H}} p(y|h)p(h)}. \tag{1.8}$$

The most important contribution of the updated model is the addition of the *size principle.* The likelihood of Shepard (1987) only considers whether or not the stimulus is contained in a consequential region:

$$p(y|h) = \begin{cases} 1 & \text{if } y \in h \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

This approach assumes that consequential regions and stimuli are sampled independently, which Tenenbaum and Griffiths (2001a) refer to as *weak sampling.* However, this does not differentiate between small and large regions. In the idealized case with a single observation, this does not have a big impact. But since the full Bayesian formulation of Tenenbaum and Griffiths (2001a) allows for multiple observations, it is important to account for the fact that one would likely not generalize beyond a narrow range in psychological space, if all consequential observations are very similar and one assumes that stimuli and consequences do not appear randomly. To account for this, the authors proposed another likelihood function in which the stimuli are sampled *from* the consequential region and where the probability for being sampled is inversely proportional to the size of the region:

$$p(y|h) = \begin{cases} \frac{1}{|h|} & \text{if } y \in h \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

Tenenbaum and Griffiths (2001a) refer to this likelihood as *strong sampling.* This updated function has the effect that the same observation is more likely given a smaller than a larger consequential region and as a consequence, both larger number and smaller variations in consequential examples lead to less generalization, which makes sense intuitively and has been shown to be true in human experiments (J. C. Lee, Lovibond, & Hayes, 2019). Given a belief state over the plausibility of consequential regions (i.e. $p(h)$ or $p(h|y)$), an agent can use *hypothesis averaging* to compute their expectation about whether a novel stimulus is part of the same region C. For this purpose, they would sum the probability of all consequential regions that contain the novel stimulus:

$$p(\tilde{y} \in C|y) = \sum_{h:\tilde{y} \in h} p(h|y). \tag{1.11}$$

Since Tenenbaum and Griffiths (2001a), incremental changes have been proposed.

Navarro et al. (2008) introduced a continuous hypotheses space and derived an analytic way to solve the model. Navarro et al. (2012) introduced *mixed sampling*, which is a weighted average of weak and strong sampling and showed that most people seem to use this approach. Soto et al. (2014) elegantly applied the same model structure to compound generalization[17] and added the idea of latent causes to the model (Gershman et al., 2015). Voorspoels et al. (2015) integrated learning from non-consequential observations and introduced *helpful sampling*. This sampling scheme assumes that stimuli and consequences are provided by a helpful mentor in a way that is most likely to result in the correct inference. This approach has a lot of face validity in the context of education where teachers are assumed to provide the most helpful examples. But all of those extensions follow the same basic structure of the model of Tenenbaum and Griffiths (2001a), which has proven to be a powerful tool.

1.2.3 The relationship with stimulus generalization

As mentioned before, the development of the Bayesian inductive reasoning approach has been almost completely independent of research on stimulus generalization until recently. One reason for this is that these models assume a deterministic outcome structure. A stimulus either always or never leads to an outcome. This is also apparent in the kind of tasks that are used to test predictions of the model (e.g. J. C. Lee, Lovibond, & Hayes, 2019; Navarro et al., 2008; Voorspoels et al., 2015). While stimulus generalization designs typically introduce an emotional experience to the study design, inductive reasoning is concerned with logical thinking under uncertainty. Typically, subjects are provided with a number of consequential examples (e.g. stimulus x has some property) and then queried on new stimuli (e.g. does stimulus y also have this property?). Until Voorspoels et al. (2015) added a treatment of learning from negative evidence (i.e. stimulus x does *not* have this property.), only positive examples were considered. The assumption of a deterministic outcome structure is deeply rooted in the idea of fixed consequential regions. Unfortunately, it is incompatible with the typically probabilistic reinforcement schedule in cognitive neuroscience, in which a stimulus is followed by a consequence only in a subset of trials.

The only attempt to unify inductive reasoning models with stimulus generalization was made by J. C. Lee, Lovibond, Hayes, and Navarro (2019). Given the similarities between the two approaches, they queried whether stimulus generalization followed the same principles as inductive reasoning using two experiments and an adapted Bayesian model. In their experiment, they found that behavior was in line with the predictions of their model and concluded that inductive reasoning and stimulus generalization rely on a similar mechanism. Their model differs from inductive reasoning models in that it does not assume deterministic consequential regions. Instead, agents learn an *association map*, which is a mapping from psychological space onto outcome probabilities. Generalization is possible via a smoothness constraint which dictates that similar stimuli have similar outcome probabilities. Interestingly, this model allows for a dimensional preference. This preference is implemented via

[17] Compound generalization is the ability to apply generalizations to novel combinations of stimuli.

parameters that govern the probability of rapid changes in outcome probabilities within each dimension. This model should in principle be able to implement dimensionality reduction by assuming very low granularity along a certain dimension and thereby foreshadows my own attempt to add representation learning to the equation.

1.3 Generalization in Reinforcement Learning

RL is a formal description of learning from experiences and rewards (Sutton & Barto, 2018). The main assumption is that agents try to maximize a time-discounted expected return in the future. That is, they learn to take actions in specific states that will generate the most beneficial outcomes while devaluing rewards that are further away in the future. Formally, RL is a collection of algorithms to solve a specific class of tasks, which are called Markov decision processes (MDPs). MDPs are characterized by a state space, an action space, a reward function, and a transition function. The state space is the set of all possible states, i.e. all possible combinations of environment variables that an agent can be in. A state could be e.g. the constellation of chess figures on the board or a physical location including all environmental variables for a more realistic scenario. The action space is the set of all possible actions that an agent can take in every state. These actions could be moving a chess piece according to the rules or choosing a stimulus in a simple neuroscientific experiment. The reward function is a mapping from states and actions to a scalar reward value. And finally, the transition function is a probability distribution over the next state given the current state and the action taken (Sutton & Barto, 2018).

RL agents solve a MDP by trying to maximize the expected return in the future, i.e. the sum of all rewards they they receive in the future while discounting those that are further away. This is done by learning the value of each action in each state, given by the *state-action value* function $Q(s, a)$, which is the expected return when taking action a in state s:

$$Q(A = a, S = s) = \mathbb{E}[R|A = a, S = s] \tag{1.12}$$

In addition, agents learn a policy π, which is a behavioral rule that dictates their behavior, conditional on the state-action value function. A policy is a probability distribution over actions in each state:

$$\pi = p(A = a|S = s) \tag{1.13}$$

Finally, given the policy and the state-action value function, we can define the *state value* function $V(S)$ by weighing the expected return of each possible action by the probability of taking said action:

$$V(S = s) = \sum_{a \in A} p(A = a|S = s)Q(A = a, S = s) \tag{1.14}$$

These two concepts, value functions and policies, are intertwined as the expected time-discounted return in the future clearly depends on the behavior of the agent, which depends on the policy. In order to learn both at the same time, agents use a behavioral algorithm called generalized policy iteration (GPI). GPI is characterized by a back and forth between the learning of a value function and a change in policy. Other approaches include off-policy learning (Sutton & Barto, 2018), e.g. Q-learning, in which agents learn the value function for the optimal policy while behaving according to any policy that allows for some amount of exploration. State-action values are updated via the Bellmann's equation using a prediction error (Sutton & Barto, 2018). The prediction error is defined as the difference between the current state-action-value and the sum of the immediate reward following this action as well as the state-value of the following state. This update equation looks like this:

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \alpha(R + V_t(S_{t+1}) - Q_t(S_t, A_t)). \tag{1.15}$$

In this context α refers to the *learning rate* and dictates how quickly agents update their value estimates. Since many experiments in cognitive neuroscience use episodic task with a single state transition, the expected reward in the future is just the expected immediate reward and the update equation simplifies to the instrumental version of the Rescorla-Wagner model[18]: (Rescorla & Wagner, 1972)

$$Q_{t+1}(S_t, A_t) = Q_t(S_t, A_t) + \alpha(R - Q_t(S_t, A_t)). \tag{1.16}$$

RL has been a success story both in AI and cognitive neuroscience (Dayan & Niv, 2008; Mnih et al., 2015; Sutton & Barto, 2018). In the latter, RL theory has been tremendously successful in characterizing dopaminergic signalling in the striatum (Schultz et al., 1997) and explaining human behavior in decision making tasks (Dayan & Niv, 2008). These results suggest that RL is a feasible approach to describe human learning in tasks that show a resemblance to a MDP.

A large proportion of research in RL uses simple designs with a finite set of stimuli and without the need to generalize. In these contexts, simple models like Rescorla-Wagner models and simplified temporal-difference learning work remarkably well as measured by their ability to explain behavioral data (Dayan & Niv, 2008). Since these models do not formally provide a way to share knowledge between states or actions, agents need to encounter every state and attempt every action in order to learn the respective expected return and to derive an optimal policy. This is feasible in the context of a small number of states and actions but quickly becomes intractable in even moderately complex problems (Sutton & Barto, 2018). Unfortunately, the real world is much more complex than the simple MDPs that are used in RL research, a problem that has been noted and led to an increased interest in generalization in the context of human research (Niv, 2019; Wu et al., 2018), AI (Lehnert et al., 2020) and somewhere in between (Flesch et al., 2022).

[18] While omitting the β parameter.

1.3.1 Generalization as abstraction

A common theme that has emerged in this literature is the idea of generalization as abstraction of state spaces or task structures (Lehnert et al., 2020; Niv et al., 2015). Conceptually, this means that agents do not learn about the full state space including variables that are irrelevant for the task at hand, but instead learn value functions on a reduced space. This approach aids in generalization as agents have to deal with a lower number of possible states and because an abstracted state space is more likely to be shared between tasks than a full state space. Abstraction *can* be achieved via dimensionality reduction. Because this has been one of the most productive approaches in human RL research, I will expand on this in the next paragraph. However, there are other ways to achieve abstraction, e.g. by exploiting correlations in the reward structure. For instance, Wimmer et al. (2012) found that subjects learned to use correlated rewards between stimuli to guide their behavior in the context of changing reward probabilities. Interestingly, this kind of learning seems to rely on the DMN, which provides an interesting perspective on the interpretation of neural findings in stimulus generalization. Using a similar, but more sophisticated approach to abstraction, Wu et al. (2018) used an extensive state space, where every state consisted of a *one armed bandit* and found that the underlying generalization can be well described as Gaussian process regression, which indicates that subjects learned to exploit the correlational structure in the state space.

1.3.2 Dimensionality reduction

The simplest way to reduce the complexity of high-dimensional spaces is to reduce the number of dimensions. Mathematically speaking this is equivalent to a projection onto a low-dimensional space using e.g. principal component analysis (PCA). Less sophisticated, but still effective is to just ignore a subset of the dimensions. This proposal is interesting, because it is backed up by a general tendency to simplify learning problems (Galdo et al., 2022). In addition, it opens up a door to the vast literature on low-dimensional neural codes (Badre et al., 2021; Bernardi et al., 2020; Bottini & Doeller, 2020; Fusi et al., 2016; Summerfield et al., 2020) and selective attention (B. A. Anderson & Yantis, 2013; B. A. Anderson et al., 2011; Brosowsky & Crump, 2021; Markovic et al., 2014; Yantis, 2008), that can be understood as neural and behavioral dimensionality reduction respectively. Indeed, dimensionality reduction is one of the main mechanisms that is being explored in the context of representation learning (Badre et al., 2021; Niv, 2019), latent causal structure learning (Eichenbaum et al., 2020; Tomov et al., 2018) and theoretical work on reusable neural codes (Bottini & Doeller, 2020; Fusi et al., 2016).

Representation learning is concerned with the mechanism by which agents discover a low-dimensional representation (Niv, 2019), that is appropriate for the task at hand (Badre et al., 2021; Brosowsky & Crump, 2021; Loose et al., 2017). While a good representation facilitates learning, it adds another step to the learning process, namely the discovery of said representation. To investigate how this is implemented in the brain, Niv et al. (2015)

used stimuli that differed on three dimensions. In each trial, subjects had to choose one of three stimuli. Unbeknownst to them, only one dimension was predictive of reward in each block. Thus, subjects had to learn the relevant dimension *and* the rewarded feature on that dimension simultaneously. Niv et al. (2015) used a RL model to quantify the subjective uncertainty about the relevant dimension and reported more activity in the FPN when there was more uncertainty. This finding is compatible with a two-step process — learning *about* dimensions and learning *on* dimensions — and implies a role of the FPN in the first step. In this context, Leong et al. (2017) suggested that the FPN is responsible to reallocate attentional resources to different stimulus dimensions, which is consistent with the results of Niv et al. (2015) and can be understood as a rescaling of psychological space. Interestingly, when summarizing this line of research, Niv (2019) emphasized the role of the FPN in the discovery of an appropriate representation and the allocation of attention, but located the actual encoding of the learned representation to the orbitofrontal cortex (OFC). While this interpretation is consistent with results that show a role of the OFC in state space representations and cognitive maps (Basu et al., 2021; Jones et al., 2012; Schuck et al., 2016) and the suggested role of the FPN in selective attention (Yantis, 2008), it is somewhat at odds with other results, that suggest that the FPN directly encodes representations (Badre et al., 2021; Bernardi et al., 2020; Flesch et al., 2022; Jackson et al., 2017). I will discuss this discrepancy in more detail in the next section.

Research on latent structure learning has extended this literature by adding a hierarchical task structure to the mix. Typically this has been done by contrasting cues with context. The latent structure to be discovered are then different cue rules, depending on the context, which maps onto different abstractions, or more specifically, the reduction of different dimensions, depending on the context. For instance, Tomov et al. (2018) conducted a study in which different stimulus features were predictive of a reward. This was dependent on the context, which was another symbol that was displayed to subjects. Corroborating the results of Niv et al. (2015), they found a structure learning signal in the FPN. They also report an encoding of the full latent structure in the posterior hubs of the FPN and the aIC. Eichenbaum et al. (2020) used a similar paradigm and could corroborate these findings quite closely. Based on their results, they suggest that the FPN is responsible for the search for and the discovery of a hierarchical task structure while the SN[19] is responsible for the encoding of this structure and the transfer to behavior. Slightly differing from these findings, Vaidya et al. (2021) used a similar cue-context paradigm, but reported that the FPN and the aIC directly encode the hierarchical task structure.

In summary, research on generalization in RL has overwhelmingly suggested a reliance on abstracted representations. This has been investigated in representation learning (Niv, 2019) and latent structure learning (Tomov et al., 2018) as well as using correlated reward structures (Wu et al., 2018).

[19] Nomenclature of brain networks is very messy (Uddin et al., 2019). While the authors call it the cingulo-opercular network, this is equivalent to the SN.

1.3.3 Neural mechanisms

Just like in the context of stimulus generalization, a consistent result has been the involvement of regions of the FPN, the SN and the DMN, albeit some contradictions about their specific role remain. In the following I will discuss those in the context of neural concepts that seem relevant, namely task representations outside of generalization (Jackson et al., 2017; Woolgar et al., 2011) and the geometry of neural representations (Badre et al., 2021; Fusi et al., 2016).

Task representations. To resolve a conflict between competing interpretations of the role of brain regions or networks, it can be helpful to consider related literature. In the present case, when debating whether the FPN and the SN discover or encode abstracted task representations *or* even both, I am going to review some examples from the literature on task and stimulus representations that are relevant to, but not specifically concerned with generalization.

Woolgar et al. (2011) used a simple design, in which subject had to learn a mapping from keys to outcomes, while the mapping depended on the context. In order to distinguish which task features were most strongly encoded, they used MVPA. This analysis revealed that both the FPN and the aIC most strongly encoded the task rule, i.e. the relevant mapping.

Jackson et al. (2017) leveraged a categorization task in order to investigate whether adaptive neural encoding is a potential neural equivalent to flexibly changing selective attention. In this task, subjects had to categorize items that differed along two dimensions. Between blocks, the relevant dimension changed. The authors found that neural representations in the FPN were flexibly updated and consistently represented the relevant dimension more strongly than the irrelevant dimension.

In a similar vein, Flesch et al. (2022) employed a categorization task as well in which stimuli differed along two dimensions, while the relevant distinction was along either axis. They corroborated the findings of Jackson et al. (2017) and found low-dimensional representations in dlPFC and IPS[20], where the relevant dimension was encoded more strongly. Interestingly, they reported converging evidence in that deep neural networks (DNNs, LeCun et al., 2015) and the frontal eye field (FEF) of monkeys showed the same pattern.

Adding to the evidence across boundaries between species, Bernardi et al. (2020) reported abstract, low-dimensional task encodings in the monkey dlPFC, anterior cingulate cortex (ACC) and HPC.

In the context of these results and other studies that report empirical evidence for the encoding of abstract task rules in the whole FPN or parts of it (e.g. Badre et al., 2010; Loose et al., 2017), it seems unlikely that the role of the FPN in generalization is constrained to the discovery of appropriate abstraction. Instead, they suggest a clear role in the encoding of those low-dimensional abstractions. That is, the FPN likely encodes the abstracted task and thereby provides this information to downstream areas.

[20] The major components of the FPN.

Geometry of neural representations. Another relevant concept, that is more strongly rooted in research on general coding properties of the brain is the geometry of neural representations (Fusi et al., 2016). In the generalization context, the most relevant property is the dimensionality of representations (Badre et al., 2021). As an example, consider a defined area in the brain of an organism. If we present different stimuli to this organism, who's brain we are measuring, we can quantify the activity of each unit of measurement. Depending on the modality and resolution of our measurement, we might consider single neurons, electrodes, or in the context of fMRI voxels. For each presented stimulus we can combine the activity of all units into a vector and thus understand the activity pattern as a point in high-dimensional space. The dimensionality of this space is equal to the number of measurement units. If we combine the vectors of all stimuli into a matrix, we can use linear algebra to quantify the rank of that matrix or to determine the number of relevant dimensions (Strang, 2021). Intuitively, if the activity pattern is the same, or a scaled version of a single pattern, to all stimuli, the rank of the matrix is 1: All patterns (points) lie on the same line. This would be the most extreme case of low-dimensional representations. If all patterns are completely different, the rank of the matrix is the number of stimuli, assuming that there are more units of measurement than stimuli. This would be the other extreme, a very high-dimensional neural representation. In practice, the matrix is always going to be full rank due to measurement noise[21], but it is still possible to determine the relevant dimensions using PCA or similar techniques.

In an elegant review, Fusi et al. (2016) discussed the role of high- and low-dimensional representations in higher cortical areas. High-dimensional representations are easily distinguishable from each other. That implies that downstream areas, i.e. neuronal assemblies that receive input from the higher cortical areas, can differentiate between them easily via linear readouts, which allows for flexibility in behavior. In contrast, low-dimensional representations are not easily differentiable and get treated as a single entity by downstream areas as they do not encode aspects of certain inputs that would be needed to distinguish them. This comes at a cost in flexibility but allows for more robustness. As an example, Fusi et al. (2016) discuss categorization tasks, where it is beneficial to only encode those features that determine class membership. Similarly, low-dimensional representations are an obviously important consideration for generalization, since representational similarity can be used by downstream areas to implement adequate behavior towards new stimuli or situations.

Summerfield et al. (2020) linked structure learning, and thus a concept that is relevant for generalization, to low-dimensional representations in the parietal cortex — parts of which are important hubs of the FPN — and the hippocampus. Like Fusi et al. (2016), they argue for the importance of low-dimensional representations in behavioral robustness towards new situations. In addition, they suggest two interesting points. First, representations in the posterior parietal cortex (PPC) tend to be often one-dimensional, i.e. *extremely*

[21] Even multiple measurements of activity to the same stimulus are practically guaranteed to be somewhat different.

low-dimensional. Some evidence suggests that this allows for generalization between task domains, e.g. the same representations could be used context-dependent and encode either direction in physical space, time or numerosity. For instance, domains that share a conceptual structure with physical space, such as time or interpersonal relationships, could be encoded in the same way. This phenomenon is called primary conceptual metaphors (Bottini & Doeller, 2020). Second, they argue for a complementary role of *cognitive maps* in the HPC and enthorhinal cortex (EC) on one and the PPC on the other hand. In this dichotomy, the HPC/EC would contain allocentric cognitive maps, i.e. encode the relationship or spatial distance between concepts or stimuli relative to each other. In contrast, the PPC would encode a more egocentric representation, i.e. encode the relationship or physical distance of stimuli to the self. These differences in localization could explain discrepancies in neural results from generalization studies when considering that for some tasks it might be more relevant to encode the similarity of stimuli relative to each other, while in other tasks the similarity of stimuli relative to the self might be more relevant.

A review on cognitive maps by Bottini and Doeller (2020) makes a very similar proposal and suggests that a hippocampal-parietal system combines allocentric and egocentric representations. They also argue for the importance of low-dimensional neural codes for the discovery of similarities between stimuli and situations. The latter point nicely illustrates the relevance of low-dimensional representations for generalization. It is possible that perceived similarity between stimuli or situations is encoded via similar neural representations and that this is the basis for generalization.

Badre et al. (2021) reviewed the importance of the dimensionality of neural representation with respect to cognitive control and task appropriate behavior. Like others, they emphasize the distinction between high- and low-dimensional representations, where high-dimensional representations allow for linear readout by downstream regions and thus behavioral flexibility while low-dimensional representations ensure robustness and generalization. Since the dlPFC, an important part of the FPN, is involved in cognitive control, they emphasize representations in this area. Interestingly, different studies have found near maximal (i.e. very high) dimensionality in this area (Rigotti et al., 2013), while others reported lower dimensionality (Bernardi et al., 2020). Badre et al. (2021) argue that this discrepancy might reflect task demands and suggest that the dlPFC can flexibly switch between a high-dimensional representation when small differences are relevant and a low-dimensional representation when generalization is more important.

The consideration of the respective utility and suspected role of low- and high-dimensional representations shows the importance of representational geometry with respect to generalization, bridges the gap between work that focusses on fundamental coding patterns in the brain vs. the neural basis of cognition and suggests that the former can help with an interpretation of research on the latter. Given the focus of representation learning on the role of the FPN and converging evidence for a involvement of this network in stimulus generalization, it is surprising that the literature on the geometry of neural representations is hardly considered in either of those disciplines.

1.4 An integrated view

In the previous chapters I have reviewed three different sub-disciplines that investigate generalization in different contexts:

1. Stimulus generalization in associative learning, often concerned with the generalization of fear learning.
2. Models of inductive reasoning in cognitive science that use rational analysis to understand mechanisms of reasoning under uncertainty.
3. Lastly, representation learning in RL that is concerned with the discovery and encoding of low-dimensional abstractions that facilitate generalization.

In addition, I summarized key findings and opinions on the geometry of neural representations and the encoding of task structures that seem relevant. In this section, I will provide an integrated view and identify commonalities in behavioral, computational and neural aspects. In this view, I will argue that there is substantial reason to assume an underlying process that is common to all three applications and that it can be expressed reasonably well in a Bayesian model that accounts for dimensionality reduction in a rational manner.

Behavioral aspects. Research designs in inductive reasoning (e.g. Navarro et al., 2008; Voorspoels et al., 2015) and stimulus generalization (e.g. Dunsmoor et al., 2011; Onat & Büchel, 2015) are quite similar to begin with. In fact, even the description of the relevant task that was given by Shepard (1987) can be considered stimulus generalization. In his example, a bird is confronted with the task of deciding which worms to eat after having had a bad encounter with a particular worm. However, the research field that has emerged from the work of Shepard (1987) has conceptualized generalization more strongly as inductive reasoning, which has led to differences in study designs and scope with respect to stimulus generalization. The typical difference in design between the two disciplines lies in the quality of an encounter: Stimulus generalization uses actual experience, e.g. an electric shock, while inductive reasoning confronts subjects with true or false statements and asks them to use those to reason about new statements. A typical example for that would be: Given that the statement „Object A has property X.“ is true, how likely is the statement „Object B has property X.“ to be true? Recently, J. C. Lee, Lovibond, Hayes, and Navarro (2019) came to the conclusion that fear generalization and inductive reasoning rely on similar mechanisms.

Similarly, designs in associative learning and RL often are not as different as it may seem because RL studies often use a very simplified task that can be modeled using associative learning models like the Rescorla-Wagner model (Rescorla & Wagner, 1972). However, studies on generalization in RL are typically more concerned with the mechanism of building an abstraction and do not use generalization gradients like stimulus generalization studies do. Instead they focus on multidimensional stimuli (Niv et al., 2015; Tomov et al., 2018) or correlated state spaces (Wimmer et al., 2012; Wu et al., 2018). This makes a comparison harder, but if we consider studies on stimulus generalization that used multidimensional stimuli, we can see that the effects are commensurate with the idea of dimensionality reduction (Ahmed & Lovibond, 2015a, 2015b; Vervliet et al., 2010). Some studies used an instrumental generalization tasks in conjunction with an associative learning phase (Lommen et al., 2010; Norbury et al., 2018; van Meurs et al., 2014). This approach can be considered somewhere in between associative learning and RL. This is because RL is concerned with learning the value of actions while associative learning is concerned with learning the predictive value of stimuli. These studies employed a purely associative learning phase, but then investigated how the learned associations translate into behavioral choices. A consistent result in those studies was that instrumental generalization gradients are closely related to associative ones.

Computational aspects. Even more interesting and enlightening are similarities in how generalization is conceptualized and in the underlying computations. Strikingly, the tasks that associative learning and RL solve are much more closely related than one might think. RL algorithms are used to solve MDPs. Associative learning omits the consideration of actions, which results in something that Sutton and Barto (2018) call Markov reward processes (MRPs). MRPs and MDPs are closely related, as are state values in RL and associative values in associative learning. Another notable parallel emerges when comparing consequential regions from inductive reasoning models with state space abstractions in RL. It seems to me that consequential regions can be understood as abstractions of psychological space. Consequently, learning the correct consequential region is akin to learning an adequate abstraction. Adding to this, the psychological space is an adequate abstraction to begin with. Shepard (1987) argues that this space is shaped by evolution, but the distinction between ontogenetic and phylogenetic learning in cognition is very hard to make and recent theories have suggested that an appropriate hypotheses set of consequential regions is at least partly learned within a lifetime (Austerweil et al., 2019). For this reason I suspect that the consequential region approach and the dimensionality reduction (or more generally abstraction) approach refer to very similar things. Lastly, the only inductive reasoning model that has been adapted to associative learning showed computational similarities between inductive reasoning and stimulus generalization (J. C. Lee, Lovibond, Hayes, & Navarro, 2019) and already implemented a notion of dimensional relevance.

Neural aspects. Since no neuroimaging studies on inductive reasoning (in the sense of Shepard (1987)) have been conducted, I will focus on results from stimulus generalization and RL. The most striking similarity is the complete overlap in brain areas and networks. Fear generalization studies consistently report positive generalization gradients in the FPN and the SN and negative gradients in the DMN (Webler et al., 2021). Studies on representation learning have emphasized those regions as well, with a role for the FPN and the aIC[22] in the discovery and encoding of task and state space abstractions (Leong et al., 2017; Tomov et al., 2018; Woolgar et al., 2011). Other studies have suggested a similar role for the HPC and vmPFC or OFC, which are parts of the DMN (Niv, 2019; Schuck et al., 2016).

[22] The aIC is a prominent part of the SN.

This discrepancy is not solved as of now, but it might have to do with study designs. In analogy to allocentric vs. egocentric cognitive maps and their respective localization in the brain (Bottini & Doeller, 2020; Summerfield et al., 2020), it is possible that other aspects of the task, like the presence of rewards or the inclusion of actions might influence where in the brain those aspects are encoded.

Interpretations of the role of those structures differ even more wildly between stimulus generalization and RL (Badre et al., 2021; Niv, 2019; Webler et al., 2021). But the strong overlap is a promising sign for a common neural mechanisms and I suspect that some of the contradictions can be resolved by making research designs more similar. For instance, there is only a single neuroimaging study on stimulus generalization that uses multi-dimensional stimuli until now (Onat & Büchel, 2015). Unfortunately, this study explicitly assumed those stimuli to be arranged on a one-dimensional subspace and analyzed the data accordingly. Given this, it is not surprising that dimensionality reduction has not been discussed in the context of stimulus generalization.

1.5 Contributions of this thesis

With this thesis I intend to contribute towards discovering a common mechanism for generalization in different contexts. I have outlined my reasoning for this assumption in the introduction. For this purpose, I will propose a novel Bayesian model that is similar to previous approaches (J. C. Lee, Lovibond, Hayes, & Navarro, 2019), but emphasizes dimensionality reduction as rational behavior. This resolves the contradiction between stimulus generalization and associative learning on the one and representation learning on the other hand while also omitting the distinction between learning *about* dimensions and learning *on* dimensions. This model is described and motivated in chapter 2. Since this approach is relying on an adequate way to control for psychological spaces and the method of Shepard (1987) using NMDS is not applicable to the kind of data I will report, I also took the liberty to develop a new hierarchical Bayesian method to estimate psychological spaces. I describe this method in chapter 3. To test the predictions of my model I also conducted three studies. The first study uses fear conditioning and is purely behavioral. In the second study, I replicate those findings and collected fMRI data to query the neural computations that underlie the behavioral mechanism. Lastly, in a third study, I used appetitive instead of aversive conditioning to show the scope of the proposed model beyond fear generalization. These studies are described in chapter 4.

2 An integrated Bayesian model of generalization

After having reviewed the literature, we now turn to the question of what a unifying model of generalization could look like. As I have argued in chapter 1, such a unifying approach to generalization needs to account for stimulus generalization, inductive reasoning and representation learning. While this list might not be exhaustive, it seems like a good starting point. Stimulus generalization and inductive reasoning have been linked before (J. C. Lee, Lovibond, Hayes, & Navarro, 2019), but an approach that accounts for all three is missing as of yet. In this section I will first argue why previous Bayesian models do not account sufficiently for the other fields. In an attempt to unify all three lines of research, I will describe a novel Bayesian model of stimulus generalization, that integrates aspects from inductive reasoning and representation learning.

2.1 Limitations of previous models

The biggest limitation of almost all previous Bayesian models (Navarro et al., 2008; Shepard, 1987; Soto et al., 2014; Tenenbaum & Griffiths, 2001a, 2001b) is the assumption of a deterministic consequential region. In the view of Shepard (1987), consequential regions map onto natural kinds. Possibly due to his focus on single-shot learning, i.e. learning from a single observation, this concept does not allow for variation in outcomes. It seems plausible that eating a certain worm sometimes, but not always, leads to a certain outcome, but the concept of deterministic consequential regions is incompatible with that. In the literature that descended from Shepard's work, the focus on generalization as inductive reasoning is even stronger (Navarro et al., 2008; Tenenbaum & Griffiths, 2001a). This can be seen in the research designs in which the statements that subjects are confronted with assume that a certain stimulus *has a property* as compared to *leads to an outcome*. A stimulus has a property or it does not, there is no variation. In other words, these models account for epistemic uncertainty with respect to the *true* consequential region. All variation in generalization gradients are due to this uncertainty. But this approach leaves out aleatoric uncertainty, i.e. uncertainty about the outcome of a stimulus that is not due to a lack of knowledge but due to inherent randomness in stimulus-outcome contingencies (Tenenbaum & Griffiths, 2001a). Such a randomness is given in almost all studies on stimulus generalization and in RL due to probabilistic reinforcement schedules. As a consequence, consequential region models are not applicable to these studies. The mathematical reason for that can be found in the likelihood function, independently of the sampling assumption (Navarro et al., 2012). The likelihood of a consequential observation (i.e. a stimulus followed by an outcome) is positive for any consequential region that contains this stimulus. Likewise, the likelihood of a non-consequential observation is 1 given any consequential region that does *not* contain this stimulus. However, a non-consequential observation given a region that *does* contain this stimulus is 0. This leads to a posterior probability of 0 for this region. Since Bayes' rule (Equation 1.3) uses multiplication to update the belief state, no further observations can change this. Beyond the pure applicability, the deterministic outcome assumption has a conceptual problem. If we account for randomness in outcomes, a next step would be to link the outcome probability to the relative position within the consequential region. To illustrate this point, imagine two glasses of water. One is very slightly contaminated with dirt and the other one is dark brown. Those glasses can be viewed as points in psychological space (Shepard, 1987). Both have a certain probability of leading to the same outcome, namely getting sick, i.e. they are members of the same natural kind. But if we gave both glasses to thirsty subjects and they had to choose one to drink from, it is very likely that most subjects would choose the less contaminated water because they would infer that it is less likely to make them sick. This difference is not due to uncertainty about the location of the consequential region but due to inherent differences in outcome probabilities. This thought experiment is evidence for a generalization process that accounts for the relative position in a consequential region. A general theory of generalization should account for that.

Some models *do* account for randomness in outcomes. Soto et al. (2014) proposed a model that explains phenomena in compound generalization with the assumption of latent causes. The learner infers the latent causes that are active in any given trial. Each latent cause leads to a certain outcome magnitude and the expected outcome is the sum of those weights over all active causes. The *real* outcome is than assumed to be drawn from a normal distribution with the expected outcome as mean and some fixed very small variance. While this accounts for some variation in outcome magnitude, the reason for it is mostly to arrive at a probabilistic model. It also does not explain different outcome probabilities and consequently does not link those to the relative position within a consequential region.

I am aware of only one model that is applicable to probabilistic reinforcement. J. C. Lee, Lovibond, Hayes, and Navarro (2019) applied an inductive reasoning model to fear generalization and proposed an alternative formulation that omits consequential region. Instead, the agent learns an *association map*, which is essentially an outcome probability for every point in psychological space. A smoothness constraint enables agents to generalize to unobserved stimuli. This model can predict behavioral data from a fear generalization experiment very well (J. C. Lee, Lovibond, Hayes, & Navarro, 2019).

The second limitation of the mentioned Bayesian models is that they do not account for representation learning. This is not a shortcoming in and of itself, but it is a lacking feature for a common mechanism of generalization. Some aspects of models can be considered as going in that direction. Shepard (1987) emphasized that psychological spaces are shaped evolutionary. This implies the possibility that irrelevant dimensions are scaled differently than relevant dimensions. But it does not explain context-dependence, which is necessary because some dimensions are relevant in some but not in other circumstances. Also missing is a mechanism to rescale those dimensions as a reaction to learning experiences. The models of Navarro et al. (2008) and Soto et al. (2014) assume a probability distribution over the size of consequential regions along every dimension. In principle this allows for different relevance of dimensions. But those models do not account for variation in outcome probabilities. Lastly, the Bayesian model of associative learning by J. C. Lee, Lovibond, Hayes, and Navarro (2019) includes a parameter that indicates the dimensional relevance. This parameter acts as the prior probability of any point being a *mutation point*, i.e. it governs the smoothness of this dimension. This approach is an important step, but it does not explicitly include dimensionality reduction.

As it turned out, this is relatively straightforward to implement in a Bayesian model. I will formulate such a model in the following sections.

2.2 Conceptual description

Conceptually, the proposed model is similar to the model of J. C. Lee, Lovibond, Hayes, and Navarro (2019) in that it omits deterministic consequential regions and instead assumes that agents learn an *associative map*. But in contrast to J. C. Lee, Lovibond, Hayes, and Navarro (2019), I impose more structure on the associative map. This comes at a cost of flexibility, but it allows for a straightforward implementation of dimensionality reduction.

The basic idea is that agents have a belief state about the midpoint and the decay of the associate map, i.e. where in psychological space the map is centered and how quickly probabilities decrease with distance to the midpoint. In addition, they learn about the outcome probability at the midpoint. The shape of the associative map needs to be defined *a priori*. Empirical results suggest either a Gaussian or exponential shape (Ghirlanda & Enquist, 2003). In line with Shepard (1987), I chose an exponential shape, although arbitrary shapes are possible (Tenenbaum & Griffiths, 2001a). Extensions of the model should implement a distribution over different shapes for a more general model. An example for an exponential associative map is shown in Figure 2.1.

Any set of values for the parameters of the associative map defines a probability of an outcome for each stimulus. This probability is part of the likelihood and is being used to update the aforementioned belief state about the parameters according to Bayes' rule. An important feature is that different values for the decay on different dimensions are mathematically equivalent to a rescaling of dimensions. This point becomes clearer in section 2.3. For extremely low values for the decay along a dimension, all points along that dimension are treated equally, as shown in Figure 2.2. This is how the model implements full dimensionality reduction. In addition it allows for partial reduction by scaling dimensions differently.

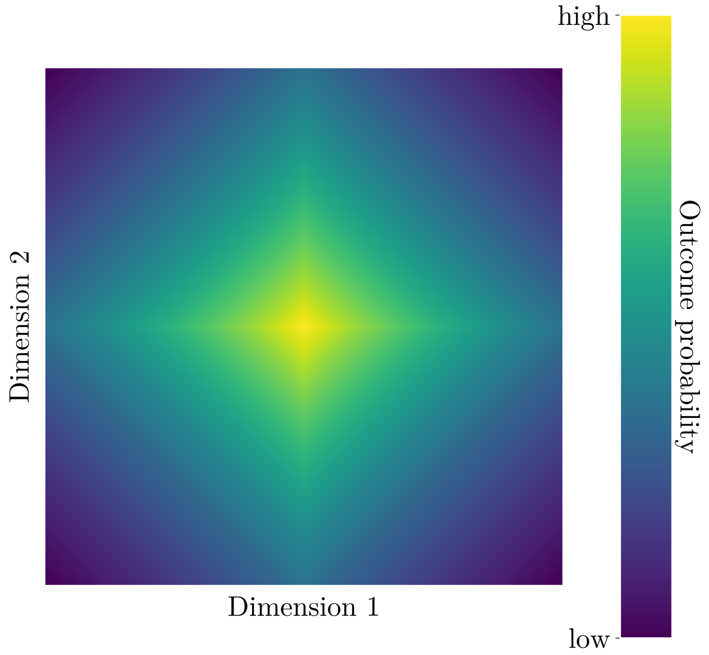


Figure 2.1: An example for an associative map in two-dimensional space. The outcome probabilities are given by the color. Probabilities decrease exponentially with distance to the midpoint. The strength of this decrease is given by the decay parameter.

Integral vs. separable dimensions. A noteworthy consideration in the context of this model is the *integrality* vs. *separability* of dimensions that has important implications on the shape of generalization (Soto & Wasserman, 2010). Integral dimensions are those that are not separable by perception. For example, the color of an object is integral, because it is not possible to distinguish the effects of hue and saturation. Separability is the opposite, i.e. dimensions that can be perceptually separated, like the size and color of an object. Those two types of dimensions lead to a different metric in psychological space. In theory, the appropriate distance between objects should be the Euclidean distance for integral dimensions and the *cityblock* (or *Manhattan*) distance for separable dimensions (Shepard, 1987). Both types of distances are special cases of the so-called Minkowski metric. Alternatively, the Minkowski metric is a generalization of both the cityblock and the Euclidean distance and is defined as

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{2.1}$$

where x and y are two points in n-dimensional space and p is the exponent. For $p = 1$, the Minkowski metric is equal to the cityblock distance, the Euclidean distance is defined as $p = 2$. This generalization allows for metrics that are neither nor, but somewhere in between for values of $1 < p < 2$. Soto and Wasserman (2010) showed in pigeons that p was closer to 1 for integral and closer to 2 for separable dimensions, but the best-fitting values

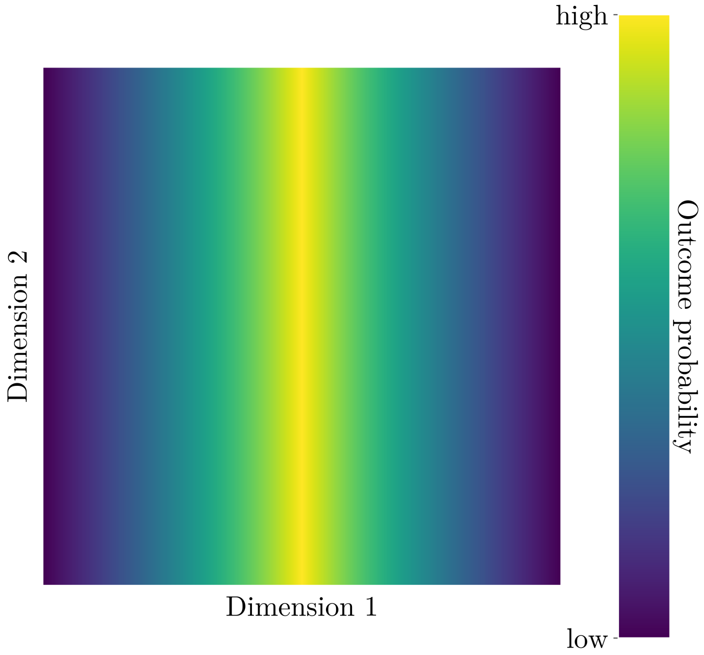


Figure 2.2: An example for dimensionality reduction. As the decay along one dimensions approaches 0, all points on this dimension have the same outcome probability. While this is visualized in a two-dimensional space, effectively both dimensions are reduced to one.

where somewhere in between.

Importantly, since integral dimensions are not distinguishable by definition, the concept of dimensionality reduction only makes sense in the context of separable dimensions. For this reason, I am only concerned with separable dimensions and accordingly use cityblock distance in the next section. The model is still generalizable to integral dimensions, if one assumes a single decay parameter and replaces the cityblock with the Euclidean distance.

2.3 Mathematical formulation

Following Shepard (1987), I consider stimuli to be arranged in a psychological space. An associative map in this space is defined by the midpoint $\boldsymbol{\mu}$, the decay $\boldsymbol{\lambda}$ and the outcome probability at the midpoint ρ. In a one-dimensional space, $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are scalars, in higher dimensional spaces they are vectors. The probability parameter ρ is always a scalar. For any given stimulus s, the outcome probability depends on the weighted *cityblock* distance[1] δ_s to the midpoint $\boldsymbol{\mu}$

[1] Intuitively, the cityblock or Manhattan distance is the sum of the absolute differences along all dimensions.

$$\delta_s = \sum_{i=1}^{d} \lambda_i |s_i - \mu_i|, \tag{2.2}$$

where d is the dimensionality of the psychological space. Note, that a change in λ_i has the exact same effect as rescaling the ith dimension. A value of 0 is equivalent to complete dimensionality reduction, while smaller values for one than the other dimension indicate partial dimensionality reduction. This way, dimensionality reduction is naturally included in the model while the fully Bayesian approach gives a rational interpretation and omits the distinction between learning *on* and learning *about* dimension since the model learns about both using the same likelihood function and rational update rule.

Accounting for perception. Assuming perfect perception, given Equation 2.2 the probability of observing an outcome r for stimulus s would then just be the exponential function of the negative weighted distance $-\delta_s$ multiplied by the outcome probability ρ:

$$p(r = 1|s) = \rho e^{-\delta_s}. \tag{2.3}$$

In reality, perception is not perfect and previous studies have established an important role for the accuracy of perception in generalization (Laufer & Paz, 2012; Laufer et al., 2016; Schechtman et al., 2010). It is therefore necessary to model the perceptual noise. Assuming Gaussian noise and a set of N stimuli, we can define a perceptual confusion matrix $\boldsymbol{A}$, that depends on the standard deviation of perceptual noise σ. In this matrix, the element $\boldsymbol{A}_{i,j}$ is the probability of perceiving stimulus s_j when the true stimulus is s_i:

$$\boldsymbol{A}_{i,j} = \mathcal{N}(s_i|s_j, \sigma^2 \boldsymbol{I})^2 \tag{2.4}$$

Because this approach is a discretization of the continuous psychological space, we need to normalize the rows, so that they sum to 1. Using the confusion matrix, we can derive the noise-accounted weighted distance δ^{noise} from a vector δ of noise-free distances:

$$\delta^{noise} = \boldsymbol{A}\delta \tag{2.5}$$

And finally, the exponential function of the negative vector multiplied with ρ gives the conditional outcome probabilities for all stimuli:

$$p(r = 1|s) = \rho e^{-\delta_s^{noise}} \tag{2.6}$$

$$p(r = 0|s) = 1 - \rho e^{-\delta_s^{noise}}. \tag{2.7}$$

Learning the belief state. So far we have treated $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and ρ as fixed, but because those are the parameters over which agents update their belief, they are variables. To account for

[2] The symbol $\boldsymbol{I}$ refers to the identity matrix. A scaled identity matrix as covariance matrix implies that perceptual noise is uncorrelated between dimensions and has the same variance along them.

the uncertainty about those variables, we need to condition on them:

$$p(r=1|s,\boldsymbol{\mu},\boldsymbol{\lambda},\rho)=\rho e^{-\delta_s^{noise}} \tag{2.8}$$

$$p(r=0|s,\boldsymbol{\mu},\boldsymbol{\lambda},\rho)=1-\rho e^{-\delta_s^{noise}}. \tag{2.9}$$

Integrating both possible outcomes into a single equation yields

$$p(r|s,\boldsymbol{\theta})=(\rho e^{-\delta_s^{noise}})^r(1-\rho e^{-\delta_s^{noise}})^{1-r} \tag{2.10}$$

where I summarize $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ and ρ in the vector $\boldsymbol{\theta}$ for readability. This is not yet the general likelihood function, because the likelihood depends on the stimulus *and* the outcome. In other words, we need to account for the sampling assumption, i.e. how the stimulus was generated. This implies that the likelihood function is a joint probability of stimulus and outcome $p(s,r)$, that can be factored into the sampling assumption and the conditional probability of an outcome according to the laws of probability (Blitzstein & Hwang, 2015):

$$p(s,r)=p(s)p(r|s). \tag{2.11}$$

The sampling assumption $p(s)$ is a probability distribution over the stimuli and depends on the context. Since the data I collected is best explained by weak sampling[3], $p(s)$ is constant and thus $p(s,r)\propto p(r|s)$, but other choice for $p(s)$ are possible and allow for a broad application of the model to different contexts. For weak sampling, Equation 2.10 *is* the likelihood, but a more general equation would be

$$p(s,r|\theta)=(\rho e^{-\delta_s^{noise}})^r(1-\rho e^{-\delta_s^{noise}})^{1-r}p(s), \tag{2.12}$$

because it accounts for different sampling assumptions, that can be included by plugging in an appropriate choice for $p(s)$.

The second crucial ingredient for the model is the prior on the parameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta})=p(\boldsymbol{\mu})p(\boldsymbol{\lambda})p(\rho). \tag{2.13}$$

This prior strongly depends on the psychological space, the relevance of dimensions and the context and is shaped by evolutionary processes and lifetime experiences. For example, there seems to be some prior knowledge about threatening stimuli that is independent of having had dangerous encounters with them (Öhman, 2009; Öhman & Dimberg, 1978; Öhman et al., 2001). Likewise, some stimulus dimensions seem to be more relevant than others with respect to certain outcomes. For instance, emotional expressions have a social signaling function and do not need to be learned (Ekman & Oster, 1979). This consideration in the context of a Bayesian model gives a rational explanation for the results of Dunsmoor et al. (2009) and Dunsmoor et al. (2011).

Given the prior and the likelihood, the model is fully specified using Bayes' rule. Because

[3] To review different sampling assumptions, see subsection 1.2.2 or Navarro et al. (2012).

there is no constraints on the number of observations, the observed stimuli and outcomes are summarized in the vectors $\boldsymbol{s}$ and $\boldsymbol{r}$. The posterior distribution over the parameters is given by

$$p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r}) = \frac{p(\boldsymbol{s},\boldsymbol{r}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{s},\boldsymbol{r})} \tag{2.14}$$

$$\propto p(\boldsymbol{s},\boldsymbol{r}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.15}$$

From belief state to generalization. The actual generalization given a belief state $p(\boldsymbol{\theta})$ or $p(\boldsymbol{\theta}|r,s)$ depends on the prior (Equation 1.6) or posterior predictive (Equation 1.7) distribution. In the context of this model, the agent can infer the probability of an observation of a new stimulus and outcome $\tilde{s}$ and $\tilde{r}$ by integrating the conditional probability $p(\tilde{s},\tilde{r}|\boldsymbol{\theta})$ over the prior or posterior distribution:

$$p(\tilde{r},\tilde{s}) = \int p(\tilde{r},\tilde{s}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{2.16}$$

$$p(\tilde{r},\tilde{s}|\boldsymbol{s},\boldsymbol{r}) = \int p(\tilde{r},\tilde{s}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r})d\boldsymbol{\theta}. \tag{2.17}$$

Typically, instead of judging the probability of a stimulus *and* and an outcome, agents need to estimate the conditional probability that an outcome follows a stimulus. This is also the case in typical experiments, in which ratings or psychophysiological measures are collected as a reaction to a stimulus. Assuming that an agent already experienced stimuli $\boldsymbol{s}$ and outcomes $\boldsymbol{r}$, this probability is given by $p(\tilde{r}|\tilde{s},\boldsymbol{s},\boldsymbol{r})$. Following Equation 2.16 and Equation 2.17, it is easy to see[4] that the conditional posterior predictive probability of an outcome can be computed by integrating $p(\tilde{r}|\tilde{s},\boldsymbol{\theta})$ over the posterior[5]:

$$p(\tilde{r}|\tilde{s},\boldsymbol{s},\boldsymbol{r}) = \int p(\tilde{r}|\tilde{s},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r})d\boldsymbol{\theta}. \tag{2.18}$$

Generating predictions. Because the posterior of this model (Equation 2.14) is intractable, there is no closed form solution. Instead, to generate predictions, the model is approximated by sampling from the posterior using Markov chain Monte Carlo (MCMC). Using this approach, we can generate samples $S(\boldsymbol{\theta})$ from the posterior distribution:

$$S(\boldsymbol{\theta}) \sim p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r}). \tag{2.19}$$

Importantly, since the samples are distributed according to the posterior, we can compute approximate posterior expectations by averaging over the samples:

[4] The only situation in which these wordings do not suck is when you're the author. The proof for this is left as an exercise to the reader.

[5] Fun fact: This principle is called *The Law of the Unconscious Statistician* (Blitzstein & Hwang, 2015).

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{\theta})] &= \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r})d\boldsymbol{\theta} \\ &\approx \frac{1}{N}\sum_{i=1}^{N} f(S_i(\boldsymbol{\theta})).\end{aligned} \tag{2.20}$$

Having implemented the model in a probabilistic programming language like `Stan` (Carpenter et al., 2017), we first need to specify the prior distributions on the parameters in order to generate samples. Because those priors are supposed to reflect assumptions about the belief state of the agent, this is extremely context-dependent and I can give no general way to do this. Specifying the priors is already enough to generate prior predictions, i.e. to predict a generalization gradient *before* conditioning. However, to generate posterior predictions, we need to specify the observations $\boldsymbol{s}$ and $\boldsymbol{r}$. Since the model only considers the distinction between outcome and no outcome, $\boldsymbol{r}$ is a binary vector. The observations $\boldsymbol{s}$ can be given as the coordinates of the stimuli in psychological space. Alternatively, one could give a vector of integers that specify the index and another object that contains the positions. Predictions for the reaction to a stimulus $\tilde{s}$ given a belief state $p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r})$ can be generated from the posterior samples $S(\boldsymbol{\theta})$ following Equation 2.20:

$$p(\tilde{r}|\tilde{s},\boldsymbol{s},\boldsymbol{r}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\tilde{r}|\tilde{s},S_i(\boldsymbol{\theta})). \tag{2.21}$$

Model fitting. If we wanted to explain generalization data from subjects using the model, i.e. fit the model to data, the property that we would try to characterize are the priors $p(\boldsymbol{\theta})$. In particular, those priors have parameters that define the prior belief state. Those are the so-called hyperpriors $\boldsymbol{\tau}$. A consideration of hyperpriors introduces new dependencies to the posterior distribution (Equation 2.14)

$$p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r},\boldsymbol{\tau}) \propto p(\boldsymbol{s},\boldsymbol{r}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\tau}) \tag{2.22}$$

and the conditional posterior predictive distribution (Equation 2.18)

$$p(\tilde{r}|\tilde{s},\boldsymbol{s},\boldsymbol{r},\boldsymbol{\tau}) = \int p(\tilde{r}|\tilde{s},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{s},\boldsymbol{r},\boldsymbol{\tau})d\boldsymbol{\theta}, \tag{2.23}$$

since different hyperpriors lead to different posteriors.

The data that we would use to fit values of $\boldsymbol{\tau}$ are the generalization gradients, i.e. either expectation ratings or psychophysiological signals, that I will call $\boldsymbol{z}$. As a consequence, we would need to define the likelihood $p(\boldsymbol{z}|\boldsymbol{s},\boldsymbol{r},\boldsymbol{\tau})$ of data given the hyperpriors and the experienced stimuli and outcomes. Unfortunately, subjects do not respond perfectly and the measurements likely do not live on the same scale as the posterior predictive probabilities[6]. It follows that we need some mapping from $p(\tilde{r}_i|\tilde{s}_i,\boldsymbol{s},\boldsymbol{r},\boldsymbol{\tau})$ to z_i. A simple solution to this would be a linear mapping assuming Gaussian errors

[6] Probabilities are naturally constrained to the range $\{0,1\}$. Psychophysiological measurements are not. Ratings could be translated to this range, but depending on the resolution of the rating scale, this is only a rough approximation and still does not solve the problem of noisy responses.

$$\begin{aligned} z_i &= \beta_0 + \beta_1 p(\tilde{r}_i|\tilde{s}_i, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\tau}) + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma), \end{aligned} \tag{2.24}$$

which would yield the likelihood

$$p(z_i|\tilde{s}_i, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\tau}) \sim \mathcal{N}(z_i|\beta_0 + \beta_1 p(\tilde{r}_i|\tilde{s}_i, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\tau}), \sigma). \tag{2.25}$$

This likelihood could be used to fit the model either using maximum likelihood estimation, Bayesian inference or a grid search. The practical problem with that is that this likelihood depends on the posterior $p(\boldsymbol{\theta}|\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\tau})$, which is intractable and itself depends on $\boldsymbol{\tau}$. Therefore, we would need to resample from the posterior for every different value of $\boldsymbol{\tau}$ that we are attempting in order to find the best-fitting value. Considering the cost of sampling from the posterior once and the number of iterations that are typically needed to generate a stable estimation of $\boldsymbol{\tau}$, this would be prohibitively expensive.

Crucially, this does not make the model useless. In fact, it is a common approach in research on inductive reasoning (J. C. Lee, Lovibond, Hayes, & Navarro, 2019; Soto et al., 2014) and outside (e.g. Behrens et al., 2007) to use intractable models to make predictions and derive specific hypotheses from them. Fitting simpler models, performing statistical tests or visually comparing empirical data to predictions can then be used to test the validity of the model.

3 An improved method for the estimation of perceptual spaces

3.1 Measuring psychological spaces

As outlined in chapter 1, there is an intricate link between psychological spaces and cognitive models of generalization, as virtually all theoretical approaches in inductive reasoning since Shepard (1987) make use of the concept of consequential regions in psychological space. The model I proposed in chapter 2 is no exception. The need to account for individual and group-level psychological spaces implies that a means to accurately measure these spaces is of crucial importance. Among the proposed solutions to this problem, two seem especially noteworthy: Multidimensional scaling (MDS) and maximum likelihood difference scaling (MLDS). In the following I will give a short description of both along with their corresponding problems. Due to those problems, I was not fully content with applying these methods to my empirical studies. For this reason, I developed a novel method based on MLDS. In this chapter I will describe this method and how it circumvents the aforementioned problems.

3.1.1 Multidimensional Scaling

MDS describes a set of mathematical techniques that are used to perform dimensionality reduction on high-dimensional data (Kruskal & Wish, 1978). The general idea of all these methods is to find an embedding of high-dimensional data points $\boldsymbol{p}$ in a low-dimensional space where the dimensionality has to be defined by the researchers a priori. All methods take a dissimilarity matrix $\boldsymbol{D}$ as an input, in which the element $\boldsymbol{D}_{i,j}$ is defined as some distance measure (e.g. euclidean distance) between data points $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. The optimal embedding is defined as the one that minimizes the deviation between the original dissimilarities and some function of the distances in the low-dimensional space.

Non-metric multidimensional scaling (NMDS), as implemented in techniques proposed by Shepard (1962) and Kruskal (1964) is especially relevant in the field of generalization. Shepard (1987) proposed it as a method to derive a psychological space from generalization data and therefore inverted the usual approach to look at generalization gradients in some predefined physical space. This approach is constrained to a special kind of data, namely the probability of showing a learned behavior towards a novel stimulus. Because the data I collected and analyzed does not adhere to this format (see chapter 4), this approach was not applicable to the presented studies. Aside from that, it suffers from a few drawbacks like the dependence on starting values in optimization-based methods (Borg & Mair, 2017) and the lack of an accepted method of arriving at a group level solution (M. D. Lee & Pope, 2003). In addition, explicit dissimilarity ratings, as would be the input to MDS, suffer from potential problems like individual judgment strategies (Schönemann & Lazarte, 1987) that are independent of the *true* psychological space. This effect could be especially bad for multi-dimensional stimuli where the dimensions differ in salience and valence.

3.1.2 Maximum Likelihood Difference Scaling

Originally, maximum likelihood difference scaling (MLDS) is a method to fit psychometric functions for sub-threshold perceptual differences along a single dimension (Maloney & Yang, 2003). As such it can't be used to estimate higher-dimensional psychological spaces. However, MLDS is an interesting starting point as the kind of data being used is less likely to be influenced by rating tendencies since no explicit dissimilarity ratings are required. Instead, MLDS uses indirect measures of stimulus dissimilarities as input. This data is typically collected using two alternative forced choice (2-AFC) tasks in which subjects are presented with two pairs of stimuli and have to rate the relative dissimilarity of the two pairs. Assuming Gaussian perceptual noise and imposing some constraints to identify the model, this problem can be formalized as a probabilistic model in which the likelihood of responses depends on a one-dimensional embedding of stimuli[1] $\boldsymbol{\psi}$ and the variance of perceptual noise σ, which are the free parameters of the model. The optimal solution is found via maximum likelihood estimation, i.e. it is defined as the parameter values for which the likelihood of the responses is maximal. Computationally, this is achieved by minimizing the negative log likelihood via a gradient-based optimization approach.

Mathematical formulation. The likelihood function for the original MLDS is $p(\boldsymbol{r}|\boldsymbol{\psi}, \sigma, \boldsymbol{s})$, where $\boldsymbol{r}$ is the set of responses of a subject and $\boldsymbol{s}$ is the sequence of quadruplets. This can be computed in a few simple steps. Initially, the dissimilarity $l_{i,j}$ between the two stimuli i and j in a pair can directly be computed from $\boldsymbol{\psi}$ as the absolute value of the difference between their positions in space:

$$l_{i,j} = |\psi_i - \psi_j|. \tag{3.1}$$

From the dissimilarities in the first pair, consisting of stimuli i and j, and the second pair, consisting of k and l, we can compute the relative dissimilarity $D_{i,j,k,l}$, i.e. the difference between the dissimilarities:

$$D_{i,j;k,l} = l_{i,j} - l_{k,l} \tag{3.2}$$

Note that this is *not* the absolute value since the sign of D indicates which pair is more similar. Assuming Gaussian perceptual noise, the likelihood of choosing the second pair to be more similar for the response of a single quadruplet q, corresponds to the standard

[1] I.e. positions of stimuli in one-dimensional space.

normal cumulative density function (CDF)[2] at D, after normalizing it using the standard deviation of perceptual noise σ:

$$p(r_q = 1|\boldsymbol{\psi}, \sigma, s_q) = \Phi\left(\frac{D_q}{\sigma}\right). \tag{3.3}$$

Intuitively this means that subjects are more likely to choose the more similar pair and the probability of choosing the more similar pair increases with the magnitude of the difference and decreases with the variance of perceptual noise. The likelihood of a set of responses $\boldsymbol{r}$ is the product of the individual likelihoods:

$$p(\boldsymbol{r}) = \prod_{q \in \boldsymbol{s}} r_q. \tag{3.4}$$

Note that this model is not identified so far as there are $N+1$ free parameters for N stimuli. For a simple example of model identification, consider the following equation:

$$x + y = 0.$$

This equation has two free parameters, x and y, i.e. there are N equations and $N+1$ parameters. As a consequence, there are infinitely many solutions to this equation. In the context of MLDS, this means that the solution is invariant to translation, scaling and reflection. Intuitively, there are an infinite number of solutions that have the same likelihood and an algorithm to find an optimal solution does not have a principled way to decide between them. Since the purpose is to fit a monotonic psychometric function, Maloney and Yang (2003) solved this problem by setting ψ_1 and ψ_N, the *endpoints* of the stimulus dimension to 0 and 1 respectively, which yields an identified model with $N-1$ parameters. In theory, ψ could be fixed for any two stimuli to identify the model.

Generalization to higher-dimensional spaces. While a generalization to higher dimensions is not necessary for the original application of MLDS, the likelihood function (Equation 3.3) can be used to estimate higher-dimensional spaces with a few modifications. This was done by Onat and Büchel (2015). First, the stimulus positions $\boldsymbol{\psi}$ are vectors, not scalars. As a consequence, the absolute value in Equation 3.1 gets replaced with the euclidean distance between the vectors

$$l_{i,j} = ||\boldsymbol{\psi_i} - \boldsymbol{\psi_j}|| \tag{3.5}$$

which does not change the one-dimensional case, but generalizes to an arbitrary number of dimensions. Second, an additional constraint is needed to identify the model as an additional invariance to rotation is present in e.g. the two-dimensional case. All invariances are displayed in Figure 3.1, using the stimulus space from Onat and Büchel (2015) as an example. To identify these models, one additional stimulus position per dimension needs to

[2] Indicated by $\Phi(x)$.

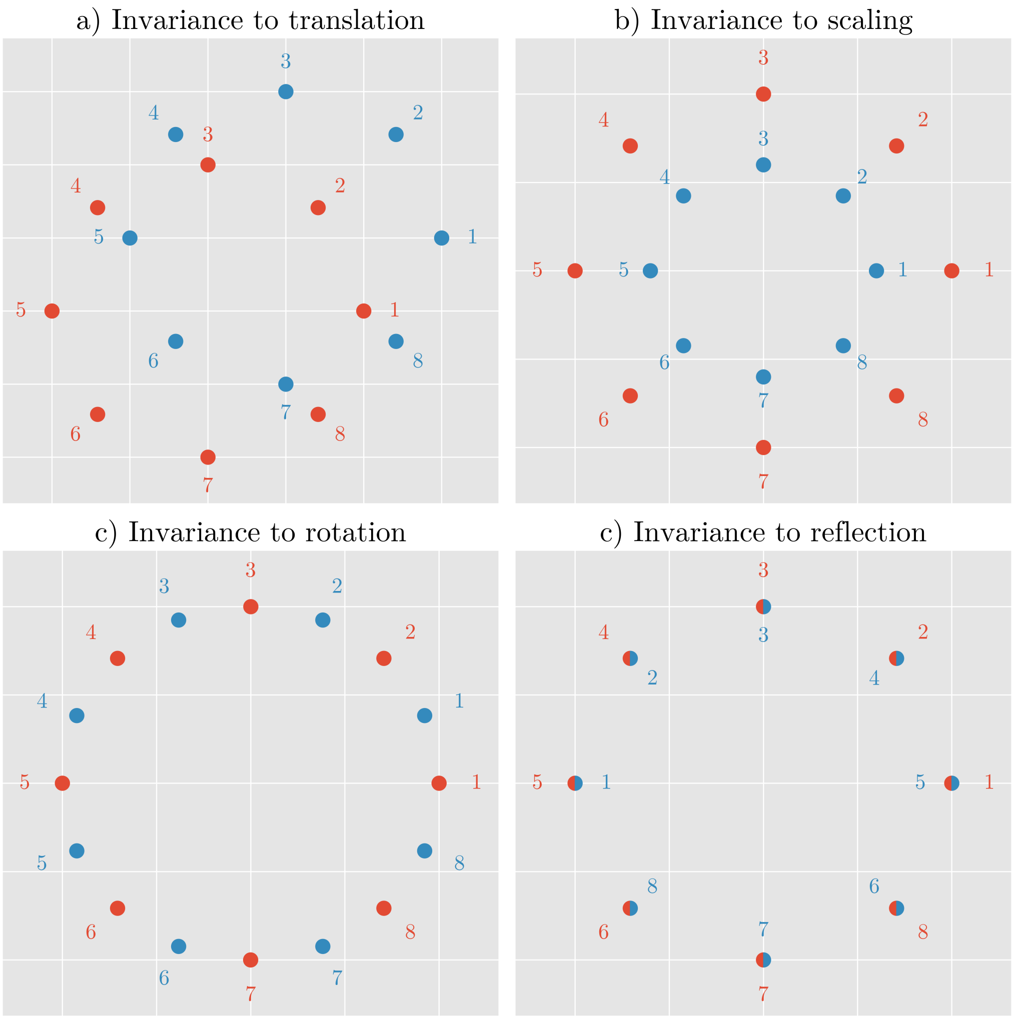


Figure 3.1: Invariances of a two-dimensional MLDS model. This figure uses the stimulus space from Onat and Büchel (2015) as an example. All depicted solutions yield the same likelihood, assuming the standard deviation of perceptual noise is adjusted for the scaled version.

be constrained. E.g. in two dimensions, this can be achieved by constraining one stimulus to lie on either side of the line that is defined by the two fixed stimulus dimensions. The necessary constraints are depicted in Figure 3.2. Note that Onat and Büchel (2015) only fixed one stimulus position, which does not constitute an identified model. Using maximum likelihood estimation, this approach can still yield a reasonable estimate, as the optimization algorithm will likely find the closest of infinite equivalent local minima. However, a Bayesian approach that relies on sampling from a full probability distribution requires a fully identified model.

3.2 Hierarchical Mean Posterior Difference Scaling

Even though generalized MLDS uses indirect dissimilarity ratings and thus solves one main problem of MDS, some concerns remain: a) There is no way to quantify the uncertainty in the estimate of positions, b) to arrive at a group solution, one needs to average estimates

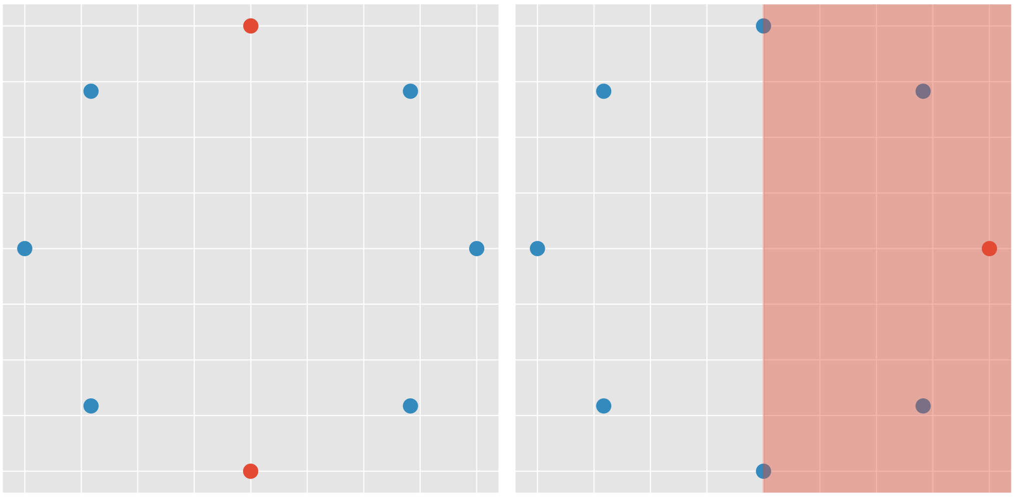


Figure 3.2: Constraints to identify a two-dimensional MLDS model. In order to account for all invariances, three stimulus conditions need to be constrained. One option, that is shown here is to fix two stimulus positions (left side) and constrain one additional position to be on a specific side of the line defined by the two fixed positions (right side).

which gives the same weight to subjects that rated randomly as to those that rated conscientiously, c) there is no way to incorporate the assumption that psychological spaces are correlated between subjects and d) prior knowledge about the psychological spaces. All of these problems can be solved by formulating MLDS as a hierarchical Bayesian model (Gelman, 2014) while keeping the MLDS likelihood function (Equation 3.3). I call this approach hierarchical mean posterior difference scaling (hMPDS).

Hierarchical parameter structure. In order to find a hierarchical Bayesian model specification for MLDS, I assumed individual stimulus positions $\psi_{subject}$ to be normally distributed around a group level stimulus position ψ_{group} with some unknown variance σ_{ψ}^2. For the x-position of stimulus i, this relationship is given by

$$\psi_{subjects,x,i} \sim \mathcal{N}(\psi_{group,x,i}, \sigma_{\psi}^2). \tag{3.6}$$

Individual perceptual noise variances were estimated on the log scale to avoid sampling problems that arise from the positivity constraint[3]. Individual log perceptual noise variances $\sigma_{subject}$ were assumed to be normally distributed around a group level log variance with an unknown between-subject variance σ_{σ}^2 and exponentiated to arrive back at the linear scale:

$$\log(\sigma_{subjects}) \sim \mathcal{N}(\log(\sigma_{group}), \sigma_{\sigma}^2). \tag{3.7a}$$

$$\sigma_{subject} = e^{\log(\sigma_{subject})} \tag{3.7b}$$

[3] Variance as the expected squared distance to the mean can only be non-negative.

Note that while a single group mean is estimated for each stimulus and dimension separately, the between-subject variance is kept constant since estimating one variance parameter per stimulus and dimension leads to identifiability issues. The interpretation for a single variance parameter is the general deviation of individual psychological spaces from the group mean. Intuitively, Equation 3.6 says that psychological spaces are likely to be similar between subjects, i.e. the same stimuli tend to be perceived as more similar. In contrast to the optimization based generalized MLDS, the aforementioned constraints (Figure 3.2) are essential as only a fully indentified Bayesian models can be estimated using MCMC[4]. However, in practice it turned out to be sufficient to apply these constraints to the group level and leave subject level parameters unconstrained, although this is likely to be depending on the amount and precision of the available data.

Weakly informative priors. Since the group level parameters act as prior distributions on the single subject parameters, we only need to specify priors on group level parameters[5]. While the model might be identified with uninformative priors (depending on the amount of data), the Bayesian approach enables us to incorporate prior knowledge about the stimulus space into the analysis to allow for a more accurate posterior estimate. I outlined the reasoning for the priors I used in my research in subsection 4.2.3, however there are a number of valid approaches to arrive at (weakly) informative priors. These can e.g. be informed by the stimulus creation process, pixel- or neural model-based estimates of similarity or results from previous studies.

Model fitting and inference. Given Equations 3.3– 3.7, the posterior of the model is fully specified. As for most somewhat complex Bayesian models, there is no closed-form solution to the equations, but it can be estimated using MCMC, Variational Bayes (VB) or a maximum a posteriori (MAP) estimate (Gelman, 2014). The specific approach I used to fit Bayesian models is described in subsection 4.1.2. To arrive at a group and individual solutions, I used the posterior mean of the group level and the subject level parameters, respectively.

[4] A model that is not fully identified has a multimodal posterior distribution with an infinite number of equivalent modes, which makes it impossible to sample from in practice.

[5] So-called hyperpriors.

4 Empirical studies

The empirical part of this thesis consists of three separate studies. These were conducted to test the predictions of the Bayesian model[1], namely if the ratings of subjects followed a model that includes dimensionality reduction and assumptions about the prior knowledge and thereby integrates cognitive Bayesian models with representation learning. The first study was purely behavioral and aimed at establishing the behavioral effects. The second study repeated the design of the first study in the fMRI scanner to replicate the behavioral findings of the first study and gain insight about their neural underpinnings. In particular, I was interested if the dimensionality aspect in stimulus generalization followed predictions from representation learning (see section 1.3) and in the neural correlates of prior knowledge. In the third study I modified the design from aversive to appetitive conditioning in order to broaden the scope of the established mechanisms from fear generalization to a more general account of generalization.

4.1 General methods

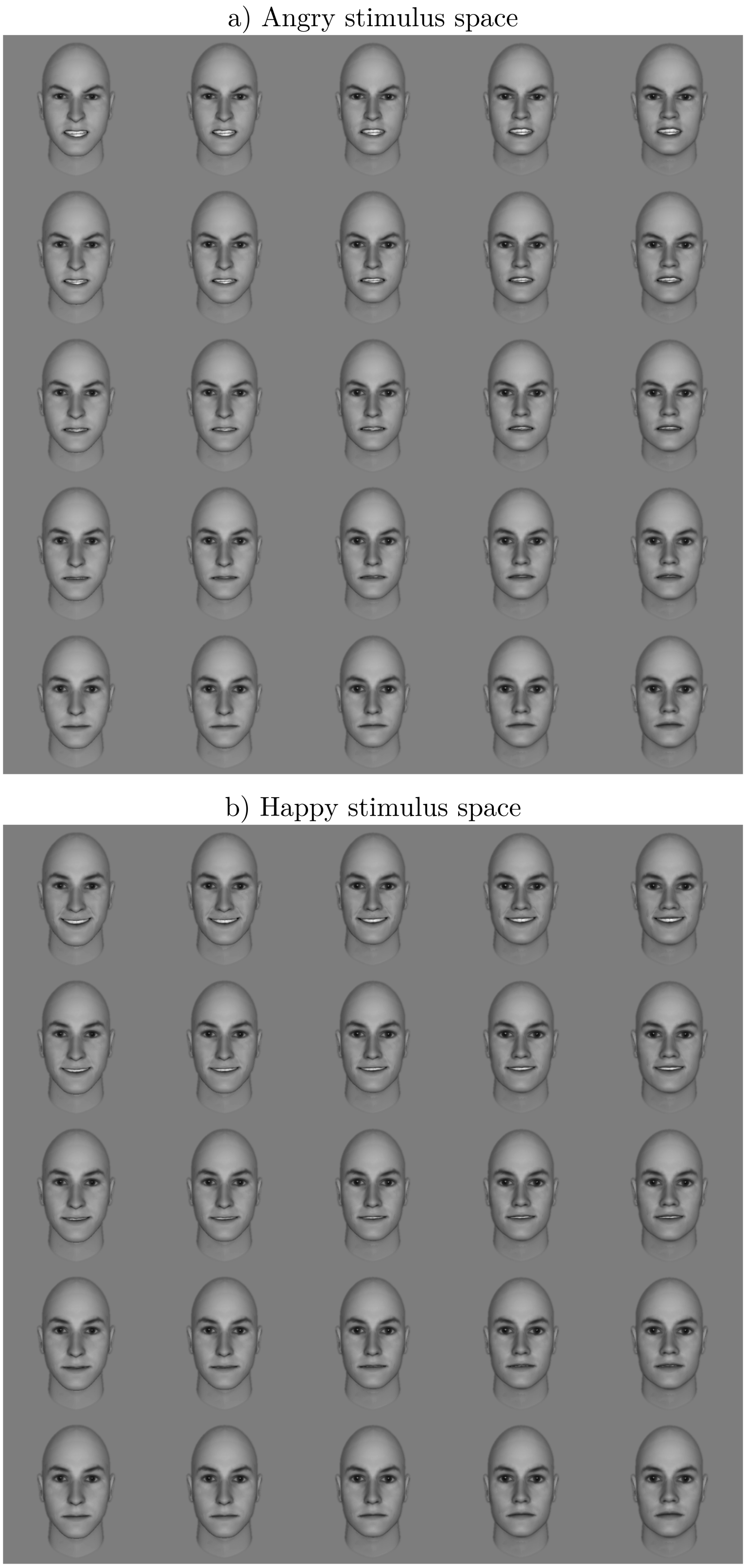
Even though the studies differed in the setup and designs, some aspects stayed constant throughout all of them. The relevant methods for these aspects are described in this section.

4.1.1 Stimulus Space

The stimulus space for all studies consisted of two-dimensional grids of computer generated faces. These faces differed on the two dimensions facial identity and emotion. For each dimension five warping steps were created by morphing two baseline identities into each other with differing relative contributions and by using different amounts of emotional facial expression. Because I compared two different emotional expressions, angry and happy, this approach resulted in two separate $5x5$ stimulus spaces (Figure 4.1).

Stimulus creation. Stimuli were created using the software FaceGen Modeller (Singular Inversion, VA). In a first step I created two distinct facial identities. Next, these two identities were morphed in 5 steps. The endpoints of this dimension were defined by the unaltered original identities. The three combined faces were created by combining the two identities with differing contributions of both faces (25%/75%, 50%/50%, 75%/25% of the first vs. second identity). In the last step, I added 5 different amounts of emotional

[1] For a detailed description, see chapter 2.



a) Angry stimulus space

b) Happy stimulus space

Figure 4.1: Stimulus spaces for all three studies. Stimulus spaces consisted of 5x5 grids of computer generated faces. Faces differed on two dimensions in five steps each: Facial identity and either **a)** angry or **b)** happy emotional expressions.

expression (0-100% in steps of 25%), which resulted in the $5x5$ grids that are displayed in Figure 4.1.

4.1.2 Bayesian model fitting

All Bayesian models were programmed in the `Stan` probabilistic programming language (Carpenter et al., 2017) as hierarchical models. `Stan` uses a MCMC approach called Hamiltonian Monte Carlo (HMC, Betancourt, 2018) to sample from the posterior distribution of the model parameters. Because sampling algorithms like HMC tend to suffer from the curvature that is typical for posteriors of hierarchical models (Betancourt & Girolami, 2013), which results in convergence issues, all hierarchical parameters can be specified in a non-centered manner (Betancourt, 2018). This approach is mathematically equivalent to the centered one, but computationally more convenient as it avoids local correlations in the posterior. All models were run using 4 chains with 2000 iterations each. Because HMC needs to find the so-called *typical set* first before the sampler has found the stationary distribution and sample expectations match the expectations of the distribution, the first 1000 iterations per chain were treated as warm-up and discarded. As a consequence, posterior expectations were computed using 4000 samples in total, unless more samples were needed for convergence diagnostics.

Convergence criteria. The convergence of the MCMC chains was assessed using the $\hat{R}$ statistic (Gelman, 2014; Vehtari et al., 2021), a quantity that measures whether different chains and different parts of the same chains have sampled from the same distribution. Successful convergence can be assumed for values of $\hat{R}$ that are close to one. Following recommendations of Gelman (2014), I used the criterion $\hat{R} < 1.1$. In cases where this criterion could not be satisfied with the aforementioned 4000 iterations, I increased the sample size by 1000 iterations per chain until convergence could be assumed. Additional criteria that were used to assess potential problems with the sampling procedure were the effective sample size (a measure of strong autocorrelation that can indicate local problems) and the number of divergent transitions. Divergent transitions are a specific problem to gradient-based MCMC samplers and indicate that the sampler is unable to sample from a specific part of the posterior, typically indicated by Neal's funnel (Betancourt, 2018), because the step size is too small to keep track of the gradient-based trajectory of the sampler particle[2]. The risk of divergent transitions is minimized by using a non-centered parameterization, but with complicated models and relatively small amounts of data they can still occur. In cases where I encountered divergent transitions, I decreased the step size for the sampler until no divergences were observed[3]. In small number of cases this

[2] HMC follows a trajectory in the posterior space by simulating a Hamiltonian system. The step size is the resolution by which this trajectory is approximated and the correct step size depends on the curvature of the posterior distribution. If the curvature is very different between different locations, the step size is likely too large to sample from regions with high curvature. For a more detailed explanation, refer to the very good conceptual introduction by Betancourt (2018).

[3] This approach is detrimental to efficiency because it needs more computations of the posterior, but it avoids divergent transitions.

was insufficient to avoid divergent transitions. For these cases the validity of the results is questionable and I have indicated those cases in the respective sections.

Posterior expectations. To use posterior samples for further use in the analysis in other models or to compute parametric modulators for model-based fMRI analysis, I computed posterior expectations for parameters and implied quantities (e.g. predicted shock expectation ratings). Posterior expectations are computed from samples by averaging over the respective values:

$$\mathbb{E}[\theta] = \frac{1}{N} \sum_{n=1}^{N} S_n(\theta). \tag{4.1}$$

4.1.3 Model comparison

In some cases I fit multiple models to the same data. To compare these models with respect to how well they explain the data while accounting for model complexity, I used leave-one-out crossvalidation (LOO-CV). This approach fits a model to all but one data points and evaluates how well the model can predict the left-out data point. Typically, this means one has to fit the model once per data point. In the case of MCMC sampling, this is often prohibitively expensive. To circumvent this problem, one can use importance sampling to approximate posteriors that do not depend on a specific data point one at a time. This approach is called Pareto smoothed importance sampling leave-one-out crossvalidation (PSIS-LOO) and was proposed by Vehtari et al. (2017) and refined by Vehtari et al. (2019). The PSIS-LOO approximation depends on the expected log predictive density (ELPD), i.e. the average log posterior predictive density of left out data points. While the approach by Vehtari et al. (2019) is very efficient because it only requires a single sampling run if everything goes smoothly, it is not robust with respect to very informative data points. These can be filtered out by observing the shape parameter of the Pareto distribution that is used to approximate the importance weights for PSIS-LOO, but unfortunately the model needs to be rerun manually for each very informative data point.

Silva and Zanella (2022) proposed an alternative method that requires an additional sampling run in which one samples from an auxiliary distribution. This is more computationally expensive than the method by Vehtari et al. (2019) is in the optimal case. But because it is robust with respect to informative data points, it is more efficient in practice as it *always* requires just two sampling runs in total. I used this implementation in all model comparisons.

Note that in contrast to classical information criteria like Akaike information criterion (AIC) or Bayesian information criterion (BIC), LOO-CV does not require explicit penalization for model complexity[4]. This is because LOO-CV is based on out-of-sample predictions, which implicitly penalizes overfitting. In addition, adding parameters to a Bayesian model

[4] Usually model complexity is approximated by the number of parameters of a model. This is problematic for Bayesian models because it does not account for the impact of narrow vs. wide priors. A model with narrow priors is more rigid despite the same number of parameters.

distributes the prior (and thus posterior) probability mass over a higher-dimensional parameter space, which reduces the ELPD and naturally penalizes model complexity.

4.2 Behavioral mechanisms of fear generalization

In the first study I intended to establish whether the proposed model can explain generalization behavior, i.e. whether the collected data was in line with the predicted behavioral effects (see subsection 4.2.2). Importantly, the model makes specific predictions for the impact of prior knowledge, the relative relevance of different stimulus dimensions and about the dynamics of these effects. An appropriate task design needs to be set up in a way that allows to probe all of these predictions.

4.2.1 Experimental design

The typical design in stimulus generalization experiments consists of three phases. A *baseline* phase, in which all stimuli, i.e. the CS+, the CS- and the generalization stimuli, are shown to collect baseline data. Following that, a *conditioning* phase, in which only two stimuli are shown to the subject and one of the stimuli (the CS+) is reinforced probabilistically while the other stimulus (the CS-) is never followed by an outcome. Finally, a *generalization* phase, in which the subject is presented with all stimuli again and generalization data is recorded.

This approach implicitly assumes that learning is constrained to the conditioning phase and that one can record a stable response in the generalization phase. For this purpose, often the CS+ is reinforced a few times to prevent extinction learning. However, one aspect that cannot be investigated with this design are the dynamics of learning. This is especially problematic because Bayesian models make specific predictions about those dynamics and an investigation can help to differentiate between data that is consistent with a perceptual model of generalization and data that is not. In addition, given the importance of perception in generalization, it is important to account for individual psychological spaces of subjects and possible changes thereof.

Top level structure of the design. To account for psychological spaces, the study started with a perceptual task[5]. The data from this task was used to estimate the subjects' psychological space using hMPDS[6]. Following this, subjects went through the associative generalization phase. To investigate the dynamics of learning, I did not follow the typical generalization design, but adopted another approach by Onat (2018). This approach omits the distinction between the three phases and replaces them with so-called *microblocks*, thereby combining conditioning and generalization phases. This allows to investigate changes to the belief state during learning by intermittently measuring the subject's response or concurrently collecting psychophysiological recordings. I used a painful electric

[5] Described in detail below.

[6] See chapter 3.

shock as UCS, which is following the typical procedure in fear generalization (Webler et al., 2021). To reduce between-subject variance, I employed a within-subject design, i.e. the same subjects came on two separate days to take part in the angry and happy condition. This top level structure is depicted in Figure 4.2.

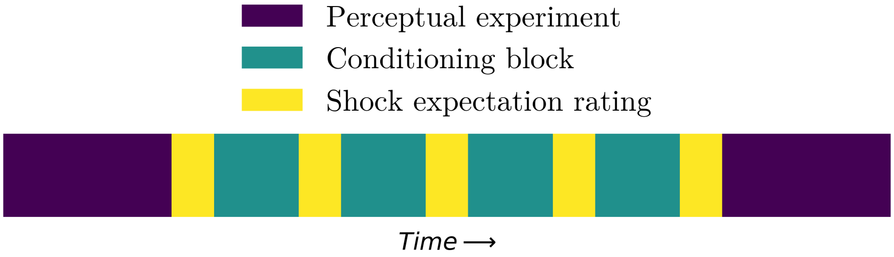


Figure 4.2: Top level structure of the experimental design. Subjects went through a perceptual task to estimate their psychological space. This was followed by the actual generalization task, which consists of conditioning in microblocks (Onat, 2018) with shock expectation ratings in between. Lastly, subjects repeated the perceptual task to control for changes in perceptual spaces.

Perceptual task. Because the concept of psychological spaces is crucial to some of the considered models of generalization, I made sure that the actual psychological spaces of subjects were accounted for and included in the analysis. Due to conceptual problems with explicit dissimilarity ratings, I decided to avoid those and instead use more indirect measures for the estimation of psychological spaces. For this purpose, subjects in all experiments went through a so called quadruplet task. In these tasks, subjects were prompted with a quadruplet of faces in each trial. A quadruplet consists of four faces where the upper and lower two faces form a pair each (Figure 4.3).

Because the full stimulus space consists of $5*5 = 25$ faces, the number of possible quadruplets is prohibitively large to run a sufficiently large proportion of them within reasonable time constraints. Since this makes it difficult to accurately estimate the psychological space, I used a reduced stimulus space by only using a $3x3$ subset of faces (Figure 4.4).

The task was programmed in `MATLAB` using `Psychtoolbox3` (Kleiner et al., 2007). In each trial, subjects had to choose the pair in which they thought the faces were more similar. Quadruplets were shown for 6 seconds. Subjects were instructed to respond intuitively within those 6 seconds. However, the response was self-paced and subjects could still make their choice after the faces had vanished from the screen. This choice was made with the `y` and `m` key on a keyboard with a German layout. The task consisted of a total of 388 trials. This number was based on another study which used generalized MLDS with similar faces (Onat & Büchel, 2015). The number of stimuli differed, but I used the same percentage of possible quadruplets. To determine an informative sequence of quadruplets, I ran 2000 simulations with a hypothetical observer and random sequences and chose the sequence for which the deviation between the assumed stimulus positions and the fitted positions was minimal. This sequence was used for all subjects. To keep attention high and ensure good data quality, the 388 trials were binned into four blocks of 97 trials each. Between blocks

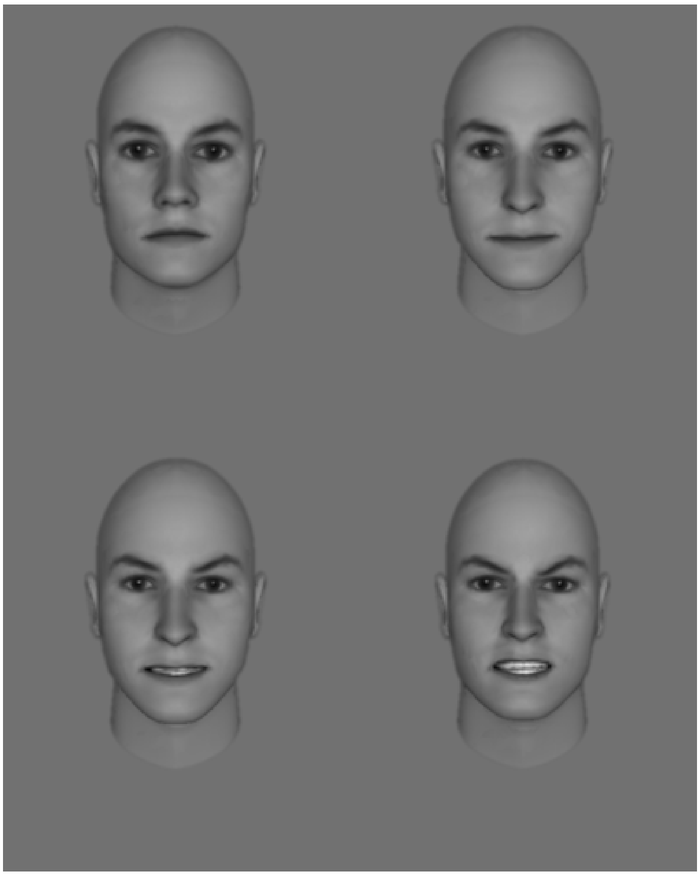


Figure 4.3: An exemplary quadruplet trial. The upper two and lower two faces form a pair each. Subjects need to decide in which pair the faces are more similar to each other.

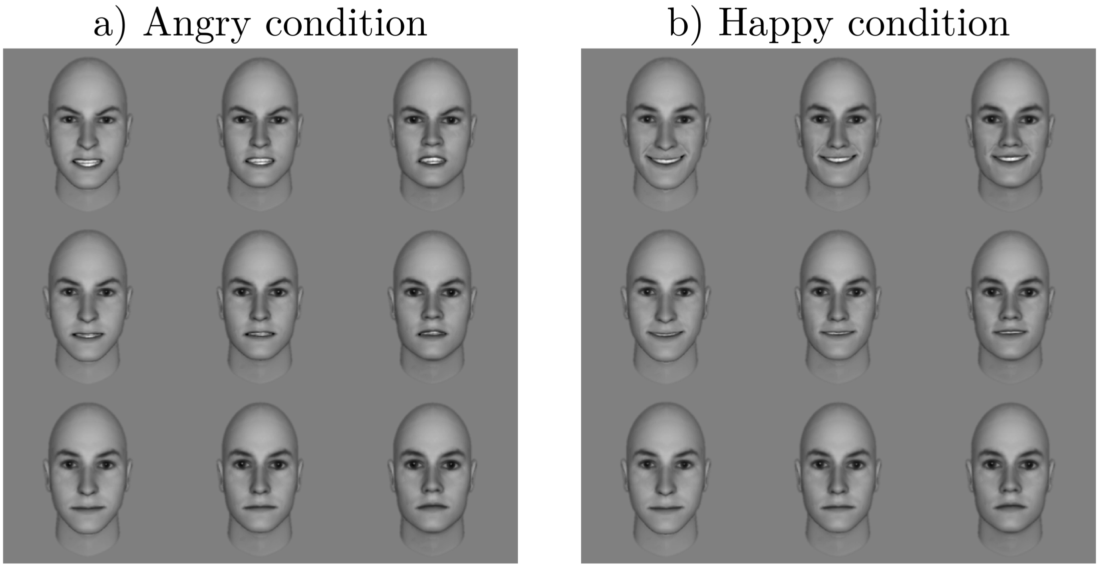


Figure 4.4: Reduced stimulus space for quadruplet task. Reduced spaces in the **a)** angry and **b)** happy condition consisted of the outer $3x3$ grid, including the center stimulus.

subjects were instructed to relax for a few moments before proceeding to the next block. Subjects repeated this task before and after the conditioning procedure (see Figure 4.2).

The point of this approach was to control for potential changes in psychological spaces that were induced by the learning experience of the conditioning paradigm.

Pain calibration. In the first study, I used an electric shock as the UCS. The shock was delivered via a direct current stimulator (Digitimer Constant Current Stimulator, Digitimer) and an electrode that was attached to the back of the left hand. Because pain perception varies substantially between individuals given the same stimulus (Amir et al., 2022), I calibrated the amplitude for each subject individually to ensure that I used a shock amplitude that was painful, but bearable. To do so, I used a QUEST procedure (Watson & Pelli, 1983). In this procedure, the subject receives a series of 12 shocks with amplitudes that are suggested by the QUEST algorithm and has to rate whether the shock was painful or not. This approach generates a pain threshold, which is defined as the amplitude that a subjects perceives as painful in 50% of the repetitions. I set the amplitude for the UCS to 1.5 times that threshold. Before starting the experiment, subjects confirmed that the shock amplitude was bearable and were informed that they could opt out of the experiment at any time.

Main experiment. Subjects were informed about the procedure using written information on the screen. Following that, they were presented with nine random faces and one oddball trial. This was to familiarize them with the experiment and practice reacting to the oddball face with the space bar. After that came the first rating block. Importantly, this block was before any conditioning and allowed me to probe the prior without being contaminated by the learning experience. Following the rating, subjects were presented with the faces and occasional reinforcement in pseudo-randomized microblocks (see below, Onat, 2018). After five microblocks, subjects gave another shock expectation rating for every face. This was repeated for a total of 20 microblocks and five rating blocks.

Outcome measures. Psychophysiological recordings like electrodermal activity and pupil dilation are notoriously noisy and have a relatively slow temporal component. For this reason, model-based approaches to deconvolve the signal like `ledalab` (Benedek & Kaernbach, 2010) or `PsPM` (Bach et al., 2013; Korn et al., 2017) are needed, especially in rapid event related designs[7]. Unfortunately, those model-based approaches need a lot of data to work reliably. Since I was interested in the dynamics of learning, i.e. could not average over trials and used a lot of different stimuli, I resorted to shock expectation ratings instead.

For the shock expectation ratings, subjects were presented with every face in random order for 1.5 seconds and had to give a rating on a 10-point Likert scale with respect to the question „How likely will this face be followed by a shock?“. The anchors of the scale were „Not at all likely“ and „Very likely“.

[7] Rapid event related designs are designs in which trials are presented in quick succession, i.e. the time between two trials is shorter than the duration of psychological responses.

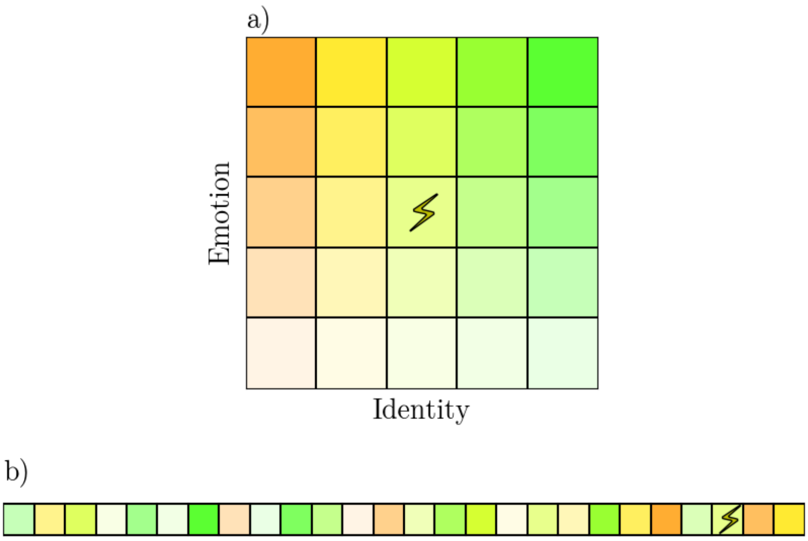


Figure 4.5: Schematic depiction of a microblock. a) Schematic view of the stimulus space with the two dimensions *Identity* and *Emotion.* The center stimulus serves as the CS+ and is probabilistically reinforced with an electric shock. b) One microblock consists of one presentation per stimulus including the CS+ and a reinforced trial with the CS+. Not shown is the null trial and an oddball trial that replaces the reinforced trial with the CS+ in 25% of the trials.

Microblocks for the dynamics of learning. Following Onat (2018), I used a microblock design. In every microblock, all stimuli are shown to the subject and the CS+ is reinforced probabilistically. The center stimulus served as the CS+ in both, the happy and angry condition (Figure 4.5a). Using concurrent psychophysiological and neural recordings (as in Onat (2018)) or intermittent ratings between microblocks (as in my case), one can investigate the dynamics of generalization. In my studies, every microblock consisted of 27 trials. One trial per stimulus including the CS+, which was not reinforced in this case, one *null trial* in which nothing was presented and either a reinforced trial with the CS+ or an oddball trial. Those were used to ensure ongoing attention and consisted of a blurred face that was not recognizable. Subjects were instructed to react to this face. A schematic depiction of a microblock is given in Figure 4.5b.

Before each trial, a fixation cross was shown in the middle of the screen for 850 ms. The face was presented for 1.5 seconds in conjunction with a fixation cross. The fixation cross jumped from the forehead to the mouth region and subjects were instructed to follow it. This way I made sure that subjects perceived the full face and accounted for individual differences in gaze patterns. In reinforced trials, the electric shock started 100 ms before the end of the presentation. Inter-trial intervals (ITIs) were 3, 4, 5 or 6 seconds, each with a probability of 25 % and equally distributed over trial types.

Preventing higher order learning. In order to prevent second-order learning (e.g. which face follows which), the order of the stimuli was pseudo-randomized using a *type 1, index 1* sequence (Aguirre et al., 2011; Nonyane & Theobald, 2006). These sequences balance second-order transitions between states. To generate such sequences, I ran the `designseqran` C++ program[8] for one hour. Full *type 1, index 1* sequences consist of as many microblocks as possible states, i.e. stimuli. For time constraints that I imposed in order to keep data quality high, I used a truncated sequence of 20 microblocks. To generate this sequence, I computed the lack of orthogonality between stimulus indicators and positions[9] for all possible sub-sequences and chose the most favorable sub-sequence. Sequences for individual subjects in all three studies were generated by keeping the indicators for null trials, oddball trials and CS+ trials constant, while replacing the rest of the sequences with random permutations of the remaining stimulus indicators.

Sample description. The sample of the behavioral study comprised a total of 53 subjects, all of which had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders. Before starting the experiment, I informed subjects about the experiment and obtained written consent.

Three subjects dropped out after the first day of the experiment, leaving me with 50 full data sets. Of these, 35 were female (15 male). The mean age was 25.96 with a range of 18-37. Because of scheduling conflicts, one subject each in the happy and angry condition did not participate in the quadruplet task. Due to the hierarchical nature of the perceptual model, I could include them in the analysis by entering their responses as missing data and effectively replacing their parameters with the group level distribution.

4.2.2 Model predictions

Knowing the experimental design and the stimuli used, I could make predictions for the five rating blocks by feeding the integrated Bayesian model (chapter 2) with the stimuli and the reinforcement history. In addition I had to choose priors for the parameters of the model that were consistent with what I believed to be the prior knowledge that subjects brought into the experiment. The important aspect of the stimuli is that they differ on two dimensions, but only one of them is informative with respect to an aversive outcome. Emotional expressions have a social signaling function (Keltner & Kring, 1998). As a result, our brains are primed to detect emotional expressions, which makes them more salient than other stimuli (Hodsoll et al., 2011; Vuilleumier, 2005). One effect of this is that emotionally salient stimuli are more easily conditioned than neutral stimuli (Dimberg & Öhman, 1996; Orr & Lanzetta, 1980). Based on this, I assumed that the emotional dimension should be more informative a priori. In the context of the Bayesian model, this should be reflected in a prior that favors larger values for λ along the emotion dimension and smaller values on the identity dimension. To implement this, I used *Gamma*[10] priors with different parameters on both dimension:

[8] Available at https://www.bioss.ac.uk/people/cmt/designseq.html.

[9] This refers to Equation 4 in Nonyane and Theobald (2006)

[10] I am using the `Stan` convention for parameterization with shape and inverse scale.

$$\lambda_{emotion} \sim Gamma(2, 1) \tag{4.2}$$

$$\lambda_{identity} \sim Gamma(1, 1.5) \tag{4.3}$$

Note that this is the same for angry and happy faces because in both cases the emotion dimension was assumed to be more salient than the identity dimension. Those priors indicate that the identity dimension is relatively irrelevant because differences on this dimension do not change the outcome expectation by much. In addition, there is higher certainty about $\lambda_{identity}$ than about $\lambda_{emotion}$. Accordingly, more new information is needed to overwrite this belief state. In contrast, the prior on $\lambda_{emotion}$ favors larger values, which indicates that the emotional dimension is more informative. Both priors are shown in Figure 4.6.

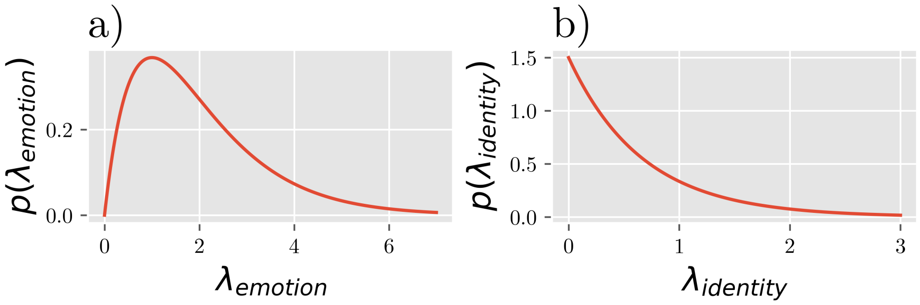


Figure 4.6: Priors on λ. The prior distribution for λ is dependent on the social function of emotional expressions and implicates that emotionality is more informative than identity with respect to an aversive outcome. As a consequence, the prior on $\lambda_{emotion}$ favors higher values than the prior on $\lambda_{identity}$, which corresponds to wider generalization along the identity dimension.

Another likely effect of the emotionality of faces are assumptions about the midpoint μ, i.e. which faces are more likely to predict an aversive outcome a priori. Intuitively, one would expect that angry faces appear more dangerous than both neutral and especially happy faces. This is in line with the idea of *preparedness* (Ohman & Mineka, 2001; Seligman, 1970). However, some findings show that just like angry faces, happy faces are more easily associated with an aversive outcome than neutral stimuli (Stussi et al., 2018, 2021). I believe this can be explained by the salience only and does not justify the assumption that happy faces are considered more dangerous. Expanding on that, Öhman and Dimberg (1978) reported diminished extinction for angry, but not for neutral and happy faces and Orr and Lanzetta (1980) reported faster aversive conditioning for angry than happy faces. Therefore, I followed the preparedness hypotheses and assumed that the prior on μ is skewed towards stronger emotional expression for the angry condition and towards neutral faces for the happy condition. I did not expect the two identities to have a differential effect (i.e. none of them appeared more dangerous a priori), so I used a uniform prior for μ on the identity dimension:

$$\mu_{angry} \sim Beta(6,1) \tag{4.4}$$

$$\mu_{happy} \sim Beta(1,6) \tag{4.5}$$

$$\mu_{identity} \sim Uniform(-1,1) \tag{4.6}$$

Those priors are depicted in Figure 4.7[11].

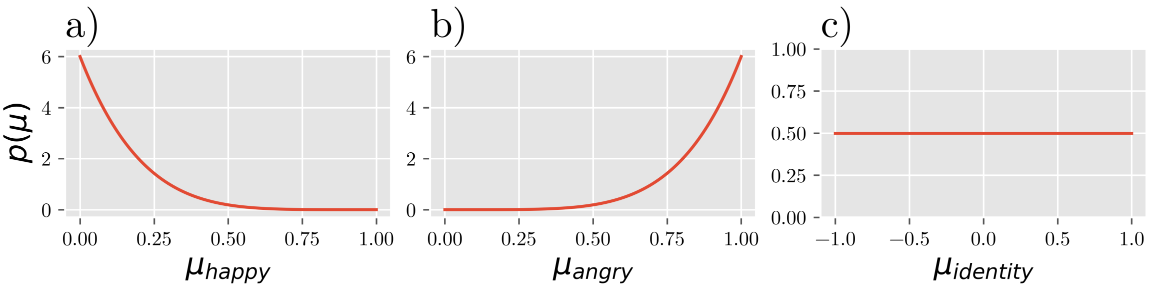


Figure 4.7: Priors on μ for aversive conditioning. The prior distribution for μ is dependent on the social function of emotional expressions. **a)** Using angry faces, it is heavily skewed towards more angry faces. **b)** In contrast, it is skewed towards more neutral faces when using happy faces. **c)** Since none of the identities is more dangerous *a priori*, the prior for μ is uniform.

Lastly, I needed to define a prior on the outcome probability ρ. Because subjects were informed that a face would be reinforced *occasionally* and because ρ is bounded between 0 and 1, I chose a *Beta* distribution that is skewed towards smaller values:

$$\rho \sim Beta(1,4) \tag{4.7}$$

This prior is depicted in Figure 4.8.

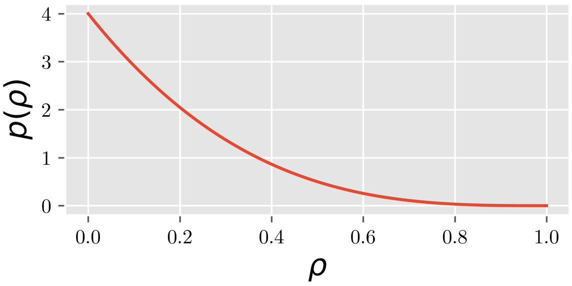


Figure 4.8: Prior on ρ. In line with the instructions of an *ooccasionally* aversive outcome, the prior is skewed towards smaller values.

Simulation. To make predictions, I simulated a Bayesian observer by implementing the model in `Stan`. The priors I outlined in the previous section define a specific support for the

[11] Note that the priors on the emotion and identity dimension have different supports because the *Beta* family of probability distributions is only defined for the range $(0,1)$ but the *Uniform* distribution I chose has support on the interval $[-1,1]$. The reasoning for this choice is given below.

different parameters. In particular, I chose a *Beta* distribution for $\mu_{emotion}$, which implies a support on the interval $(0, 1)$, but a $Uniform$ distribution for $\mu_{identity}$ with a support on the interval $[-1, 1]$. Intuitively, the support of those priors is the extent of the psychological space. The choice for different supports on both dimensions has two reasons. First, there are natural endpoints for the emotional expression dimension as there are completely neutral and completely emotional faces. The same is not true for identity, since faces that are more dissimilar to either of the faces are imaginable. Second, my model as well as other Bayesian models of inductive reasoning have an unintuitive property that was first pointed out by Navarro et al. (2008): A uniform prior on the support of a midpoint parameter does *not* imply a uniform generalization gradient, if the support is constrained to a specific interval. This is because on average points in the middle of this interval are closer to the midpoint of all possible consequential regions or associative maps than points at the edges, which results in a generalization gradient that is peaked in the middle of the dimension. This effect is incompatible with my assumptions about the information of facial identities. Therefore, I chose a uniform prior on the identity dimension with a support that is wider than the range of faces I considered. In particular, I considered faces with values in the interval $[-0.5, 0.5]$ on the identity dimension but a support of $[-1, 1]$ for the prior. Besides the priors, I needed to choose a standard deviation for perceptual noise. I used a value of $\sigma = 0.05$ for all simulations. Predictions were generated by feeding the same sequence of stimuli and outcomes into the model that were also used in the experiment. Predictions that assumed the priors I outlined above are depicted in Figure 4.9. I inferred the following

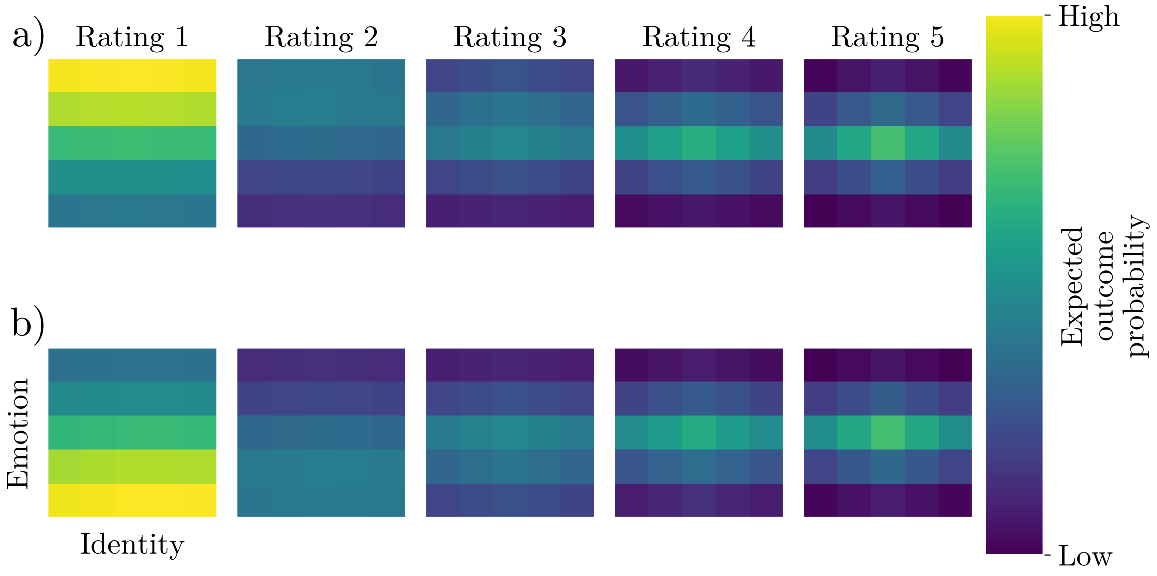


Figure 4.9: Model predictions for partial dimensionality reduction. Predicted shock expectation ratings for the five ratings in the **a)** angry and **b)** happy condition. These predictions assume two effects of prior knowledge, the first one being that the emotion dimension is considered more informative a priori, which is visible in **a)** and **b)**. **a)** The midpoint of the associative map on the emotion dimension μ_{angry} is shifted towards more angry faces and **b)** μ_{happy} is shifted towards more neutral faces.

hypotheses from these predictions:

1. Ratings are initially primarily driven by the amount of emotional expression. Ratings should be higher for angry than neutral faces and lower for happy than neutral faces.
2. This effect should decrease with increasing experience as new information is integrated.
3. (a) With ongoing conditioning, the proximity to the CS+ should become more relevant.
   (b) This effect should be stronger along the emotion than the identity dimension.
   (c) The width of the proximity-based gradient should decrease with increasing experience and thus increased certainty.
4. The extent of belief updating should become slower from rating to rating.

Finally, another possible implementation of dimensionality reduction is a *complete* one, where the identity dimension is completely ignored. This is equivalent to the assumption of a one-dimensional space in which faces only differ on the amount of emotional expression. I implemented such a model by removing the parameters $\mu_{identity}$ and $\lambda_{identity}$ from the model and only considering information about the emotional expression, not the identity of faces. The predictions of this model are depicted in Figure 4.10.

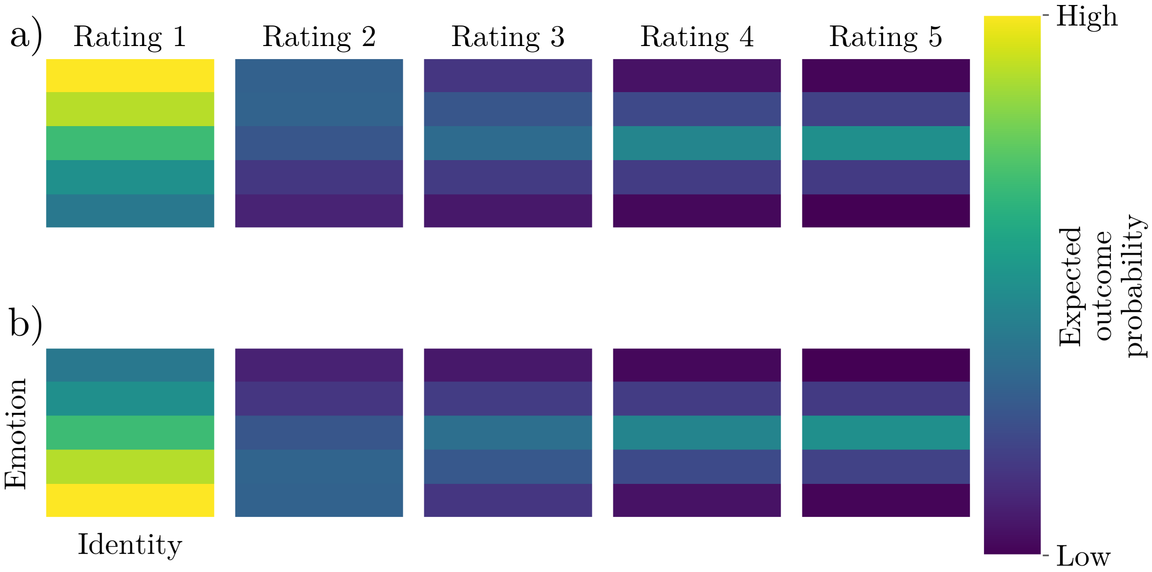


Figure 4.10: Model predictions for full dimensionality reduction. Another possible prediction for the ratings is that the identity dimension is completely ignored, which results in full dimensionality reduction.

4.2.3 Computational modeling

Both the fitting of perceptual spaces and shock expectation ratings rely on hierarchical Bayesian models. Perceptual spaces were fitted using hMPDS while behavioral data was fitted using heuristic approximations to the Bayesian model. These are outlined in the following paragraphs.

Perceptual spaces. The data from the quadruplet task was fit using hMPDS to generate both single subjects psychological spaces and a group level estimate that accounted for differences in estimation certainty between subjects. To identify the model I imposed the necessary constraints[12] by fixing the positions of two faces, for which I chose the minimum and maximum emotional expression versions of the morphed identity (first and last row of the middle column in Figure 4.4) and defined the positions to be $\{0.5, 0\}$ and $\{0.5, 1\}$ respectively. In addition I constrained the position on the identity dimension for the medium emotional expression version of the first identity (first column, second row in Figure 4.4) to be below 0.5, i.e. to the left of the line defined by the two fixed positions. These constraints were defined on the group level, while leaving all single subject stimulus positions unconstrained. I used information from the stimulus creation process to inform and constrain the posterior. For this purpose I chose the positions in the *optimal* grid (Figure 4.11) as prior means on group level stimulus positions. Priors were defined as normal distributions with $\sigma^2 = 1$:

$$\psi_{group} \sim Normal(\mu_\psi, 1) \tag{4.8}$$

This might sound narrow at a first glance but allows striking deviations from the expected solution that is implied by the prior means since the stimulus positions in the *optimal* solution are constrained to the set $\{x, y \in \mathbb{R}^2 \mid 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Perceptual noise

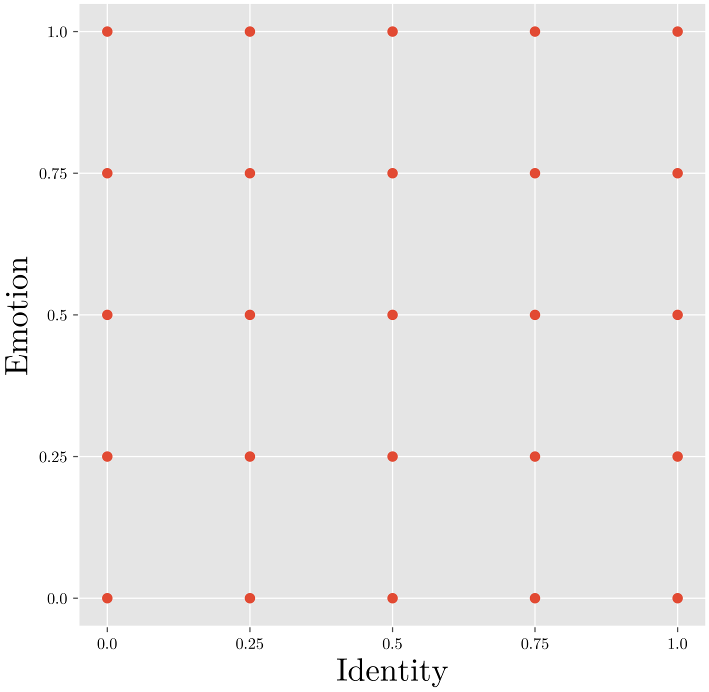


Figure 4.11: Optimal perceptual space. Based on the stimulus creation, the optimal perceptual space is a 5x5 grid of equidistant points in the range from 0 to 1.

[12] See Figure 3.2.

standard deviations were sampled on the log scale with a standard normal prior on the group level log standard deviation and a half normal prior with $\sigma^2 = 2$ for the between-subject variance.

$$\log(\sigma_{group}) \sim \mathcal{N}(0, 1) \tag{4.9}$$

$$\sigma_{psi} \sim HalfNormal(0, 2) \tag{4.10}$$

$$\log(\sigma_{subject}) \sim \mathcal{N}(\log(\sigma_{group}), \sigma_{psi}) \tag{4.11}$$

Interpolation of perceptual spaces. Perceptual spaces were fitted using a subset of all stimuli because the number of possible quadruplets for 25 faces is very large[13] and it was unfeasible to sample a sufficient subset of those. Instead only 9 faces were used in the quadruplet task. To arrive at the full perceptual space, I interpolated subjects' perceptual spaces using a second order polynomial linear regression. To be more precise, I computed the mapping from the optimal positions of those 9 stimuli (and the square of those positions) to the individually fitted positions. Applying the same mapping to the full grid of 25 stimuli gave me a complete perceptual space per subject. Interpolated spaces were then procrustes aligned to the optimal solution to generate the spaces that I used in further analyses.

Behavioral models. As argued in chapter 2, the full Bayesian model is intractable and it is practically impossible to fit it to data. Instead, I derived heuristic approximations that include parameters which can be interpreted with respect to the hypotheses that are implied by the model predictions. All of these models take into account the individual perceptual spaces. That is, the values for emotionality, identity and proximity to the CS+ for each subject are based on the fitted perceptual spaces.

Model 1a: The simplest model integrates the hypotheses about the impact of emotionality and proximity to the CS+ including simple temporal dynamics. This model assumes that the ratings in each block t can be expressed as a linear combination of a baseline shock expectation (β_0), the impact of emotion (β_{emo}), the impact of identity (β_{id}) and the impact of proximity to the CS+ (β_{prox}). The error model is $Normal$, i.e. residuals are assumed to be normally distributed around 0. As an equation, the predicted rating $\hat{r}_{s,f,t}$ for subject s, face f in block t was computed as

$$\hat{r}_{s,f,t} = \beta_0 + \beta_{emo,s,t}\mathrm{emo}_{s,f} + \beta_{id,s,t}\mathrm{id}_{s,f} + \beta_{prox,s,t}\mathrm{prox}_{s,f} \tag{4.12}$$

$$r_{s,f,t} \sim \mathcal{N}(\hat{r}_{s,f,t}, \sigma_s) \tag{4.13}$$

where $emo_{s,f}$ refers to the value on the emotional dimension for stimulus f and subject s and similarly for $id_{s,f}$ and $prox_{s,f}$. The subscript s on parameters and independent variables is needed because perceptual spaces and parameters were fitted for every subject

[13] To be precise, there are 303600 possible quadruplets.

in a hierarchical Bayesian framework. To implement the temporal component, I assumed a linear time dependence of parameters

$$\beta_t = \beta_0 + \beta_1 t \tag{4.14}$$

where I omit the subscripts for subjects and predictor for readability. The same linear time dependence was applied for all parameters. On the group level I used $Normal$ priors on the model components:

$$\beta_{0,group} \sim \mathcal{N}(5, 5) \tag{4.15}$$

$$\beta_{1,group} \sim \mathcal{N}(0, 5) \tag{4.16}$$

Subject level parameters were normally distributed around the group level parameter with standard deviation σ_β:

$$\beta_s \sim \mathcal{N}(\beta_{group}, \sigma_\beta) \tag{4.17}$$

$$\sigma_\beta \sim HalfNormal(0, 2) \tag{4.18}$$

Importantly, this model only takes euclidean distance to the CS+ into consideration, i.e. a distance measure that depends on both dimensions equally. This is by design because it allows for a comparison with models that split the distance along the dimensions to test for dimensionality reduction. Still, the model on its own allows for a test of hypotheses 1, 2 and 3a.

Model 1b: To include the idea of partial dimensionality reduction, model 2a splits the distance to the CS+ into two components, one for each dimension. The model is otherwise identical to model 1a. Thus, the predicted rating $\hat{r}_{s,f,t}$ for subject s, face f in block t was computed as

$$\begin{aligned}\hat{r}_{s,f,t} = \beta_0 &+ \beta_{emo,s,t}\text{emo}_{s,f} + \beta_{id,s,t}\text{id}_{s,f} \\ &+ \beta_{prox-emo,s,t}\text{prox-emo}_{s,f} + \beta_{prox-id,s,t}\text{prox-id}_{s,f}\end{aligned} \tag{4.19}$$

where *prox-emo* and *prox-id* are the distances to the CS+ along the emotional and identity dimensions, respectively. This model allows for a test of hypothesis 3b.

Model 1c: Finally, model 1c drops the identity altogether and is supposed to test the hypothesis of full dimensionality reduction. Note that this assumption is incompatible with the Bayesian model. Dropping the information about identity from Equation 4.12 yields the following equation for the predicted ratings:

$$\hat{r}_{s,f,t} = \beta_0 + \beta_{emo,s,t}\text{emo}_{s,f} + \beta_{prox-emo,s,t}\text{prox-emo}_{s,f} \tag{4.20}$$

Model 2a: All models so far assume a simple linear temporal dynamic of effects. Model 2a relaxes this constraint by adding another parameter λ that allows for non-linear dynamics. Instead of Equation 4.14, the temporal update is

$$\beta_t = \beta_0 + \beta_1 t \lambda^t. \tag{4.21}$$

This results in a decrease of belief updating for $\lambda < 1$, an increase for $\lambda > 1$ and constant updating for $\lambda = 1$. Accordingly, λ allows for a direct test of hypothesis 4. The model is otherwise identical to model 1a. Due to the strong effect of deviations from $\lambda = 1$ that stems from the non-linear dynamic, I used a narrower prior than for the other parameters.

$$\lambda \sim \mathcal{N}(1, 0.5). \tag{4.22}$$

This choice still allows for a wide range of values and is bordering on uninformative. Apart from the temporal dynamics, that allow for a test of hypotheses 4, the model is identical to model 1a.

Model 2b: This model is identical to model 2a except that it uses the split distance measure of model 1b.

Model 2c: Model 2c combines the full dimensionality reduction of model 1c with the non-linear temporal dynamics of model 2a. The model is otherwise identical to model 2a.

Model 3a: Model 3a incorporates hypothesis 3c, i.e. that the proximity-based part of the generalization gradients should become narrower over time, as there is less uncertainty left about the fact that only the CS+ predicts a shock. To implement this, I added another parameter ρ that is itself exponentially time dependent:

$$\begin{aligned} \rho_t &= \rho_0^t \\ prox_t &= prox_0^{\rho_t}. \end{aligned} \tag{4.23}$$

Here, $prox_0$ is the proximity that is implied from the fitted perceptual spaces. For values of $\rho < 1$, generalization around the CS+ becomes wider over time, for values of $\rho > 1$ it becomes narrower, which is the prediction of hypothesis 3c. Similar to λ, ρ_0 has non-linear effects and small deviations from 1 have a strong effect. To account for that, the prior is narrower than for the β parameters:

$$\begin{aligned} \rho_0 &\sim \mathcal{N}(1, 0.5) \\ \sigma_\rho &\sim HalfNormal(0, 0.5) \end{aligned} \tag{4.24}$$

Model 3b: This model is similar to 2b, except that it adds the time-dependent proximity parameter of model 3a.

Model 3c: Likewise, this model implements full dimensionality reduction like model 2c, but adds the time dependent proximity parameter of model 3a.

4.2.4 Results

Perceptual spaces. I compared two models for perceptual spaces: one that assumes that perceptual spaces are constant and one that assumes that they change over time due to the conditioning process. This is especially important since some studies have suggested changes in perceptual accuracy around negatively reinforced stimuli (e.g. Laufer et al., 2016). While the model comparison I conducted can not account for local changes in perceptual accuracy, it can account for fundamental distortions in perceptual spaces. The model comparison is based on the ELPD of left out data points[14]. As shown in Table 4.1, the model that assumes constant perceptual spaces explained the data better in both conditions[15]. An inspection of the group level solutions revealed that the perceptual spaces were fairly close to the intended grid structure. From this I conclude that the perceptual spaces were approximately constant over time and adhered to the intended structure. Group level solutions for the angry and happy condition and an example interpolation for a single subject are shown in Figure 4.12.

	Model	
Condition	Constant	Dynamic
Angry	**-20472.03**	-20562.04
Happy	**-20164.23**	-20240.1

Table 4.1: Model comparison for perceptual spaces in behavioral study. In both conditions, the model with a constant perceptual space fits the data better than the model with a dynamic space as indicated by the ELPD. The winning model is indicated by bold numbers.

Behavioral models. Mean ratings for the five rating blocks in both conditions are depicted in Figure 4.13a-b. At a first glance, the ratings seem to follow the predictions of the Bayesian model strikingly well. In both conditions, ratings are initially driven by the amount of emotionality. As the conditioning progresses, ratings become more and more driven by the proximity to the CS+, where the proximity along the emotion dimension seems to be more relevant. This is evident from the fact that subjects generalized more strongly along the identity dimension.

The results of the model comparison for the behavioral study are shown in Table 4.2. In general, models that assume partial dimensionality reduction (models 1-3b) fit the data better than models that assume full dimensionality reduction (models 1-3c) and models that assume no dimensionality reduction (models 1-3a). This is in line with the predictions of the Bayesian model. The best fitting model was model 3b, which contains parameters for non-linear temporal dynamics of the generalization gradients and a change in the width of the proximity-based part of the generalization gradients over time.

[14] See subsection 4.1.3 for a detailed description.

[15] A *higher* ELPD implies better model fit.

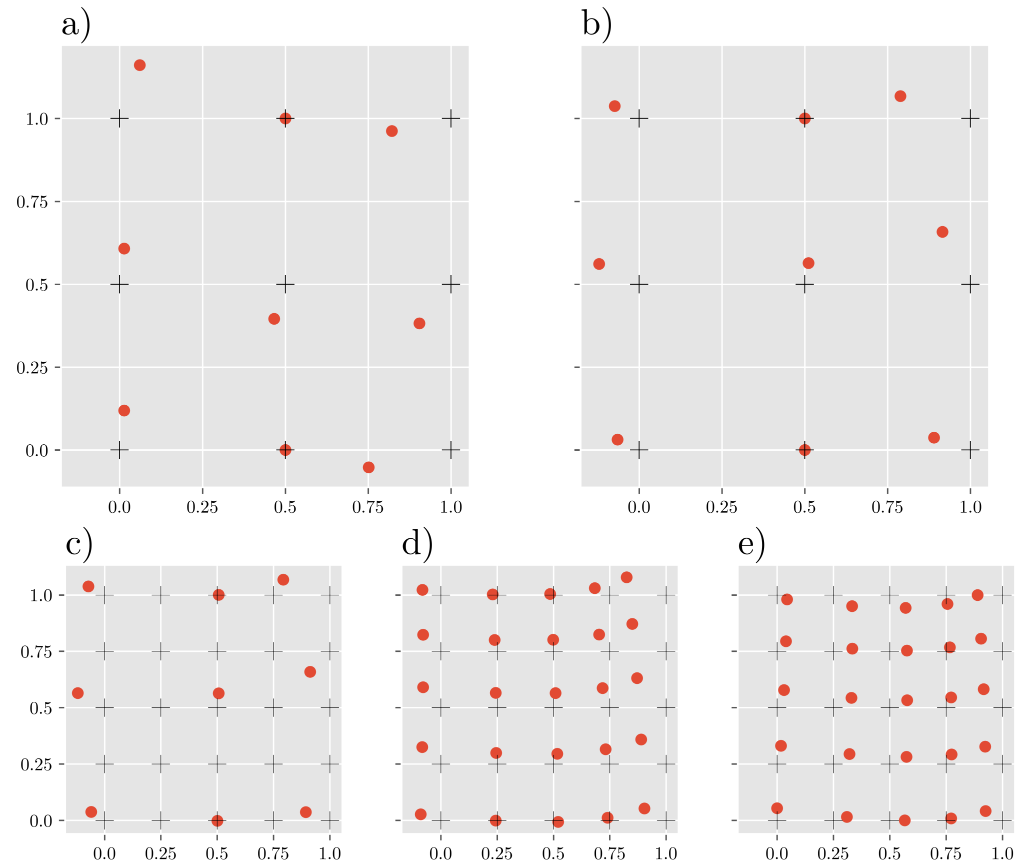


Figure 4.12: Group level perceptual spaces in behavioral study. Group level perceptual spaces for **a)** the angry and **b)** happy conditions are aligned reasonably well with the intended grid structure. **c)** Single subject perceptual spaces **d)** were interpolated to the full stimulus set and **e)** procrustes aligned with the optimal grid structure. Black crosses indicate the optimal $3x3$ and $5x5$ solution. Subfigures **c)-e)** show the space of a representative subject from the happy condition.

	Condition	
Model	**Angry**	**Happy**
1a	-11630.39	-11201.91
1b	-11483.62	-11091.93
1c	-11668.69	-11238.94
2a	-11282.82	-10906.31
2b	-11268.31	-10782.11
2c	-11512.20	-11003.92
3a	-11404.42	-10778.36
3b	**-11224.97**	**-10581.73**
3c	-11451.33	-10822.38

Table 4.2: Model comparison for behavioral models in behavioral study. The ELPD for each model is given for the angry and happy conditions. The best fitting model for each condition is model 3b, which is indicated in bold. This model assumes partial dimensionality reduction and comprises all hypotheses that were derived from the Bayesian model.

A visual inspection of the posterior predictive check for this model revealed that model 3b fits the data well (Figure 4.13c-d). To test the different hypotheses more directly, I looked

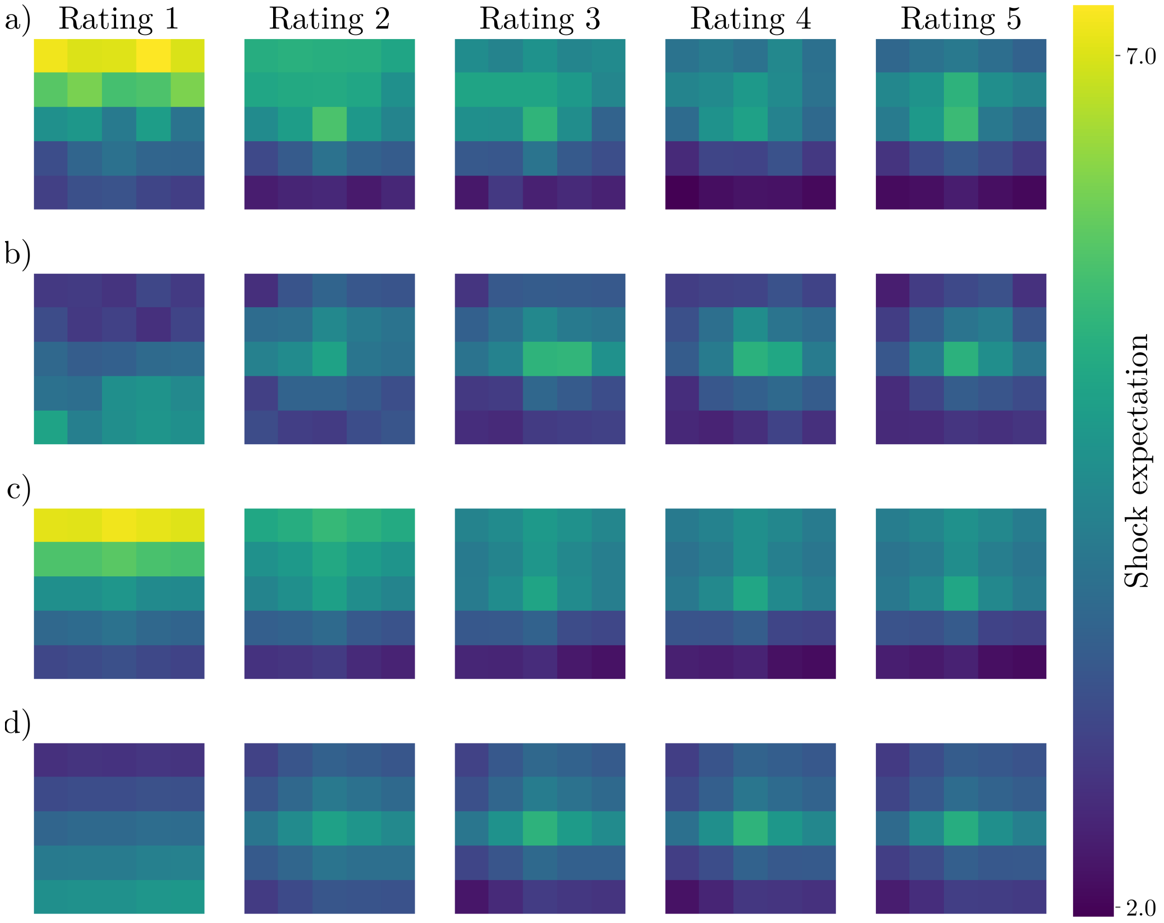


Figure 4.13: Mean ratings and posterior predictive checks for the behavioral study. Mean ratings in both **a)** the angry and **b)** the happy condition follow the model predictions fairly closely. The posterior predictive checks for **c)** the angry and **d)** the happy condition show that the model fits the data well.

at posterior distributions on the model parameters that can be interpreted with respect to the predictions of the model. Group level posterior distributions of the relevant parameters and the implied contribution of the different components for every rating are depicted in Figure 4.14.

In both conditions ratings were initially driven by the amount of emotionality, but not identity or by the proximity to the CS+[16], corroborating hypotheses 1a. The initial effect of emotion was stronger in the angry than the happy condition. As the conditioning went on, emotionality became less influential and ratings became more and more influenced by the proximity to the CS+. There was a clear discrepancy between the dimensions as this effect was much stronger along the emotionality dimension in both conditions. These findings were predicted by hypotheses 2 and 3a-b. The changes from rating to rating became smaller over time and the proximity-based part of the gradients became narrower, confirming hypotheses 4 and 3c respectively.

[16] Because these ratings are before any conditioning, subjects can not have learned about the CS+. Therefore I attribute the slight deviations from 0 in the initial effect of proximity to inflexibilities in the model.

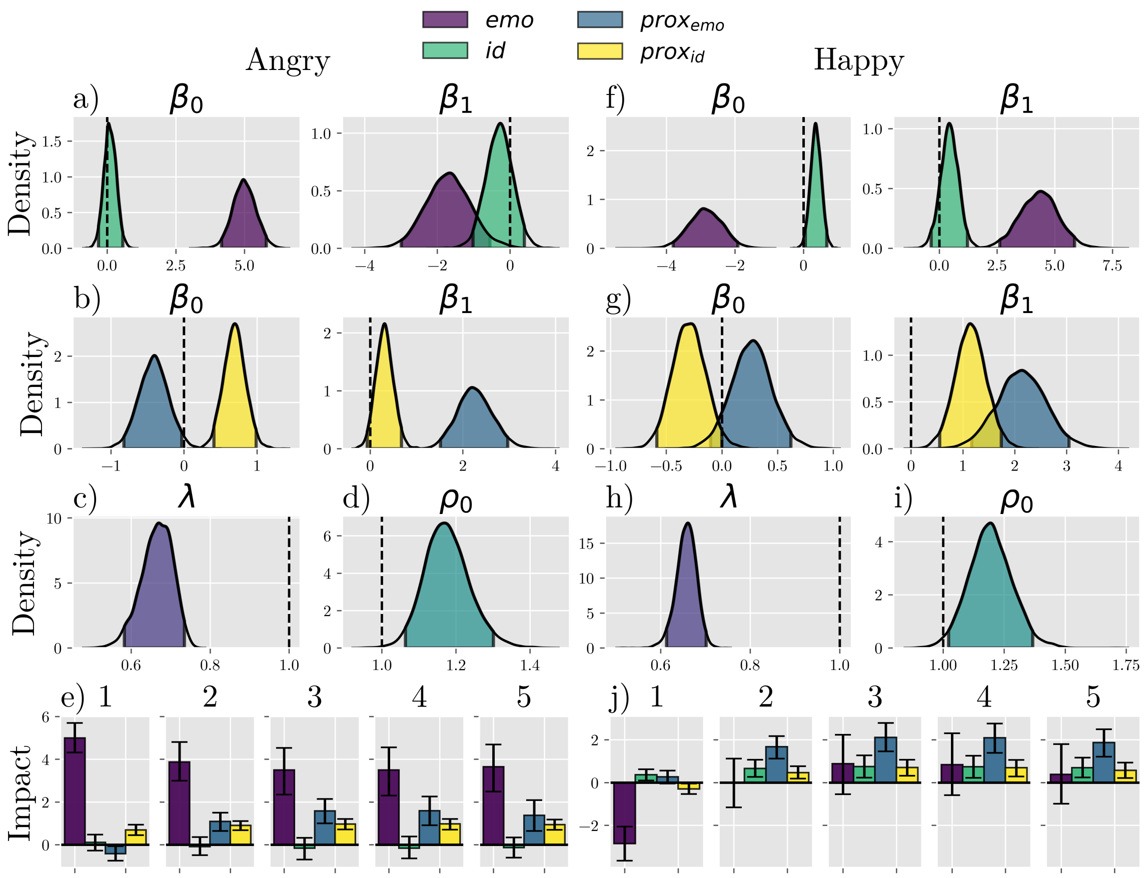


Figure 4.14: Posterior distributions of the model parameters for the behavioral study. **a)** Ratings in the angry condition are initially driven by the amount of emotionality. This effect decreases with time. In contrast, identity is not relevant at all. **b)** Proximity to the CS+ is not very relevant in the beginning. As the conditioning progresses, proximity becomes more relevant, with a much stronger effect along the emotion dimension. **c)** The belief updating from rating to rating and **d)** the width of the proximity-based part of the generalization gradients decrease with time. **e)** Bar plots show the contribution of the different components to the five ratings. **f)** In the happy condition, ratings are initially driven by the amount of emotionality as well, but the effect is in the opposite direction and weaker than in the angry condition. Identity is irrelevant here as well. **g)** Proximity to the CS+ does not play a significant role in the beginning, but becomes more relevant with time and especially along the emotion dimension. Like in the angry condition, **h)** the belief updating and **i)** the width of the proximity-based part of the generalization gradients decrease with time. **j)** Bar plots show the resulting contribution of all features in all ratings.

4.2.5 Interim discussion

The results of the behavioral study are in line with the predictions of the Bayesian model. Ratings followed the predictions of the model fairly closely. A model comparison of the different model approximations revealed that the best fitting model was model 3b, which assumes partial dimensionality reduction. Parameter inspections further corroborated that the results are well in line with the hypotheses that I derived from the Bayesian model.

In summary, the data supports the assumptions of the Bayesian model. While this study has a somewhat limited scope due to being a fear generalization study, the results encouraged me to continue with a fMRI study to investigate the neural mechanisms.

4.3 Neural mechanisms of fear generalization

The second study aimed to corroborate the findings from the previous study and to investigate the neural mechanism that underlie the proposed model. To this end, I replicated the first study, but collected fMRI data from the participants.

4.3.1 Behavioral methods

Apart from the fact that subjects lied in the scanner, the behavioral methods were almost identical to the first study. Some minor differences were unavoidable. During pain calibration, I ran a dummy echo-planar imaging (EPI) sequence. This sequence was not included in the analysis, as the only purpose was to account for possible effects of the scanner noise on pain perception. Additionally, subjects gave their responses using a button box instead of the keyboard.

Sample description. I collected behavioral and fMRI data from 62 subjects. Because I only analyzed data from subjects that didn't show problems in either of the two sessions, I had to discard 12 data sets because of the following issues (# of subjects):

- structural abnormalities in the anatomical image (1)
- fell asleep (2)
- excessive head movement (2)
- dropped out after the first session (3)
- electrode dispatched (4)

This left me with 50 complete data sets that were included in the analysis. In addition, three subjects in the angry and two subjects in the happy condition skipped the quadruplet tasks due to scheduling conflicts. Those were treated as described in paragraph 4.2.1.

Behavioral data analysis. The analysis of behavioral data was identical to the first study. I excluded one subject in the happy condition because they gave a constant shock expectation rating of 1 for all ratings. Two further subjects kept the behavioral Bayesian models from converging in the happy condition. As a result, I fitted all behavioral models with data from 50 subjects in the angry and 47 subjects in the happy condition.

4.3.2 fMRI methods

MRI data collection. FMRI data were collected on a Siemens 3T PRISMA scanner (Siemens, Erlangen, Germany) using a 64 channel head coil. To reduce the acquisition time, I employed a multiband sequence (multiband factor = 3, TE = 30ms, TR = 1.526, flip angle = 60°, FOV = 225 mm, GRAPPA PAT factor = 2, reference lines = 48). The field of view consisted of 63 slices per volume with an isotropic voxel size of 1.5 mm. In addition, I collected a B0 field map to correct for inhomogeneities in the magnetic field. Before the functional scans I also obtained a high-resolution (isotropic voxel size = 1 mm) anatomical image using a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence.

fMRI preprocessing. Preprocessing was conducted in `MATLAB` using `SPM12` (Wellcome Trust Center for Neuroimaging, London, UK). I discarded the first four EPI images per run to avoid artifacts. The actual preprocessing consisted of spatial motion correction (realignment and field map correction) and slice timing correction. Mean EPIs over all runs were calculated and segmented to generate native tissue maps. Those were used to calculate a flow field from the native subject space to standard Montreal Neurological Institute (MNI) space in the `DARTEL` toolbox of `SPM12`. I did not normalize raw functional images, but instead computed firstlevel analysis in native space and normalized the resulting images (e.g. beta images or correlation maps). The normalized images were smoothed with a Gaussian kernel (FWHM = 6mm) before entering them into the secondlevel analysis.

Univariate analysis. Univariate analysis was conducted in SPM12 using two general linear models (GLMs). Both GLMs use outputs of behavioral models and can therefore be considered model-based fMRI. Since model 3b showed the best fit to the behavioral data, I used this model to generate the parametric modulators for both GLMs. The first GLM was used to assess correlates of the modeled shock expectation. This is roughly equivalent to the approach of typical fear generalization studies in which researchers are looking for brain areas that show a generalization tuning (Webler et al., 2021), but it includes the temporal dynamics. In the second GLM I exploited the fact that the shock expectation was modeled as a linear combination of different features (e.g. the impact of emotion and proximity). This allowed me to assess the contribution of those features independently and probe the role of the FPN with respect to the different dimensions. In both GLMs I included one intercept per run and six motion parameters (three translations and three rotations) as regressors of no interest. Beta images from the firstlevel analysis were normalized to MNI space and smoothed with a Gaussian kernel (FWHM = 6mm). Using these images, I conducted a secondlevel analysis using a fixed effect analysis of variance (ANOVA) in SPM12. Secondlevel analyses contained images from both sessions (angry and happy). The rationale behind that choice is increased statistical power and that I was interested in effects that are independent of the emotion. After estimating the model, contrasts were computed as directed t-tests. I computed both negative and positive contrasts and applied a voxel-wise α-threshold of $p < 0.025$. This threshold refers to the whole brain family-wise error (FWE) corrected p-values to account for multiple comparisons and is equivalent to a two-sided t-test at $\alpha = 0.05$. Differential effects were computed as differential t-contrasts with a FWE corrected threshold of $p < 0.05$. In some cases I had specific hypotheses about the role of the FPN and used a small volume correction (SVC) within a mask of the FPN based on the segmentation by Yeo et al. (2011).

GLM 1: Shock expectation. For the first GLM I extracted model-predicted shock expectation ratings from the behavioral model. In the model those are defined for five ratings. To use them as parametric modulators on every trial, I interpolated them to 20 microblocks. In particular, in the model, time was modeled as the vector $t = [0, 1, 2, 3, 4]$. The interpolated values were then calculated by using the vector that contained 20 linearly spaces values from 0 to 4[17] instead. This approach assumes that the belief state within a microblock is constant. While this is technically not true, it is a reasonable approximation. The GLM included regressors for the onset of faces and the onset of shocks. The interpolated shock expectations were then entered as parametric modulators on the face onsets.

GLM 2: Feature contribution. Similar to the first GLM, I interpolated the features that were used to model the shock expectation to 20 microblocks. The second GLM included the same regressors and all features were entered as parametric modulators on the face onsets. The features were:

- **Baseline shock expectation**: β_0
- **Impact of emotion**: $\beta_{emo} *$ emo
- **Impact of identity**: $\beta_{id} *$ id
- **Impact of proximity along the emotional dimension**: $\beta_{prox-emo} *$ prox-emo
- **Impact of proximity along the identity dimension**: $\beta_{prox-id} *$ prox-id

Multivariate analysis. For the multivariate analysis I used RSA. RSA relies on beta images from a firstlevel analysis. Because I wanted to investigate the time component of representations, I used a least squares separate (LSS) approach (Mumford et al., 2012) to generate one beta image for every single trial in order to limit the problems that can arise from autocorrelation that is due to the slow BOLD signal. In the LSS approach, one computes as many GLMs as there are trials. In each GLM, there is one predictor for the trial of interest and one predictor for all other trials.

In RSA, one correlates different representational dissimilarity matrices (RDMs) with each other. A neural RDM that describes the dissimilarities between the neural representations of different stimuli and a model RDM that describes the dissimilarities between the stimuli based on a model or different properties. The neural RDMs are computed from the beta images that were generated from the LSS GLM.

Since the purpose of my multivariate analysis was to investigate the dimensionality of representations of faces and how they relate to the behavioral effects, I generated two different model RDMs per subject. Those were based on the perceptual dissimilarities of the faces along either the emotional or the identity dimension. The perceptual dissimilarities were calculated using the absolute value of the difference between the values of faces along that dimension[18]. Neural dissimilarities for each voxel were calculated using the correlation

[17] `linspace(0, 4, 20)` in `MATLAB`.

[18] This is equivalent to both the cityblock and the Euclidean distance, since those are the same in one-dimensional space.

distance

$$d_{corr}(x, y) = 1 - \mathrm{corr}(x, y) \tag{4.25}$$

between the representations of different faces x and y. To compute the correlations between the neural and the model RDMs, I ran a searchlight analysis with a custom kernel and a radius of 5 mm as implemented in the `brainiak` toolbox in `Python`. Correlations between model and neural RDMs were calculated as Spearman's rank correlation coefficient. The resulting correlation maps (one correlation per voxel) were then Fisher Z-transformed, smoothed with a Gaussian kernel (FWHM = 6 mm) and entered into a secondlevel analysis in `SPM12`. The secondlevel analysis was equivalent to the one described in the univariate analysis.

Visualization of fMRI results. All fMRI results are visualized as thresholded t-maps. I chose an uncorrected threshold of $p < .0005$ for visualization because the distribution of sub-threshold activation is still informative. This is especially important in the context of a network view because those sub-threshold activations give an idea about the overlap between activity and brain networks. The chosen threshold corresponds to an uncorrected two-sided t-test at $\alpha = .001$. Colormaps indicate the direction of the effect. Red indicates positive effects and blue indicates negative effects. Significant activations and deactivations are indicated by circles or region names. All fMRI plots were generated using the `nilearn` package in `Python`.

4.3.3 Behavioral results

Perceptual spaces. Subjects in the fMRI study went through the same perceptual task as participants in study 1, including the repetition after conditioning. To investigate whether spaces were subject to distortions due to the conditioning, I again fitted the two models to the data, that I described for the previous study. Corroborating the findings from the first study, the model that assumes constant perceptual spaces fit the data better than the model that assumes that spaces change over time. This can be seen in the higher ELPD as displayed in Table 4.3. A visual inspection of group level positions that were implied by the model that assumes constant spaces indicated that the fitted spaces were very similar to the one of participants in the first study. This can be seen in Figure 4.15, where I display the posterior expectations of group level positions. In the fMRI study, the positions of

	Model	
Condition	Constant	Dynamic
Angry	**-19037.65**	-19091.26
Happy	**-18672.03**	-18800.57

Table 4.3: Model comparison for perceptual spaces in fMRI study. Results in the fMRI study corroborate the results from the behavioral study. In both conditions a constant perceptual space explains the data better than a dynamic space.

faces were close to the intended positions along a perfect grid. Importantly, the dimensions spanned an approximately similar range of values. This indicates that differences in the width of generalization along these dimensions were due to prior assumptions and not due to differences in discriminability.

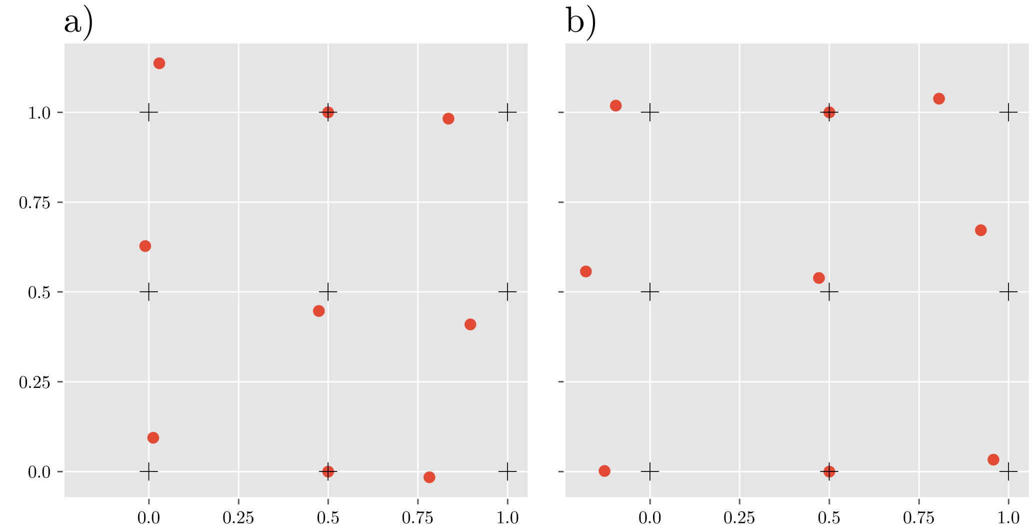


Figure 4.15: Perceptual spaces in fMRI study. The perceptual spaces that were fitted to the data in the fMRI study for the **a)** angry and **b)** happy condition. The positions of the faces are close to the intended grid structure and very similar to those from the behavioral study.

Behavioral models. As can be seen in Figure 4.16a-b, mean ratings in the fMRI study were similar to the ratings from the first study at a first glance. The major difference seems to be that subjects learned faster and the initial bias towards emotional faces subsided quickly. Apart from that, all the major hallmarks that were visible in the first study emerged again in the fMRI study. Ratings were initially dependent on the emotional expression of faces but depended increasingly on the proximity to the CS+ over time. As in the first study, this effect was more pronounced along the emotional dimension. Another feature that I could replicate from the first study was that the initial bias was stronger for angry than happy faces, which is likely due to the congruence between negative emotions and pain (Seligman, 1970).

More formally, I fitted the same models to the data as in the first study. The results of the model comparison (Table 4.4) confirmed the findings from the first study, as model 3b was the best-fitting model and in general models that included separate distance measures along both dimensions (i.e. models 1-3b) fit the data better than models that assume only proximity along the emotion dimension (i.e. models 1-3c) and models that include the joint cityblock distance along both dimensions (i.e. models 1-3a). Using the posterior predictive check in Figure 4.16c-d, I concluded that model 3b explained the data very well.

A consideration of posterior distributions on model parameters corroborated the impressions from the visual inspection of mean ratings and replicated the results from the first study. As depicted in Figure 4.17, posteriors on the parameters for the initial impact of

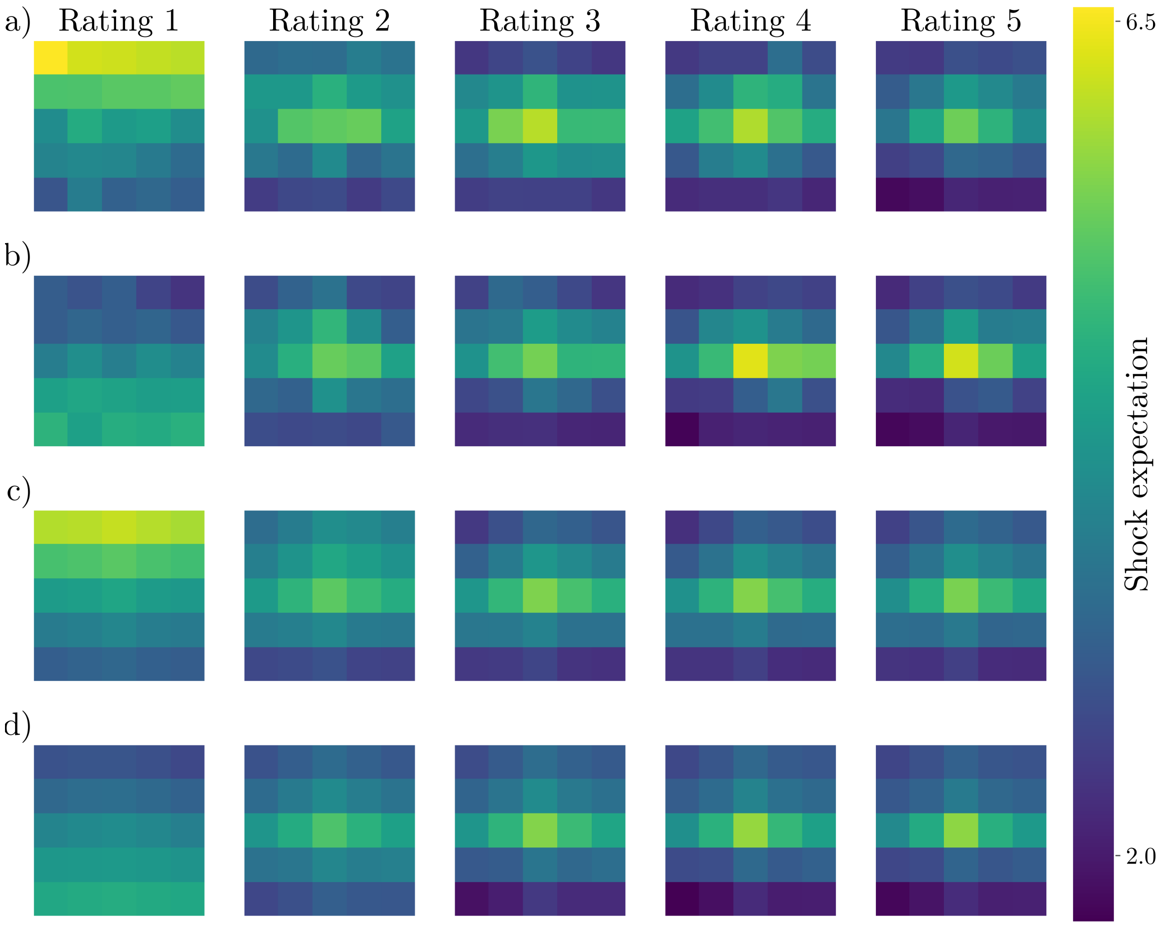


Figure 4.16: Mean ratings and posterior predictive checks for the fMRI study. a) While subjects in the fMRI study learned about the relevance of proximity to the CS+ and unlearned the initial impact of emotionality faster than in the first study, the pattern is otherwise very similar. Initially, ratings depended on emotionality. Proximity – especially along the emotion dimension – became more important over time. **b)** Similarly, subjects in the happy condition learned faster than in the first study, but ratings showed a strong resemblance to the ones from the first study. The differences in learning speed might be attributed to the change in setting, as being in a fMRI scanner might have made subjects more attentive. **c)-d)** Posterior predictive checks show a strong resemblance to the mean ratings and indicate that the model fits the data well, both in the **c)** angry and **d)** happy condition.

the different features were around or close to 0 for identity and the proximity along both dimensions, but clearly different from 0 for emotion. Again, this effect was stronger in the angry condition. Parameters for the change of the impact indicate that the initial impact of emotion vanished, the impact of identity did not change and the impact of proximity increased along both dimensions, but much stronger so along the emotional dimension. With respect to the non-linear temporal dynamics, the change from rating to rating went down and the width of the proximity component went down. All of those findings are in the line with the suggested hypotheses and the results from the first study.

	Condition	
Model	**Angry**	**Happy**
1a	-12143.48	-12366.98
1b	-12038.73	-12149.18
1c	-12224.96	-12337.97
2a	-11834.45	-12055.02
2b	-11684.78	-11814.15
2c	-11899.48	-12036.83
3a	-11760.16	-11981.41
3b	**-11562.81**	**-11724.3**
3c	-11804.61	-11915.1

Table 4.4: Model comparison for behavioral models in fMRI study. Model comparison corroborated the finding from the first study, namely that the best fitting model is model 3b, the one that includes all hypotheses that were based on the Bayesian model.

4.3.4 fMRI results

The analysis of fMRI data analysis consists of a univariate approach using two GLMs and multivariate analysis using RSA. Those approaches are reported separately in the following paragraphs.

Univariate analysis. Since model 3b provided the best fit to the data, I used this model to generate parametric modulators for the model-based fMRI analysis. In particular, I interpolated model-predicted shock expectations and the features that comprise those expectations to 20 microblocks and entered posterior expectations of those quantities as parametric modulators into GLMs. The interpolation and separation of the features is described in more detail in subsection 4.3.2 and visualized in Figure 4.18.

GLM1. GLM1 used the interpolated shock expectations as parametric modulators (Figure 4.18c). The results from the second-level ANOVA are depicted in Figure 4.19. I found two distinct networks that either showed a positive and a negative generalization tuning[19]. Virtually all positive tunings were located in regions of either the FPN or the SN (Figure 4.19a). Negative activations were exclusively located in the DMN (Figure 4.19b). The specific areas with significant positive and negative generalization tuning are listed in the legend of Figure 4.19. These results corroborate findings from a meta-analysis on neural correlates of fear generalization (Webler et al., 2021). However, unlike the studies that were included in that meta-analysis, my analysis does not only measure the correlates of a final behavioral outcome in a generalization phase, but instead tracks the temporal dynamics of the generalization process. Still, due to the correlation nature, these findings suffer from the same limitations as previous publications. Especially when viewing the activations through the lens of brain networks, it is important to keep in mind that the analysis can

[19] In the context of this analysis, a positive tuning refers to a positive correlation of BOLD activity with interpolated shock expectations, whereas a negative tuning refers to a negative correlation.

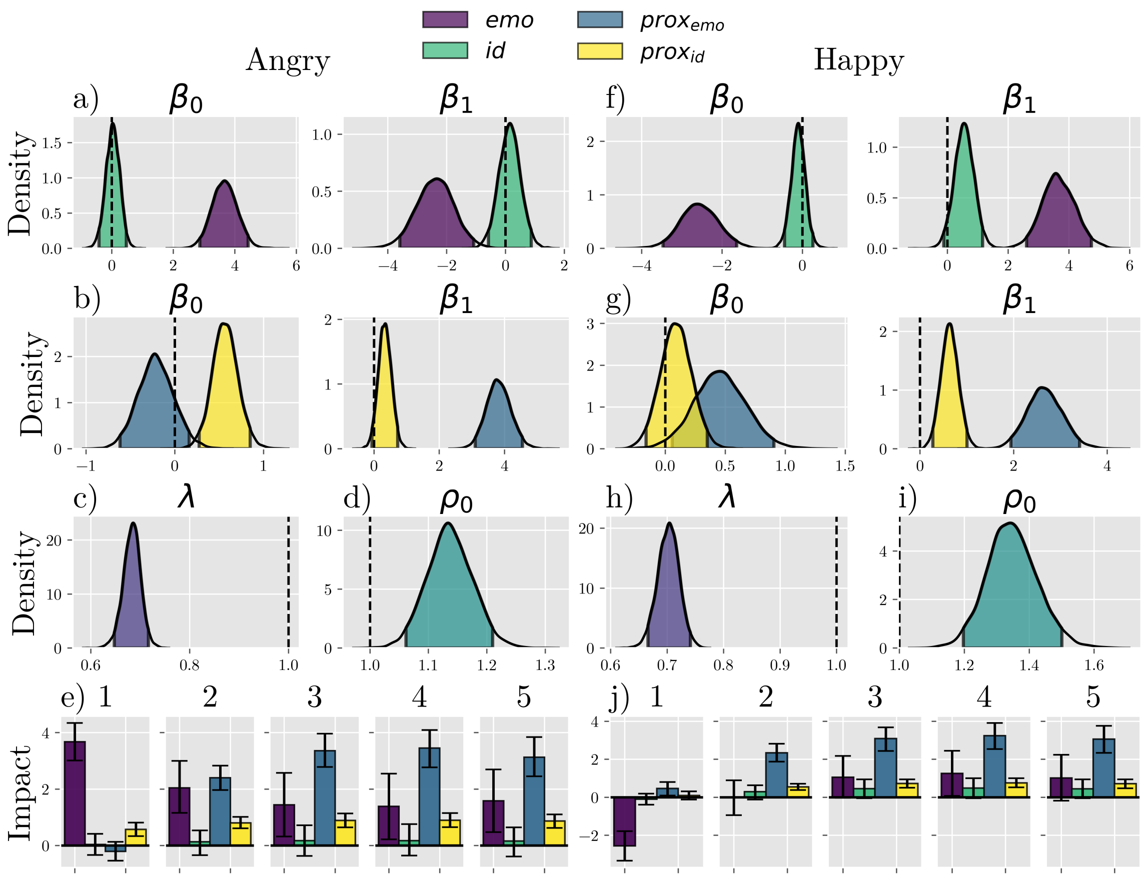


Figure 4.17: Posterior distributions of the model parameters for the fMRI study. Overall, posterior distributions are similar to those of the first study. In the angry condition, **a)** the initial impact of emotion is clearly different from 0 and dimishes over time while the impact of identity is relatively static around 0. **b)** The impact of proximity is negligible at the beginning but increases over time, especially along the emotional dimension. **c)** From rating to rating, ratings stabilize, and **d)** the width of the impact of proximity decreases. **e)** When viewing the amount of dependence of the different ratings on the different features, the expected pattern of an initial impact of emotion and an increasing impact of proximity along the emotion axis emerges. **f)** In the happy condition, the initial impact of emotion is negative and trends towards 0 over time. **g)-i)** Apart from that, the pattern of parameter estimates mirrors both the angry condition and the first study. **j)** The relatively weaker initial impact of emotion leads to a faster, stronger dependence on proximity along the emotional dimension and a faster decline of the impact of emotion.

not distinguish between the generalization process itself and the results of it. For instance, stimuli for which subjects have a higher subjective outcome expectation are more salient, which would lead to a SN mediated switch from DMN to FPN activity (Goulden et al., 2014) – a pattern that is well in line with the reported activations.

GLM2 Due to the usage of interpolated features instead of the full shock expectation, the second GLM provides a more fine-grained view and helps with the interpretation of the role of different areas and networks. In addition, it allowed me to investigate how the dimensional preference for the emotion dimension correlates with brain activity. This

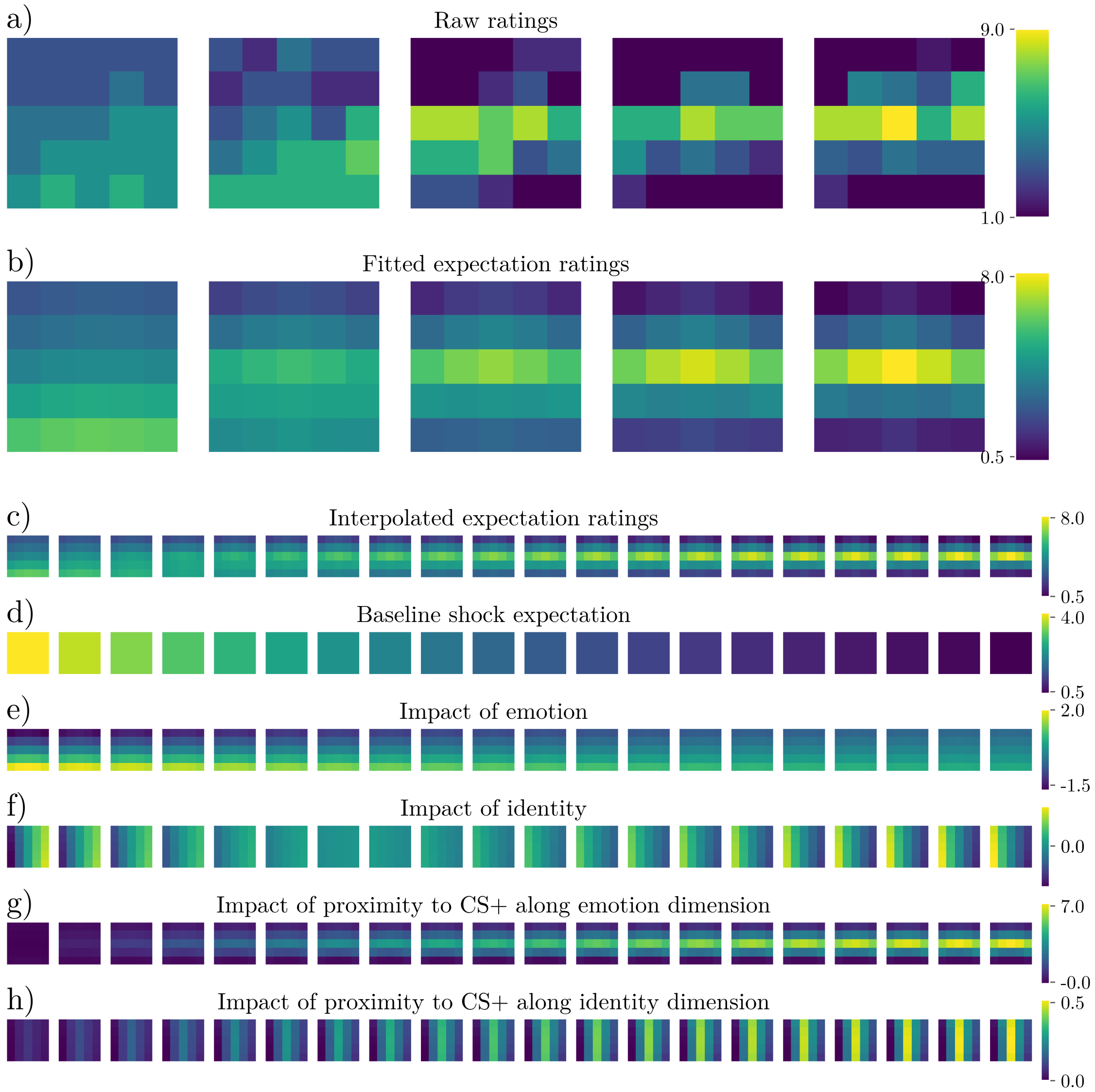


Figure 4.18: Interpolation of behavioral models to the fMRI study. a) Raw ratings of a subject in the happy condition. **b)** The model predictions for the data in **a)**. For the fMRI analysis, these model predictions needed to be interpolated, because there are only five ratings, but 20 microblocks. This resulted in the predicted shock expectations for all micoblocks in **c)**. For the second GLM, I interpolated the features that additively make up the model predictions separately. The resulting features are **d)** a baseline shock expectation per microblock, **e)** the impact of emotion, **f)** the impact of identity, **g)** the impact of proximity along the emotional dimension and **h)** the impact of proximity along the identity dimension.

is especially interesting due to the established role of the FPN in representation learning (Niv, 2019). A second level ANOVA and directed *t*-contrasts revealed different significant correlations with the distinct features.

I found a significant correlation of baseline shock expectation in the left PCU (Figure 4.20a). The PCU has been shown to be important in the context of social anxiety disorder (SAD), as SAD patients showed both increased gray matter volume (X. Wang et al., 2018) and functional and network deficits (Yuan et al., 2018) in this region. Those find-

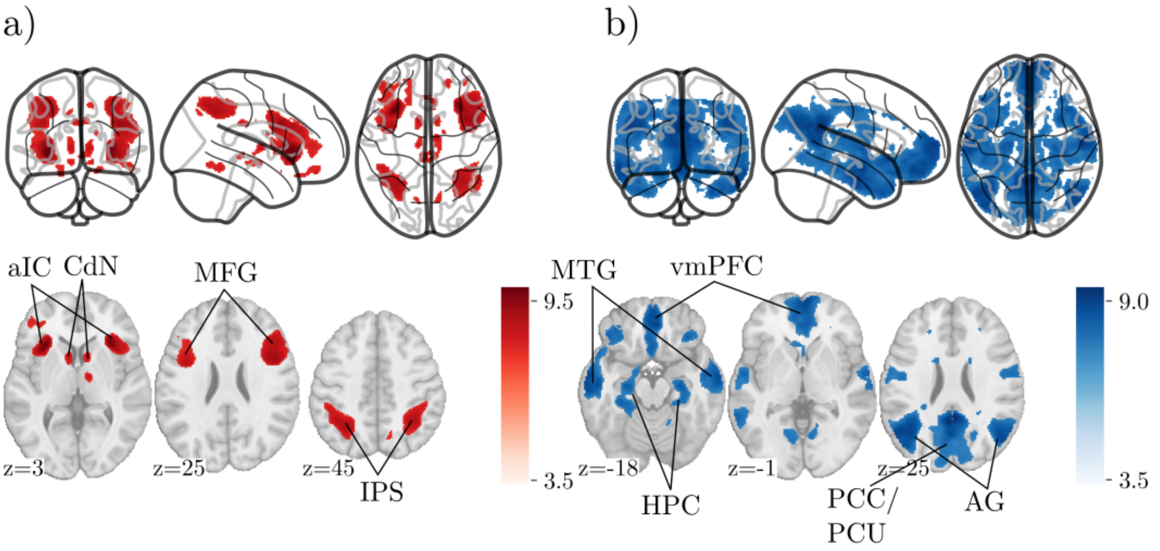


Figure 4.19: GLM1 results. Positive and negative correlations of BOLD responses with interpolated shock expectation ratings were found in two distinct networks. The first row shows glass plot brain to give an idea about the distribution of those networks. **a)** Positive generalization gradients were found in regions of the FPN – MFG, IPS and CdN – and in the aIC, a major hub of the SN. **b)** Negative gradients appeared in all important regions of the DMN: vmPFC, HPC, MTG, PCC, PCU and angular gyrus (AG).

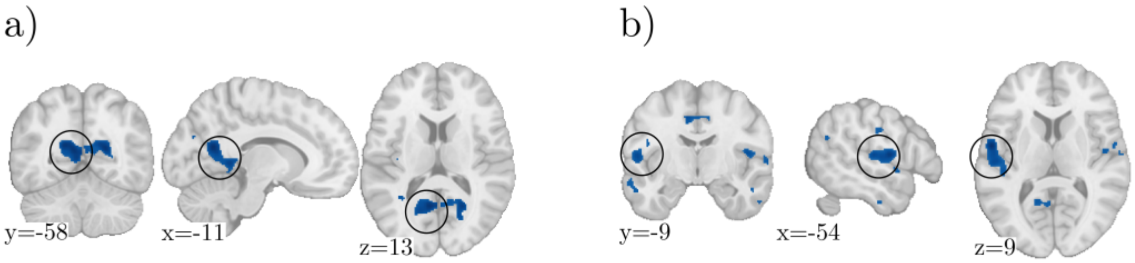


Figure 4.20: GLM2: fMRI correlates of baseline shock expectation and emotion. a) The baseline shock expectation was negatively correlated with BOLD activity in the left PCU. **b)** The impact of emotion was negatively correlated with BOLD activity in the left parietal operculum (PO), corresponding to the human secondary somatosensory cortex (S2).

ings indicate that the deactivation in my study might be specific to the use of facial stimuli. A single significant negative correlation with the impact of emotion was found in the left parietal operculum (PO), which corresponds to the secondary somatosensory cortex (S2) in humans. This contrast is especially interesting, because it defines the neural correlates of the prior knowledge about the predictive value of emotional expressions. A view at the relevant literature suggests some interpretations. Adolphs et al. (2000) found that the S2 is involved in the processing of emotional facial expressions. They investigated impairments in emotion recognition in patients with brain legions and found lesions in S1 and S2 to be associated with a deficit in the recognition of emotional expressions. Zeidan et al. (2015) found that correctly cued pain stimuli led to greater activation in the PO than violated expectations. This suggests a role of the PO in the assessment of congruency of outcomes with expectations. Drevets et al. (1995) reported results from a positron emission tomography (PET) study, in which participants expected somatosensory stimulations. In this study, blood flow in S2 decreased bilaterally during the anticipation of stimulations, among other regions. Gijsen et al. (2021) conducted an electroencephalography (EEG) study in which participants had to learn statistical regularities in patterns of somatosensory stimulations. The found that the mismatch negativity[20] correlated with Bayesian surprise, i.e. an uncertainty weighted prediction error. Interestingly, they used source reconstruction and could locate the source of the mismatch negativity in S2. This would indicate that deactivations in S2 are used as a signal to update a belief state rather than an anticipatory signal. Of course it is possible that the PO is involved in both processes. It is interesting to note that the PO is involved in the processing of both emotional facial stimuli and somatosensory anticipations. Those two processes together make S2 a prime candidate for the computation of a prior expectation of pain given the emotionality of faces. This further suggests that the encoding of prior expectations depends on the modalities of stimuli and outcomes.

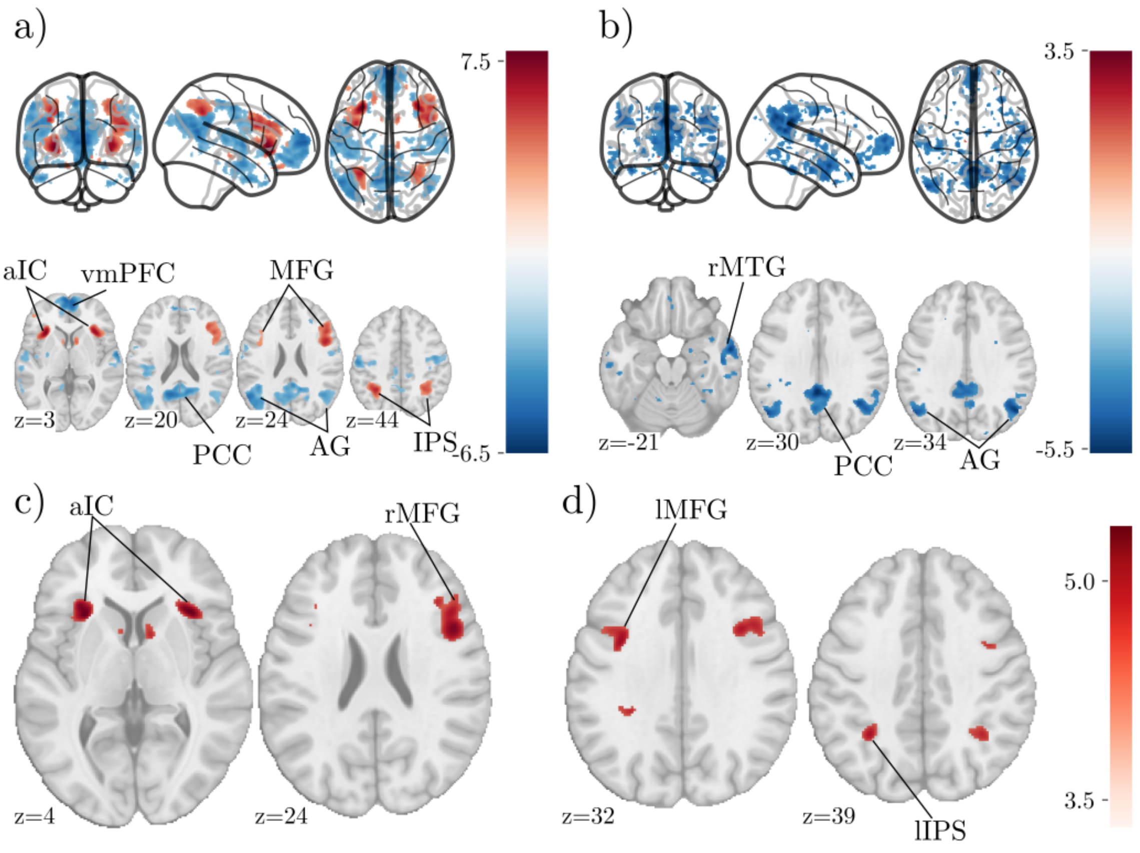


Figure 4.21: GLM2: fMRI correlates of the impact of proximity to the CS+. a) Neural correlates of the impact of proximity to the CS+ along the emotion dimension closely mirrored the activations that were associated with interpolated shock expectation ratings with negative correlation in the DMN (vmPFC, PCC and angular gyrus) and positive correlation in the FPN (MFG and IPS) and the aIC. **b)** The impact of proximity along the identity dimension showed negative correlations in the DMN as well (rMTG, PCC and AG), but virtually no positive correlations.

[20] A common EEG response to violations in expectations.

Correlations with the impact of proximity to the CS+ showed a very interesting pattern. Activations related to the proximity along the emotional dimension closely mirrored the correlations with the full interpolated shock expectation in GLM1, with negative correlations in the DMN and positive correlations in the FPN and the aIC (Figure 4.21a). In stark contrast, activations related to the proximity along the identity dimension were negatively correlated with activity in the DMN, but positive correlations were virtually absent, even at a liberal statistical threshold (Figure 4.21b). To statistically confirm this pattern, I computed a directed contrast to identify areas that showed a stronger correlation with the impact of proximity along the emotion than the identity dimension. This contrast revealed a significant difference in the bilateral aIC and the right MFG (Figure 4.21c). Because I had a special interest in the FPN, due to its importance in representation learning, I repeated this analysis but constrained the FWE correction to this network, using a mask of the FPN (Yeo et al., 2011) for SVC. This analysis indicated additional significant clusters in the left MFG and IPS (Figure 4.21d).

Note that the close resemblance of the activation patterns in GLM1 and the impact of proximity along the emotional dimension is not surprising, since subjects learned relatively quickly in this study and as a consequence, ratings were primarily dependent on the proximity on the emotion dimension for most of the time. Still, this fact can not explain the discrepancy between the dimensions, namely why negative correlations with the DMN persisted along both dimensions, while positive correlations appeared exclusively along the emotional dimension. Due to the correlational structure of this analysis, it is hard to draw definite conclusions, but this result is well in line with the role of the FPN and the aIC in representation learning and could indicate that those regions encode a reduced representation of stimuli, that is purely based on emotional expression.

Multivariate analysis. To further investigate, whether the results of GLM2 are based on low-dimensional representations in the FPN, while circumventing the limitations of correlational univariate approaches, I used a multivariate approach. One possible explanation for the discrepancy of correlations with the impact of proximity along the two dimension would be that the FPN and the aIC encode one-dimensional representations of the stimulus space. To probe this, I constructed two different model RDMs per subject. Those RDMs depended either on the emotionality or the identity of faces and for each RDM, only one dimension was considered. When coding the stimuli with respect to the emotionality (displayed in Figure 4.22a) and computing the difference between those values between all stimuli, one can obtain the corresponding RDM (Figure 4.22b). The same procedure was applied when coding stimuli according to their identity (Figure 4.22c-d). I constructed two RDMs for each subject, a step that was necessary because I used individually fitted perceptual spaces instead of a canonical space.

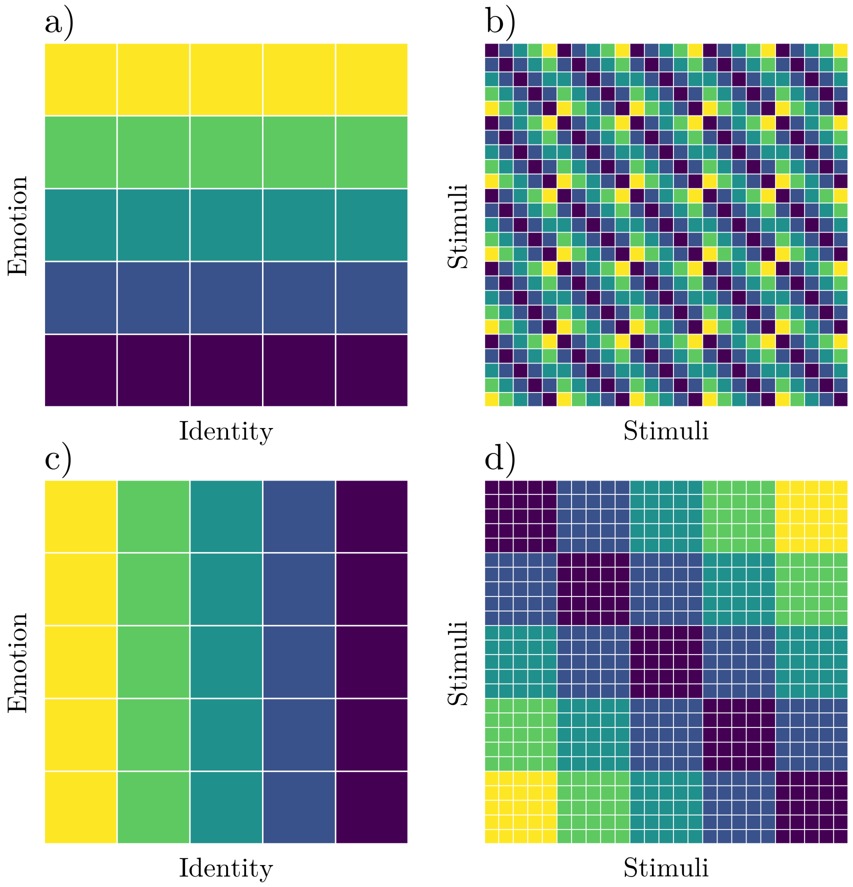


Figure 4.22: Depiction of model RDMs. Model RDMs depend on the value of stimuli on one dimension only. **a)** Schematic of stimulus space, when only considering the value on the emotion dimension. **b)** The resulting model RDM depends on the absolute values of differences between these values. **d)** When only considering the value on the identity dimension, **c)** the resulting model RDM looks like this. Note that the dimensionality of stimulus spaces in **a)** and **c)** is $5x5$ while the dimensionality of the model RDMs is $25x25$ as it is based on all possible combinations of stimuli. The model RDMs result when traversing the stimulus space in *column* → *row* order. The depicted spaces and RDMs are based on optimal grid. In practice I computed them based on individually fitted perceptual spaces.

Static representations. In a first step, I averaged beta images across all repetitions of a stimulus and used a searchlight analysis to compute correlation maps[21] for each subject, RDM and condition. I computed a second level ANOVA and specified a directed t-contrast to identify brain areas in which the representation of stimuli dependent more strongly on emotion than identity. This was the case *only* in the bilateral MFG, which is the anterior part of the FPN. The differential activations are depicted in Figure 4.23. This result supports the aforementioned interpretation of the results of GLM2 and (at least one aspect

[21] A correlation map is a brain image, where each voxel contains the Spearman correlation between the model RDM and the neural RDM computed from the beta images.

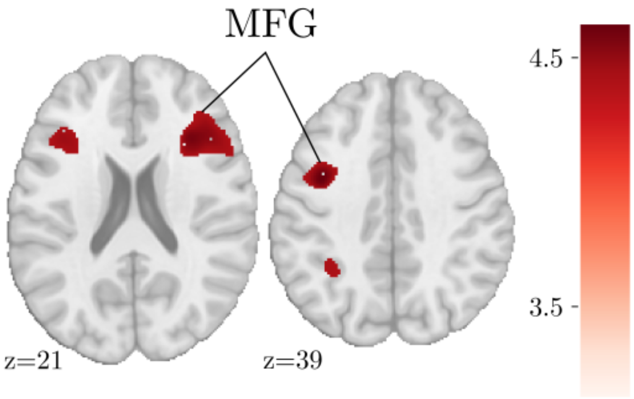


Figure 4.23: Static representations. Representations in the bilateral MFG are more strongly dependent on emotional differences than on differences in the facial identity.

of) the role of the FPN in stimulus generalization. However, due to the averaging over all presentations of stimuli, this analysis does not allow to draw conclusions about the temporal component.

Dynamic representations. If the dimensionality of representations in the FPN is relevant for the behavioral effects, one would expect that the representation of stimuli in the FPN changes over time. In particular, subjects did learn that the position of stimuli on the identity dimension was relevant, albeit slower and less pronounced than on the emotion dimension. This behavior makes sense in the context of the Bayesian model, assuming that subjects have a strong prior with respect to the irrelevance of the identity dimension. Still, this belief update needs to be encoded in brain activity and representations, and if the FPN plays the role of providing adequately dimensional stimulus representations, one would expect a shift from a representation that only depends on emotion to one that comprises both stimulus dimensions.

To test this hypothesis, I ran an additional searchlight analysis. Unlike in the previous one, I did not use beta images that were averaged over the whole experiment, but instead averaged them over all presentations of stimuli *within* a run[22]. This analysis resulted in 16 correlation maps per subject, one per run, model RDM and condition. To investigate the temporal dynamics of representations, I first entered those maps into a second level ANOVA and specified a generic F-contrast with one row per run and model RDM, while averaging over conditions. This contrast was meant to identify brain areas in which the representation of stimuli depends on at least one dimension in at least one run, independently of the condition.

This contrast revealed a single significant cluster in the left MFG. Two more significant clusters in the right MFG and the left IPS emerged when using SVC in a mask of the FPN. Those clusters are depicted in the left column of Figure 4.24. To further investigate the

[22] One fMRI run consists of 5 microblocks, i.e. 5 presentations of each stimulus.

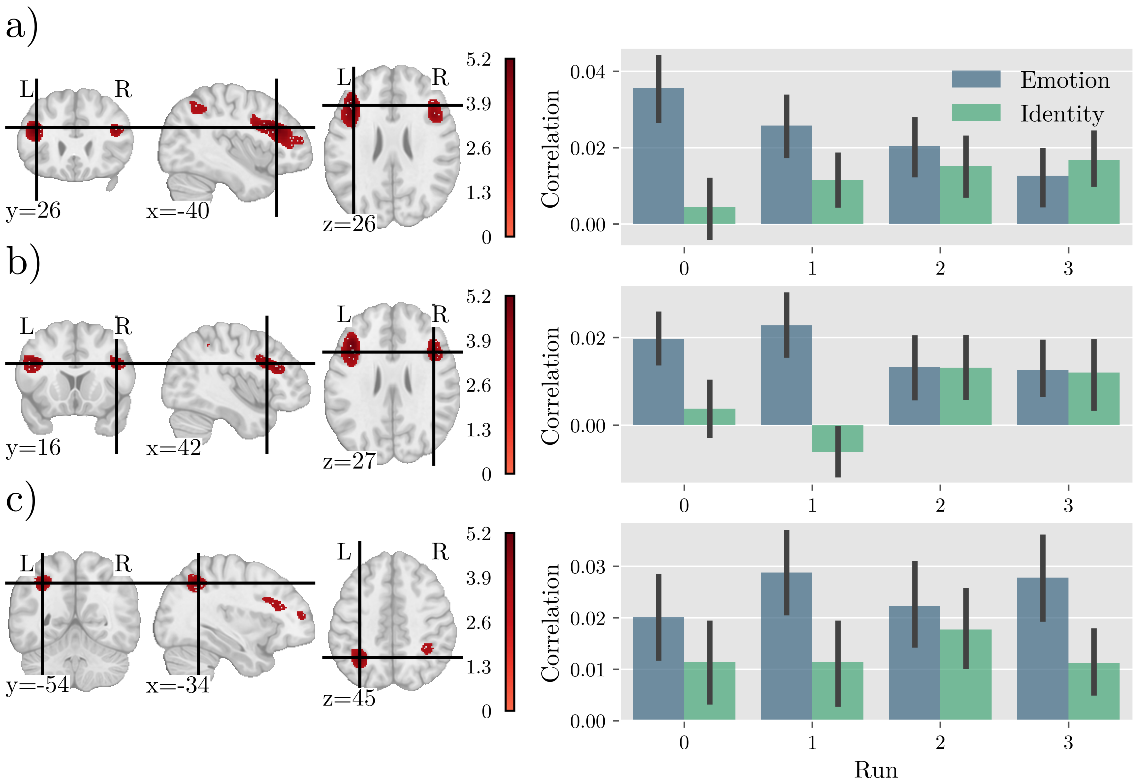


Figure 4.24: Dynamic representations. I investigated the dynamics in the dependence of neural representations on stimulus dimension in three RSA derived brain regions by computing the correlation between neural RDMs and the different model RDMs for each run. **a)** Representations in a peak voxel in the left MFG were initially dependent on emotion, but not on identity. Over the blocks, there is an almost linear increase in the dependence on identity and a slight decrease in the dependence on emotion. **b)** Representations in the right MFG showed a similar, but more noisy pattern. **c)** In contrast, representations in the left IPS were initially dependent on both stimulus dimensions and fairly static over time.

temporal dynamics of representations, I extracted the Spearman correlations between the model RDM and the neural RDM for each run, condition and subject. When averaging those correlation over conditions[23], correlations in the left MFG showed a pattern of initial correlations with the emotion RDM that decreased over time, while correlations with the identity RDM were not present initially but emerged over time almost linearly (Figure 4.24a, right column). A similar pattern was observed in the right MFG (Figure 4.24b, right column). In the left IPS, correlations were more static and above zero for both model RDM (Figure 4.24c, right column).

To test those patterns more formally, I modeled the time course of correlations with a Bayesian hierarchical linear regression[24]. In this model, correlations were modeled as a linear function of time, that is the Spearman correlation between the model RDM m and

[23] This is equivalent to the F-contrast that was used to identify the clusters.

[24] Note that formally the assumption of zero mean Gaussian residuals is violated, since correlations are bounded in the range $[-1, 1]$. However, the observed values were not close to the boundaries. Accordingly, this is unlikely to be a problem and due to the familiarity, simplicity and ease of interpretation I decided to use a linear regression model.

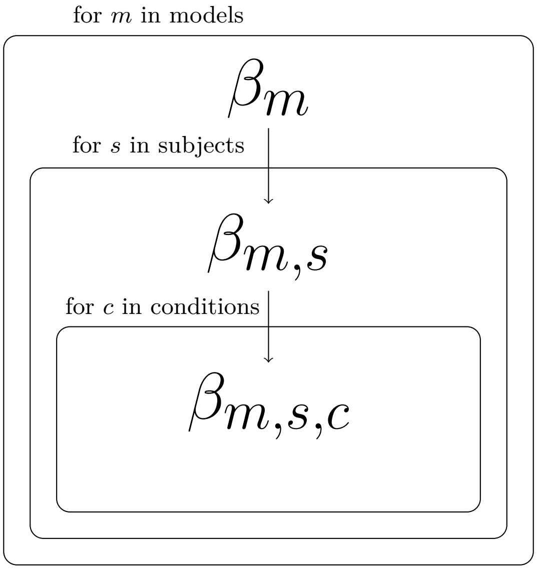


Figure 4.25: Hierarchical structure of regression model. Parameters β_0 and β_1 were nested in three levels. Subject level coefficients in group level distributions and condition level coefficients in subject level distributions. Models were treated as independent.

the neural RDMs at time t, for subject s in condition c is given by

$$\begin{aligned} \rho_{s,t,c,m} &= \beta 0_{s,c,m} + \beta 1_{s,c,m} \cdot t + \epsilon \\ \epsilon_{s,t,c,m} &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{4.26}$$

I used $Normal(0, 0.1)$ priors on all group level β coefficients, which I consider weakly informative given the range of observed correlations, and standard normal priors on hierarchical variance parameters. To account for the structure of the data, I included three levels of hierarchy for all β parameters: subjects in group and conditions in subjects. This structure is depicted in Figure 4.25. This analysis confirmed that representations in the left MFG were exclusively dependent on emotion initially (Figure 4.26a, first row). With time, the dependence on emotion went down, but the dependence on identity increased[25] (Figure 4.26a, second row). A similar pattern was observed in the right MFG, albeit the posteriors probabilities are a bit more uncertain about the effect (Figure 4.26b). In contrast, posteriors for the left IPS indicate that while the representation depends more strongly on emotion, it is more static than in the MFG and depends on both dimensions from the beginning (Figure 4.26c). These dynamics of representations in the MFG are well in line with the role of the FPN in representation learning. The Bayesian model gives a rational explanation for the observed change in dependence on the different dimensions of frontal representations. Still, if these changes relate to the behavioral effects and their explanation via the Bayesian model, I expected to see a correspondence between neural effects and behavioral effects on a subject level. In particular, there should be a covariance between how much representa-

[25] While the 90% posterior highest density interval (HDI) includes zero for $\beta 1_{identity}$, the bulk of the probability mass indicates an increase. Bayesian analysis is not supposed to be used as dichotomous decision criterion.

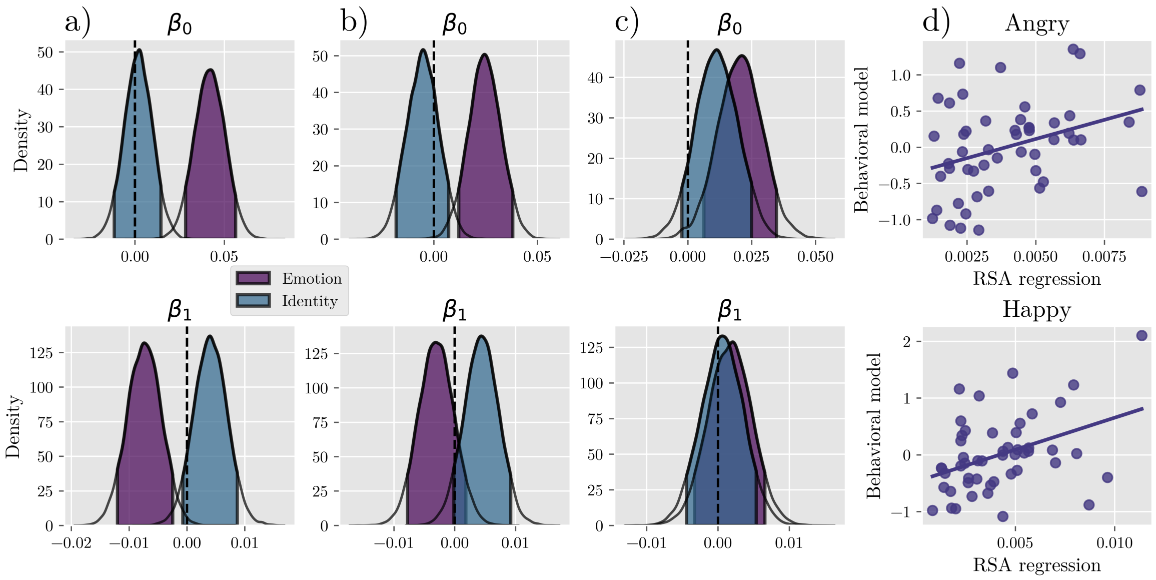


Figure 4.26: Posterior distributions of regression coefficients. **a)** In the left MFG, the representation is initially dependent on emotion ($p(\beta0_{emotion} > 0) = 1$), but not on identity ($p(\beta0_{identity} > 0) = 0.59$). With time, the dependence on emotion decreases ($p(\beta1_{emotion} < 0) = 1$) and the dependence on identity increases ($p(\beta1_{identity} > 0) = 0.92$). **b)** Posterior distribution for the left MFG indicate the same direction of effects, but are more uncertain about the magnitude of the effects ($p(\beta0_{emotion} > 0) = 1$, $p(\beta0_{identity} > 0) = 0.75$, $p(\beta1_{emotion} < 0) = 0.85$, $p(\beta1_{identity} > 0) = 0.93$). **c)** Representations in the left IPS depend on emotion ($p(\beta0_{emotion} > 0) = 0.99$) and identity ($p(\beta0_{identity} > 0) = 0.91$) from the beginning and are less dynamic than those in the MFG ($p(\beta1_{emotion} > 0) = 0.71$, $p(\beta1_{identity} > 0) = 0.58$). **d)** The increase of the dependence on identity of representations in the left MFG is significantly correlated with the increase of the influence of proximity along the identity dimension in behavior in both conditions.

tions increasingly depend on identity in the MFG and how much the proximity to the CS+ along the identity dimension increasingly influences shock expectation ratings. To test this, I extracted the corresponding parameters from the Bayesian regression of RSA correlations and the best fitting behavioral model. The corresponding parameters are $\beta1_{identity}$ from the regression model[26] and $\beta1_{identity}$ from the behavioral model, which I extracted from the model fits for individual subjects. I then correlated those two measures to investigate whether there is a relationship between the two. As depicted in Figure 4.26d, the correlation was significant for both the angry ($r = 0.34$, $p = 0.015$) and the happy condition ($r = 0.4$, $p = 0.004$). This indicates that the dimensionality of representations in the MFG is related to the width of generalization along stimulus dimensions.

4.3.5 Interim discussion

This study provides two important lines of evidence. First, I was able to fully replicate the behavioral results from the first study in an independent sample. While this is not exactly surprising as the study design was virtually the same, other than the measurement of fMRI

[26] Note that because the behavioral models were fitted to conditions separately, I also extracted condition specific parameters from the RSA regression.

data, it demonstrates the robustness of the results. This seems especially important in the light of the replication crisis in psychology (Maxwell et al., 2015; Open Science Collaboration, 2015). In addition, it gives reason to believe that the model-based investigation of neural effects in the second study is based on a solid foundation.

Second, my neural results seem important with respect to the question of the role of the FPN in stimulus generalization and how this relates to its role in representation learning. Famously, the FPN is thought to be involved in the discovery or encoding of low-dimensional representations (Niv, 2019; Tomov et al., 2018; Woolgar et al., 2011). So far this role has been ignored in the context of stimulus and more specifically fear generalization (Webler et al., 2021). My results indicate that the FPN plays a similar role when it comes to the learning and encoding of an appropriately abstracted stimulus space in the context of stimulus generalization. These results further question whether stimulus generalization and representation learning, i.e. a relevant concept for generalization in RL are really as distinct as they are often treated. In addition, the proposed Bayesian model provides a rational explanation for behavioral and neural effects. A correlation between neural and behavioral effects provides further support for the suggested role of the FPN and the proposed model and implies behavioral relevance of representations in the MFG. I will expand on those findings in the context of the big picture of this thesis in the general discussion.

4.4 Appetitive generalization in an online sample

The third study had the purpose to test whether the results from aversive conditioning translate to appetitive conditioning as well. This is an important aspect as the proposed mechanism of generalization is supposed to be a general one. If, as I propose, fear generalization is not a fundamentally different process from other forms of generalization, then a similar behavioral pattern should emerge for generalization of learned appetitive associations.

4.4.1 Methods

In contrast to the first two studies, this study was an online study. Apart from this and the change from aversive to appetitive conditioning, the procedure was very similar to the other studies. Some obvious differences were unavoidable due to the change of valence in the UCS and the context of an online study.

Sample description. Subjects were acquired using the services of Prolific (https://www.prolific.co/). The experiment itself was hosted on Pavlovia (https://pavlovia.org/). Because this platform is incompatible with `MATLAB` and the `PsychToolbox`, I programmed the task in `Javascript` using the `jsPsych` library (https://www.jspsych.org/). I imposed some constraints on eligibility for subjects. Only subjects that reported German as their first language, had completed at least 50 tasks on Prolific and had an approval rate of at least 95% were allowed to take part. Subjects that chose to participate in the study were redirected to Pavlovia for the experiment. After a successful completion, they were redirected back to Prolific. The payment was handled by Prolific and participants received 12£/h for their participation.

Because of the difficulty to conduct a within-subject study on Pavlovia, subjects were instead randomly assigned to the happy and angry condition. In total, I collected data from 104 subjects. The random allocation resulted in a somewhat unbalanced distribution of 48 subjects in the angry and 56 subjects in the happy condition. Of those 104 participants, 44 identified as female (60 as male). The mean age was 30.9 years with a range of 18 to 59.

Experimental procedure. After being forwarded to Pavlovia, subjects were informed about the study and gave informed consent for their participation. After that, subjects took part in the same quadruplet task as in the other two studies. I used the same sequences and stimuli for this.

For the main generalization task, the stimuli, sequences of stimuli, stimulus durations and ITIs were the same as in the first study. Instead of an electric shock as UCS, I used money as reinforcement. Due to its cultural significance, monetary reward acts as a powerful secondary appetitive reinforcer with a similar neural profile to primary reinforcements (Knutson et al., 2001). In rewarded trials, after the offset of the CS+, I presented the CS+ face in conjunction with a dollar bill for 2 seconds. Subjects were informed that each reinforced trial would result in a potential monetary gain of 0.25£ on top of the baseline compensation. Since online studies have an additional concern with attention, I used the oddball trials to ensure that subjects were vigilant. Subjects were informed that they needed to react to the oddball trial with pressing the space bar and that the bonus they received was directly dependent on their reactions. Subjects that reacted quickly enough to all oddball trials received the whole bonus. If they reacted to half the oddball trials, they received half and so on. Since there were 16 reinforced trials and thus a maximum bonus of $16*0.25 = 4$£ and 4 oddball trials, every successful reaction resulted in a 1£ bonus.

The expectation ratings differed in the question relative to the aversive studies. Due to the different reinforcement, I collected reward expectation ratings instead of shock expectation ratings. The procedure was the same as in study 1, but subjects had to rate faces with respect to the question „How likely is this face going to reward you with money?“.

Because I had established a constant perceptual space before in the first two studies and in order to keep data quality high, I omitted the second quadruplet task. As a consequence, I also did not compare different models for the perceptual spaces and instead fitted a single model. The implied positions of stimuli from these models were used in the behavioral models as described in the previous studies.

4.4.2 Model predictions

The predictions of the Bayesian model are depending on the valence of the emotional expression and the reinforcement. Accordingly, the predictions for the first and second study[27]

[27] See subsection 4.2.2.

do not fit for the third study since the valence of the reinforcement is changed. According to the preparedness hypotheses (Seligman, 1970), the valence of the expected outcome and of the predictor need to be congruent. Still, emotionality is the more salient dimension and the one that is *a priori* more likely to be informative with respect to an appetitive outcome. For the prior distributions, this implies that the priors on λ, i.e. the strength of exponential decay stayed the same as for aversive conditioning. The same is not true for μ, the midpoint of associative maps. A simple, but reasonable way to include the change in valence in the model predictions is to flip the priors on $\mu_{emotion}$ so that happy faces are more likely than neutral faces to predict a reward while angry faces are less likely:

$$\mu_{angry} \sim Beta(6, 1) \tag{4.27}$$

$$\mu_{happy} \sim Beta(1, 6) \tag{4.28}$$

The resulting predictions are exactly the same as in the first two studies, except that they are for the other emotion respectively[28].

4.4.3 Results

As a check for data quality, I looked at the distribution of the number of successful oddball hits. This is depicted in Figure 4.27 and indicates ongoing attention in almost all subjects.

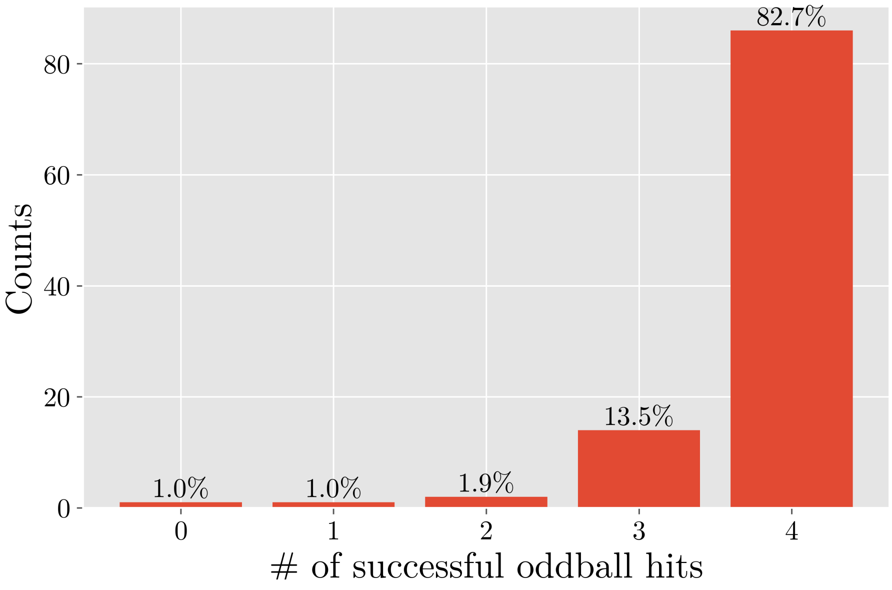


Figure 4.27: Distribution of the number of successful oddball hits. The vast majority of subjects reacted to all oddball trials, with only two subjects reacting to less than half of the trials.

Perceptual space. Perceptual spaces in the online study (Figure 4.28) are very similar to those in the other two studies. As in those, the fitted group level positions are aligned fairly

[28] Refer to Figure 4.9 and mentally exchange the rows with each other.

closely with the grid that indicates the expected positions. Single subject positions were included in the behavioral models to account for individual differences in the perceptual spaces. Note that these spaces are based on a single iteration of the quadruplet task, while positions in the first two studies were based on two iterations.

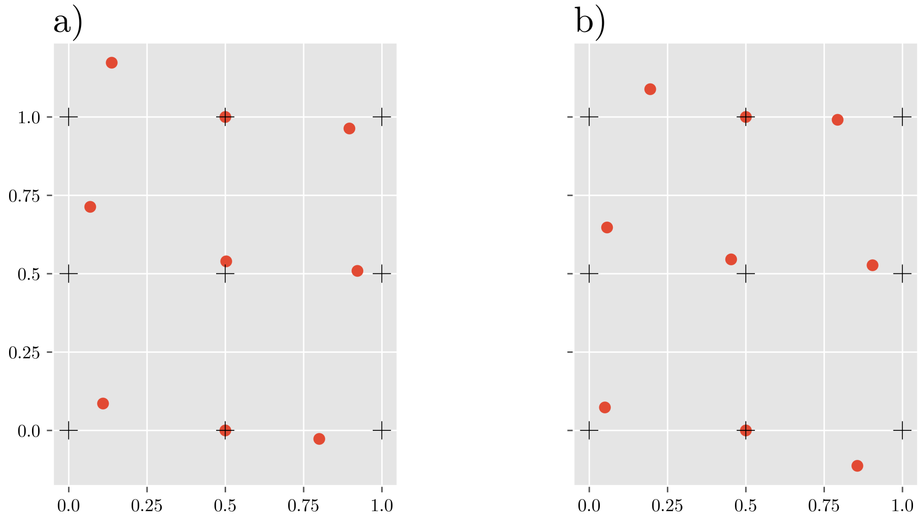


Figure 4.28: Perceptual space of the online study. As in the other studies, those are fairly closely aligned with the grid that indicates the expected positions. This indicates that data quality in the online study is comparable to data that was collected in the lab.

Behavioral models. Since the only difference in the rating data was the valence of the reinforcement and consequently also the expected outcome (i.e. reward expectation ratings instead of shock expectation ratings) and because I'm suggesting a general process of generalization that is independent of the valence, I fitted the same models to the data as in the other two studies. Given my assumption of generalization as a valence-independent process, I expected a similar pattern of results in the model comparison and parameter inspections. Results from model comparison are shown in Table 4.5.

The emerging pattern suggests, that models that include proximity to the CS+ separately for each dimension explained the data better than models with a compound or a single measure for proximity. Additionally, models that included non-linear temporal dynamics for the strength of updating and the impact of proximity fitted the data better than models that assumed linear dynamics and a static impact of proximity. The best fitting model was model 3b, which is the same as in the first two studies. While the validity of the ELPD is questionable for a subset of those models[29], the overall pattern closely follows the studies with aversive conditioning, which is in line with my assumption of a general process of generalization.

A striking difference in mean ratings when comparing them to the other two studies is the change in the initial impact of emotionality. As suggested by the preparedness hypothesis (Seligman, 1970), I observed a flipped pattern, where happy faces were initially rated as

[29] Indicated with an asterisk in Table 4.5.

	Condition	
Model	**Angry**	**Happy**
1a	-11555.03	-13207.22
1b	-11491.83	-13022.28
1c	-11796.53	-13157.57
2a	-11296.61	-12812.19*
2b	-11093.63	-12596.48*
2c	-11532.55	-12764.98*
3a	-11154.05	-12802.83
3b	**-11077.13**	**-12558.48**
3c	-11409.12	-12703.5

Table 4.5: Model comparison for behavioral models in online study. Like in the other two studies, model 3b showed the best fit. Importantly, this is despite the fact that the value of reinforcement is different. Note that I observed divergent transitions in the fitting of some models. Those cases are marked with an asterisk.

more likely to be rewarding than neutral faces. In contrast, angry faces were rated as less likely to lead to a reward. Apart from that, mean ratings were very similar to those in the other studies with an initial impact of emotionality that was followed by a gradual decrease, while the impact of proximity became more relevant with time. This effect again was stronger along the emotional dimension, indicating partial dimensionality reduction. Additionally, the initial effect of emotionality was stronger in the congruent, i.e. the happy condition. Recall that this effect was stronger in the angry condition in aversive conditioning. Posterior predictive checks indicated that ratings could be explained well by the best-fitting model.

To corroborate those initial findings and strengthen the overlap with the other two studies, I again investigated the posterior distributions on parameters of the best-fitting model. Those can be seen in Figure 4.30. Besides the fact that the impact of emotion was flipped with respect to the different emotions when compared to the aversive conditioning studies, overall a very similar pattern of posterior distributions with the same model emerged in the present study. A detailed description of the parameter estimates is given in the legend of Figure 4.30. Again, the invariance to the valence of the reinforcements indicates a general process and is in line with the assumption that fear generalization is not fundamentally different from other forms of generalization, e.g. in the context of appetitive generalization.

4.4.4 Interim discussion

The present study investigated the scope of the proposed mechanism by testing it in a new context. In contrast to the widely held assumption that fear generalization is a special case of generalization, among other reasons for the supposed dependence on the amygdala, I showed that results from a study design that only differed in the valence of the reinforcements from the aversive design in the first two studies led to very similar results. In particular, the dynamics of ratings, the dependence on the emotionality of faces and the increased relevance

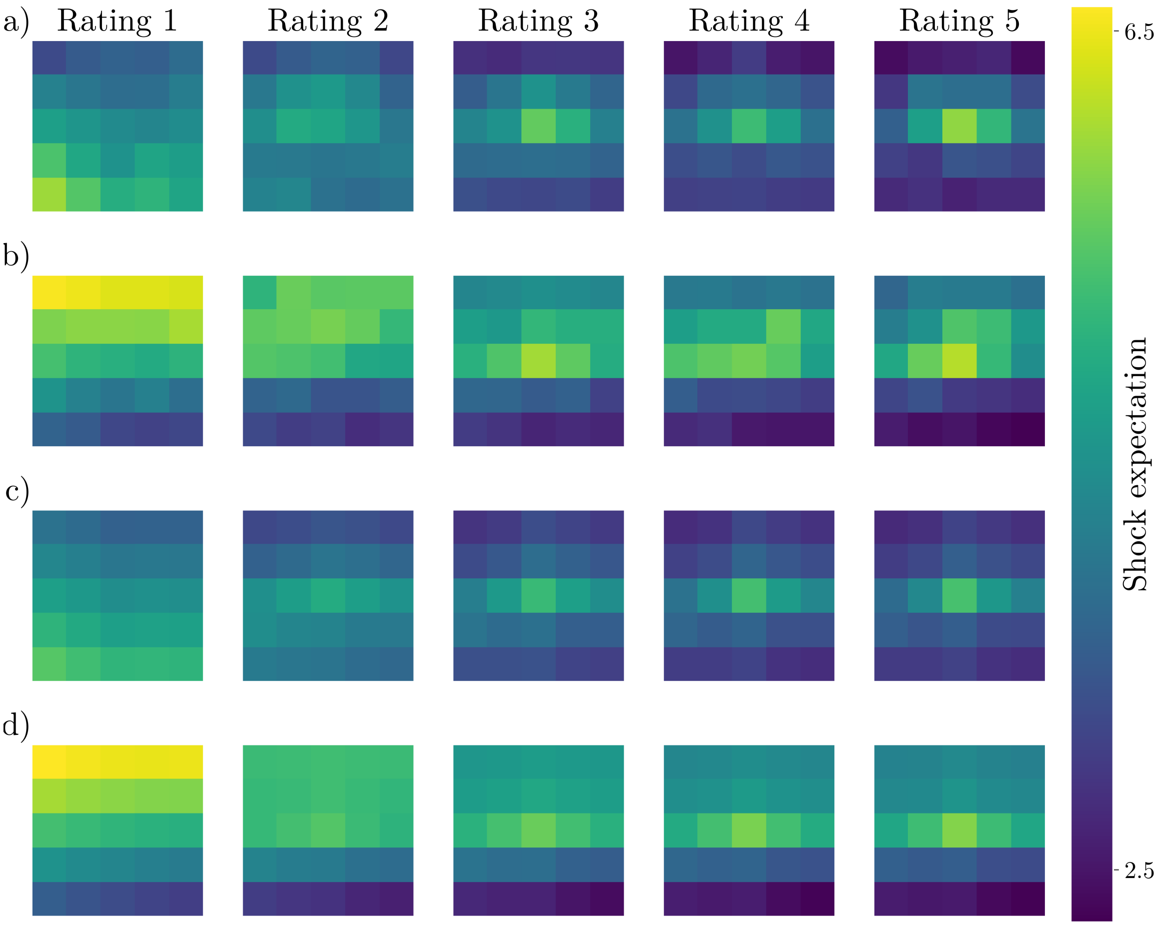


Figure 4.29: Mean ratings and posterior predictive checks for the online study. a) Mean ratings for the angry condition show an opposite and weaker effect of emotionality in the first rating, when compared to the other studies. Other than that, the structure and dynamics of the ratings are very similar. **b)** Likewise, ratings in the happy condition show an opposite and stronger effect of emotionality and are otherwise comparable to those of the first two studies. **c)-d)** Posterior predictive checks indicate that ratings could be explained well by the winning model in both the **c)** angry and **d)** happy condition.

of the emotion dimension showed a striking invariance to the valence of reinforcements. Solely the direction of the initial impact was reversed, which is in line with the preparedness hypothesis and emphasizes the importance of prior knowledge over and above the salience of dimensions. This finding in combination with the dynamics of belief updating strongly suggest a Bayesian mechanisms that is independent of the kind of reinforcement.

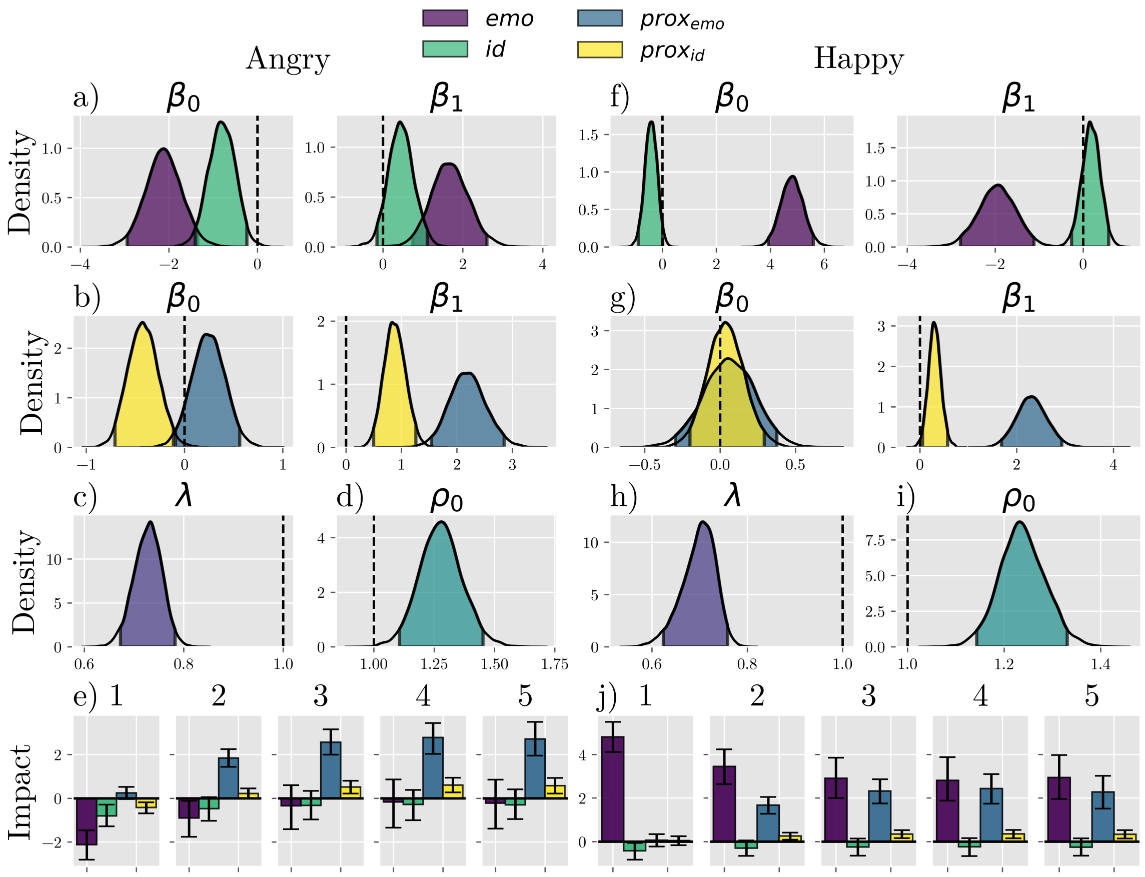


Figure 4.30: Posterior distributions of the model parameters for the online study. **a)** In the angry condition, ratings were initially dependent on the emotionality of faces, with a slight unexpected effect of identity. Ongoing conditioning reversed the effect of emotionality. **b)** Unexpectedly, proximity to the CS+ in both axis had small initial impact on ratings. Since nothing had been learned for the first ratings and this effect is not visible in mean ratings, I attribute this to inflexibility in the model. Impact on both dimensions increased, with a stronger effect along the emotion dimension. **c)** Updating from rating to rating slowed down and **d)** the impact of proximity became narrower with time. **e)** As a result, ratings were initially driven mostly by emotion, but towards the end most strongly dependent on proximity along the emotion dimension. **f)** Similarly, in the happy condition, ratings were initially driven by emotionality and this effect decreased with time. The direction of the effect is opposite to the angry condition and the previous studies. **g)** Proximity to the CS+ was not relevant in the beginning but increased, especially along the emotion dimension. **h)** The belief state became more stable over time, but **i)** the impact of proximity became narrower. **j)** Over the ratings, the stronger impact of emotionality, compared to the angry condition stayed more relevant over time, but like in the angry condition, proximity along the emotion dimension became a strong driver of ratings.

5 General discussion

The current thesis aimed to contribute of our understanding of human generalization. How do people deal with an everchanging world and are able to apply knowledge that they have acquired in certain situations and about certain stimuli to other situations and stimuli? After having reported the theoretical, methodological and empirical work that I conducted to this end, I now want to summarize the contributions of my thesis. In particular I want to recapitulate on the specific questions I was trying to answer, briefly summarize and interpret the results I have obtained and discuss their implication.

5.1 Recapitulation of the research question

As outlined in the introduction, I have identified three main subfields in which generalization is discussed in different contexts: (1) Stimulus generalization, (2) generalization in RL and (3) inductive reasoning. I have also given my reasons to believe that there is the possibility of a common mechanism that would describe phenomena in all fields. That is not to say, that these are the exact same thing, but it seems reasonable to assume that they rely at least partly on the same mechanism. In order to proceed in that direction, I derived my main research question that I wanted to answer in the course of this thesis: „How can results and explanations from all subfields be integrated into a single model and what is the neural implementation?“. It seems important to note that a fully satisfying answer to this question is way beyond the scope of this thesis and will require a lot of further work by people smarter than me.

5.2 Summary of theoretical and empirical results

Theoretical results

A formal attempt to propose a common mechanism and thereby unify those fields needs to be able to explain a bulk of results from all fields and to comprise mechanisms that have been shown to be a good match for empirical data. It therefore needs to be Bayesian in nature to account for the vast and fruitful literature on inductive reasoning (J. C. Lee, Lovibond, Hayes, & Navarro, 2019; Shepard, 1987; Tenenbaum & Griffiths, 2001a). It needs to include some form of abstraction, e.g. dimensionality reduction, as this is a primary mechanism in RL (Badre et al., 2021; Niv, 2019; Tomov et al., 2018). To provide an integration, it should ideally provide a rational explanation for this abstraction in the context of a Bayesian model. Lastly, it needs to be applicable to probabilistic associative learning to account for the literature on stimulus generalization in general (Ghirlanda & Enquist, 2003) and fear generalization in particular (Dymond et al., 2015) and include a treatment of perceptual inaccuracy (Schechtman et al., 2010; Zaman et al., 2019).

To my knowledge, such an attempt has not been made before, since previous approaches aimed to integrate either inductive reasoning and stimulus generalization (J. C. Lee, Lovibond, Hayes, & Navarro, 2019) or generalization in instrumental learning (i.e. RL) and stimulus generalization by including an instrumental generalization phase (Norbury et al., 2018; van Meurs et al., 2014).

Luckily, some theoretical parallels between the subfields that I believe have been overlooked so far, can be used to facilitate this process. An obvious parallel between inductive reasoning models that are based on the idea of a psychological space and consequential regions (Shepard, 1987; Tenenbaum & Griffiths, 2001a) on the one side and RL, that is concerned with state spaces on the other side, is the idea of a spatial configuration. To go beyond that, the learning of an adequate state space representation and the learning of a psychological space seem very similar to me. I assume that work that is concerned with either of those two processes (e.g. Austerweil et al., 2019; Collins & Frank, 2013) will be helpful in the future to extend and refine approaches that try to unify them.

The model that I have developed in this thesis is heavily influenced by the work of Tenenbaum and Griffiths (2001a), Soto et al. (2014), and J. C. Lee, Lovibond, Hayes, and Navarro (2019). The most important addition that I have made to the existing models is a rational treatment of partial or full dimensionality reduction. The implementation of this is based on a rescaling of perceptual dimensions according to their currently assumed importance. Importantly, this provides a bridge between inductive reasoning models, parts of the literature on generalization in RL (Leong et al., 2017; Niv et al., 2015; Niv, 2019; Tomov et al., 2018) and the role of low-dimensional neural representations as a neural mechanism (Badre et al., 2021; Fusi et al., 2016). One important aspect that the full Bayesian treatment adds is the omission of the distinction between learning *about* the relevant dimensions and learning about the values *on* those dimensions. In addition, the model is able to account for prior knowledge and provides a way to include it when making predictions.

To account for perceptual inaccuracy, the model includes a mechanism that implements misidentification of stimuli that is based on the idea of a perceptual confusion matrix and Gaussian perceptual noise.

Empirical results

Behavioral results. Even just looking at the raw data from the experiments, it is clear that the data closely followed the predictions of the proposed model in all three studies. To go beyond a purely qualitative assessment, I fitted a series of models to the data. Those models incrementally include hypotheses that I derived from the full Bayesian predictions. Empirically, the model that includes the most hypotheses fits the data best across all three studies and all conditions. I interpret this as a strong indication that the model is a good match for the data. Importantly, I provide evidence for partial dimensionality reduction in all three studies. This effect was previously understudied in the context of stimulus generalization. While this is not proof of the validity of the model, it is a strong indication that the actual process of generalization shares some important properties with the mechanism of the model.

Neural results. Since previous neural results in stimulus generalization have been overwhelmingly descriptive (Webler et al., 2021), I intended to investigate whether those findings can be integrated with the literature on neural mechanisms of abstraction in RL. While there are some inconsistencies in this literature, a common theme that has emerged is the involvement of the FPN and the aIC in the process of either the discovery (Niv, 2019) or the encoding (Loose et al., 2017; Tomov et al., 2018; Woolgar et al., 2011) of relevant dimensions. Correlational results that I reported were equivalent to the typical neural generalization tunings in fear generalization (see Webler et al., 2021 for a meta-analysis). To investigate the role of those areas further, I conducted another correlational analysis which revealed a distinction between stimulus dimensions, with activity in the aIC and the FPN being correlated with generalization along the emotional but not the identity dimension, whereas the DMN did not show such a distinction. Seeing this as a first corroboration of a similar role of the aIC and the FPN in the process of stimulus generalization as in the context of RL, I used RSA to investigate this further. This analysis showed that the bilateral MFG[1] more strongly encoded the emotional dimension than the identity dimension of stimuli. Taking the temporal dynamics of those representations into account, the left MFG, but to a certain extent also the right MFG, showed the expected pattern in which representations were initially more strongly correlated with the emotional dimension and then drifted towards a two-dimensional representation. It turned out that individually, the increase of representational dependency on the identity dimension in the left MFG was correlated with decreased behavioral generalization along this dimension. This leaves the dimensionality of representations in the MFG as prime candidate for the neural implementation of the dimensionality reduction effect in behavior.

5.3 Implications for the field

5.3.1 Generalization as a perceptual process

The role of perceptual accuracy in stimulus generalization is still an ongoing debate (Laufer & Paz, 2012; Laufer et al., 2016; Schechtman et al., 2010; Struyf et al., 2015, 2017; Zaman et al., 2019). Although a number of studies have empirically contradicted the assumption of generalization as a purely perceptual process (Kampermann et al., 2021; Onat & Büchel, 2015; Tuominen et al., 2019; Zaman et al., 2021), the fact that the idea of generalization as a perceptual process has been around for at least 65 years is a testament to its resilience. As a consequence, it seems reasonable to probe this idea with every study on generalization.

[1] As a reminder: The MFG is an important hub of the FPN.

The idea of a purely perceptual process rejects an active Bayesian process[2] and thereby a formal mechanism to incorporate prior knowledge. As a consequence, popular models of fear generalization provide no way to implement prior knowledge (Lissek, 2012; Lissek et al., 2014). A second consequence is that generalization cannot differ between dimensions, as long as they are matched with respect to their perceptual similarity. My results clearly show that both of those assumptions are wrong. There is a very clear effect of prior knowledge along the emotional dimension. In addition, the dimensionality reduction effect prevails until the end of the conditioning phase. Those results add to the decisive evidence that generalization is not merely a perceptual process.

5.3.2 Fear generalization as a distinct phenomenon

Especially in the last two decades, the idea of fear generalization as a distinct phenomenon has been gaining traction. This is, at least implicitly, apparent from multiple observations. Most of the research on stimulus generalization has been conducted in the context of aversive conditioning (Dymond et al., 2015), there is a lot of debate about the overgeneralization of fear in the context of anxiety disorders (Berg et al., 2020; Greenberg et al., 2013b; Lissek, 2012) and a neural model for fear generalization has been proposed, that is centered on the amygdala (Lissek et al., 2014; Webler et al., 2021). In addition, changes in perceptual tuning happen exclusively in the context of negative reinforcement (Laufer & Paz, 2012; Laufer et al., 2016; Schechtman et al., 2010).

To directly address this issue, my work contrasted generalization in the context of aversive and appetitive conditioning. The results showed a remarkable similarity of outcome expectation ratings between the two types of reinforcement and further question the idea of fear generalization being different from other forms of generalization. Going beyond that, neural evidence from my fMRI study suggests that fear generalization and generalization in RL rely on similar neural mechanisms, namely the discovery and encoding of relevant dimensions. I will advance on this point in the next paragraphs.

5.3.3 A common mechanism of generalization

The fundamental question of my thesis was whether there is a common mechanism in human generalization that spans different applications. While I do not mean to imply that this is a question that I was sufficiently able to answer, I believe that the presented conceptual, theoretical and empirical evidence supports the notion that this is indeed the case.

Conceptually, all of those different applications need to be implemented in the brain somehow. I have argued that a consideration of research that is concerned with how the brain works in general can be helpful when identifying the neural mechanisms of stimulus generalization and generalization in RL (Badre et al., 2021; Bottini & Doeller, 2020; Fusi et al., 2016). The fact that low dimensionality in neural representations is applicable to all of those different applications is a strong argument in favor of a common mechanism.

[2] Other than those that implied in e.g. predictive coding.

Theoretically, I have derived commonalities between models of generalization in RL and inductive reasoning, both of which rely on a spatial model of the world and adequate abstractions. This observation motivated me to derive a Bayesian model of generalization that integrates ideas from RL and inductive reasoning and is applicable to research designs in stimulus generalization. This enabled me to propose a draft of how the suggested common mechanism could look like and to make predictions for a specific study design.

Empirically, I have shown that rating data from aversive and appetitive generalization studies closely followed the predictions of the model, including those derived from the dimensionality reduction approach in RL.

5.3.4 Neural mechanisms of generalization

FMRI studies in RL have emphasized the role of the FPN (Niv, 2019; Tomov et al., 2018) and the aIC (Tomov et al., 2018; Woolgar et al., 2011). In addition, relevant research has suggested a role of cognitive maps in the OFC (Niv, 2019; Schuck et al., 2016), HPC and PPC (Bottini & Doeller, 2020; Summerfield et al., 2020). Despite some contradictions with respect to its role[3], an involvement of the FPN seems to be the most consistent finding in this literature.

This interpretation of the role of the FPN is missing in the literature on fear generalization, instead the classical view of the interplay between brain networks has emerged in recent publications (Berg et al., 2020; Webler et al., 2021). This discrepancy between the fields seems to arise from multiple factors, among them a focus on biomarkers for anxiety disorders and the overwhelming use of one-dimensional stimuli in fear generalization. The latter omits the need to think about dimensional relevance. In fact, I am not aware of a single neuroimaging study on fear generalization that emphasized that aspect of generalization and therefore its neural implementation.

Starting from the assumption that a common behavioral mechanism most likely relies on common neural underpinnings, I probed the role of the FPN in fear generalization within a two-dimensional stimulus space. My results support the idea of a common mechanism and provide evidence for the behavioral relevance of the dimensionality of representations in the FPN, with a strong focus on the MFG.

These results can also help to solve the contradictions in the literature on the role of the FPN in RL. In contrast to the notion that the FPN is not involved in the encoding of the learned abstraction (Niv, 2019), my results suggest that it is.

Taken together, these results are in line with a common neural mechanism in stimulus generalization and representation learning and a consideration of other relevant literature suggest low-dimensional neural representations as a plausible candidate.

[3] Niv (2019) implied the FPN in the discovery of relevant dimensions, other authors (Tomov et al., 2018; Woolgar et al., 2011) propose that it encodes them.

5.4 Limitations

While I believe that there is value in the presented work, there are some limitations that should be considered.

First, the proposed model and the empirical studies deal with associative learning. There are some strong parallels with generalization in instrumental learning, as the dimensionality preference and the role of the FPN. Still it is not trivial to extend the model to instrumental learning and multiple possible actions per state. In addition, the model only accounts for a single latent cause[4] and a single outcome intensity. Any study design that violates those assumptions would need to extend the model, which in principle is possible.

Second, the hypotheses I derived from the model are fairly generic. I found evidence for them in the data, but the same set of hypotheses could have been derived from a different model that is not necessarily Bayesian. The principle of multiple realizability (Bechtel & Mundale, 1999) states that the same cognitive state can be realized by different neural mechanisms. As a consequence, the fact that data looks like model predictions does not imply that the data was generated *by* that mechanism. However, this is very general criticism of cognitive neuroscience and the Bayesian framework, that my model is based on, has been very successful in explaining a vast amount of phenomena. In addition, even if the brain does not perform Bayesian inference per se, Bayesian models are still a useful model for the actual process.

Third, the work on generalization in RL and coding principles of the brain is vast and a literature search is not easy since the work that is relevant does not necessarily cover generalization itself, but rather mechanisms that *enable* generalization. This is why a lot of relevant papers do not have the term *generalization* in their title. Because they are often not considered in the work that is directly concerned with generalization, it is hard to have a good overview of the complete literature. One thing I want to mention is that my consideration of research in RL has likely been biased towards those studies that focus on the FPN. This is in parts because it is most well-aligned with my own results and also because work on mechanisms that are likely useful for the understanding of generalization focus on other aspects. In particular, I believe the work on cognitive maps in different brain regions deserves more attention in following work (Bottini & Doeller, 2020; Schuck et al., 2016; Summerfield et al., 2020).

5.5 Future outlook

I hope that this work will prove to be useful in extending our understanding of generalization. In this last section of my thesis I want to discuss some possible directions for future work that I would deem interesting. Essentially, those are approaches that fill the gaps that I identified in the previous section.

On a computational level, I think the model could be extended to account for aspects

[4] See Soto et al. (2014).

that it does not account for yet. In particular, those are variations in outcome strength, multiple latent causes and a consideration of instrumental learning. The latter would imply a consideration of multiple actions per state which is not trivial. In contrast, the extension to multiple latent causes should not be too complicated as one can take inspiration from the elegant work by e.g. (Soto et al., 2014).

With respect to the neural mechanisms, I think future work should try to disentangle the role of cognitive maps and the FPN in generalization. To provide a small teaser, Summerfield et al. (2020) and Bottini and Doeller (2020) suggested an interplay between allo-centric cognitive maps in the HPC and ego-centric cognitive maps in the posterior parietal cortex. Likewise, the distinction between the encoding of the learned abstraction in either the OFC (Schuck et al., 2016) or the FPN (Woolgar et al., 2011) could be due to differences in task demands and potentially be resolved by a more detailed analysis of study designs.

Lastly, I think that research in all three subfields that I was concerned with would hugely profit from a consideration of work outside of their narrow focus to arrive at a more complete picture of generalization in the brain. I hope to have contributed towards such a more unifying approach with this thesis.

Bibliography

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. R. (2000). A Role for Somatosensory Cortices in the Visual Recognition of Emotion as Revealed by Three-Dimensional Lesion Mapping. *Journal of Neuroscience*, *20*(7), 2683–2690. https://doi.org/10.1523/JNEUROSCI.20-07-02683.2000

Aguirre, G. K., Mattar, M. G., & Magis-Weinberg, L. (2011). De Bruijn cycles for neural decoding. *NeuroImage*, *56*(3), 1293–1300. https://doi.org/10.1016/j.neuroimage.2011.02.005

Ahmed, O., & Lovibond, P. F. (2015a). The impact of previously learned feature-relevance on generalisation of conditioned fear in humans. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 59–65. https://doi.org/10.1016/j.jbtep.2014.08.001

Ahmed, O., & Lovibond, P. F. (2015b). The Impact of Instructions on Generalization of Conditioned Fear in Humans. *Behavior Therapy*, *46*(5), 597–603. https://doi.org/10.1016/j.beth.2014.12.007

Amir, C., Rose-McCandlish, M., Weger, R., Dildine, T. C., Mischkowski, D., Necka, E. A., Lee, I.-s., Wager, T. D., Pine, D. S., & Atlas, L. Y. (2022). Test-Retest Reliability of an Adaptive Thermal Pain Calibration Procedure in Healthy Volunteers. *The Journal of Pain*, *23*(9), 1543–1555. https://doi.org/10.1016/j.jpain.2022.01.011

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108

Anderson, B. A., & Yantis, S. (2013). Persistence of Value-Driven Attentional Capture. *Journal of experimental psychology. Human perception and performance*, *39*(1), 6–9. https://doi.org/10.1037/a0030860

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates, Inc.

Andreatta, M., & Pauli, P. (2019). Generalization of appetitive conditioned responses. *Psychophysiology*, *56*(9), e13397. https://doi.org/10.1111/psyp.13397

Austerweil, J. L., Sanborn, S., & Griffiths, T. L. (2019). Learning How to Generalize. *Cognitive Science*, *43*(8), e12777. https://doi.org/10.1111/cogs.12777

Bach, D. R., Friston, K. J., & Dolan, R. J. (2013). An improved algorithm for model-based analysis of evoked skin conductance responses. *Biological Psychology*, *94*(3), 490–497. https://doi.org/10.1016/j.biopsycho.2013.09.010

Baddeley, R. J., Osorio, D., & Jones, C. D. (2007). Generalization of color by chickens: Experimental observations and a Bayesian model. *The American Naturalist*, *169 Suppl 1*, S27–41. https://doi.org/10.1086/510142

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Basu, R., Gebauer, R., Herfurth, T., Kolb, S., Golipour, Z., Tchumatchenko, T., & Ito, H. T. (2021). The orbitofrontal cortex maps future navigational goals. *Nature*, 1–4. https://doi.org/10.1038/s41586-021-04042-9

Bechtel, W., & Mundale, J. (1999). Multiple Realizability Revisited: Linking Cognitive and Neural States. *Philosophy of Science*, *66*(2), 175–207. https://doi.org/10.1086/392683

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, *47*(4), 647–658. https://doi.org/10.1111/j.1469-8986.2009.00972.x

Berg, H., Ma, Y., Rueter, A., Kaczkurkin, A., Burton, P. C., DeYoung, C. G., MacDonald, A. W., Sponheim, S. R., & Lissek, S. M. (2020). Salience and central executive networks track overgeneralization of conditioned-fear in post-traumatic stress disorder. *Psychological Medicine*, 1–10. https://doi.org/10.1017/S0033291720001166

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954–967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Betancourt, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*. Retrieved July 9, 2020, from http://arxiv.org/abs/1701.02434

Betancourt, M., & Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *arXiv:1312.0906 [stat]*. Retrieved July 9, 2020, from http://arxiv.org/abs/1312.0906

Blitzstein, J. K., & Hwang, J. (2015). *Introduction to probability*. CRC Press/Taylor & Francis Group.

Borg, I., & Mair, P. (2017). The Choice of Initial Configurations in Multidimensional Scaling: Local Minima, Fit, and Interpretability. *Austrian Journal of Statistics*, *46*(2), 19–32. https://doi.org/10.17713/ajs.v46i2.561

Bottini, R., & Doeller, C. F. (2020). Knowledge Across Reference Frames: Cognitive Maps and Image Spaces. *Trends in Cognitive Sciences*, *24*(8), 606–619. https://doi.org/10.1016/j.tics.2020.05.008

Boyle, S., Roche, B., Dymond, S., & Hermans, D. (2016). Generalisation of fear and avoidance along a semantic continuum. *Cognition and Emotion*, *30*(2), 340–352. https://doi.org/10.1080/02699931.2014.1000831

Brosowsky, N. P., & Crump, M. J. C. (2021). Contextual recruitment of selective attention can be updated via changes in task relevance. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *75*, 19–34. https://doi.org/10.1037/cep0000221

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Darlington, T. R., Beck, J. M., & Lisberger, S. G. (2018). Neural implementation of Bayesian inference in a sensorimotor behavior. *Nature Neuroscience*, *21*(10), 1442–1451. https://doi.org/10.1038/s41593-018-0233-y

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

de Voogd, L. D., Murray, Y. P. J., Barte, R. M., van der Heide, A., Fernández, G., Doeller, C. F., & Hermans, E. J. (2020). The role of hippocampal spatial representations in contextualization and generalization of fear. *NeuroImage*, *206*, 116308. https://doi.org/10.1016/j.neuroimage.2019.116308

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*(4), e1005508. https://doi.org/10.1371/journal.pcbi.1005508

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*, *20*(2), 149–182. https://doi.org/10.1007/BF02253869

Drevets, W. C., Burton, H., Videen, T. O., Snyder, A. Z., Simpson, J. R., & Raichle, M. E. (1995). Blood flow changes in human somatosensory cortex during anticipated stimulation. *Nature*, *373*(6511), 249–252. https://doi.org/10.1038/373249a0

Dunsmoor, J. E., Kroes, M. C. W., Braren, S. H., & Phelps, E. A. (2017). Threat intensity widens fear generalization gradients. *Behavioral Neuroscience*, *131*(2), 168–175. https://doi.org/10.1037/bne0000186

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology*, *89*, 300–305. https://doi.org/10.1016/j.biopsycho.2011.11.002

Dunsmoor, J. E., Mitroff, S. R., & LaBar, K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, *16*(7), 460–469. https://doi.org/10.1101/lm.1431609

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: How humans generalize fear. *Trends in Cognitive Sciences*, *19*(2), 73–77. https://doi.org/10.1016/j.tics.2014.12.003

Dunsmoor, J. E., Prince, S. E., Murty, V. P., Kragel, P. A., & LaBar, K. S. (2011). Neurobehavioral mechanisms of human fear generalization. *NeuroImage*, *55*(4), 1878–1888. https://doi.org/10.1016/j.neuroimage.2011.01.041

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., & Hermans, D. (2015). Fear Generalization in Humans: Systematic Review and Implications for Anxiety Disorder Research. *Behavior Therapy*, *46*(5), 561–582. https://doi.org/10.1016/j.beth.2014.10.001

Eichenbaum, A., Scimeca, J. M., & D'Esposito, M. (2020). Dissociable Neural Systems Support the Learning and Transfer of Hierarchical Control Structure. *Journal of Neuroscience*, *40*(34), 6624–6637. https://doi.org/10.1523/JNEUROSCI.0847-20.2020

Ekman, P., & Oster, H. (1979). Facial Expressions of Emotion. *Annual Review of Psychology*, *30*(1), 527–554. https://doi.org/10.1146/annurev.ps.30.020179.002523

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258–1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*, *21*(4), 500–508. https://doi.org/10.1038/mp.2015.88

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, 101508. https://doi.org/10.1016/j.cogpsych.2022.101508

Gelman, A. (2014). *Bayesian data analysis* (Third edition). CRC Press.

Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. *PLOS Computational Biology*, *11*(11), e1004567. https://doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50. https://doi.org/10.1016/j.cobeha.2015.07.007

Ghirlanda, S. (2002). Intensity generalisation: Physiology and modelling of a neglected topic. Retrieved July 9, 2020, from http://cogprints.org/5273/

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36. https://doi.org/10.1006/anbe.2003.2174

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2021). Neural surprise in somatosensory Bayesian learning. *PLOS Computational Biology*, *17*(2), e1008068. https://doi.org/10.1371/journal.pcbi.1008068

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: Combining reinforcement learning theory with fMRI data. *WIREs Cognitive Science*, *1*(4), 501–510. https://doi.org/10.1002/wcs.57

Goulden, N., Khusnulina, A., Davis, N. J., Bracewell, R. M., Bokde, A. L., McNulty, J. P., & Mullins, P. G. (2014). The salience network is responsible for switching between the default mode network and the central executive network: Replication from DCM. *NeuroImage*, *99*, 180–190. https://doi.org/10.1016/j.neuroimage.2014.05.052

Greenberg, T., Carlson, J. M., Cha, J., Hajcak, G., & Mujica-Parodi, L. R. (2013a). Neural reactivity tracks fear generalization gradients. *Biological Psychology*, *92*(1), 2–8. https://doi.org/10.1016/j.biopsycho.2011.12.007

Greenberg, T., Carlson, J. M., Cha, J., Hajcak, G., & Mujica-Parodi, L. R. (2013b). Ventromedial Prefrontal Cortex Reactivity Is Altered In Generalized Anxiety Disorder During Fear Generalization. *Depression and Anxiety*, *30*(3), 242–250. https://doi.org/10.1002/da.22016

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79–88. https://doi.org/10.1037/h0046219

Hodsoll, S., Viding, E., & Lavie, N. (2011). Attentional capture by irrelevant emotional distractor faces. *Emotion*, *11*, 346–353. https://doi.org/10.1037/a0022771

Huggins, A. A., Weis, C. N., Parisi, E. A., Bennett, K. P., Miskovic, V., & Larson, C. L. (2021). Neural substrates of human fear generalization: A 7T-fMRI investigation. *NeuroImage*, *239*, 118308. https://doi.org/10.1016/j.neuroimage.2021.118308

Jackson, J., Rich, A. N., Williams, M. A., & Woolgar, A. (2017). Feature-selective Attention in Frontoparietal Cortex: Multivoxel Codes Adjust to Prioritize Task-relevant Information. *Journal of Cognitive Neuroscience*, *29*(2), 310–321. https://doi.org/10.1162/jocn_a_01039

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science (New York, N.Y.)*, *338*(6109), 953–956. https://doi.org/10.1126/science.1227489

Kaczkurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., Sponheim, S. R., & Lissek, S. (2016). Neural Substrates of Overgeneralized Conditioned Fear in PTSD. *American Journal of Psychiatry*, *174*(2), 125–134. https://doi.org/10.1176/appi.ajp.2016.15121549

Kampermann, L., Tinnermann, A., & Büchel, C. (2021). Generalization of placebo pain relief. *Pain*, *162*(6), 1781–1789. https://doi.org/10.1097/j.pain.0000000000002166

Kausche, F. M., Zerbes, G., Kampermann, L., Büchel, C., & Schwabe, L. (2021). Neural signature of delayed fear generalization under stress. *Psychophysiology*. https://doi.org/10.1111/psyp.13917

Kausche, F. M., Zerbes, G., Kampermann, L., Müller, J. C., Wiedemann, K., Büchel, C., & Schwabe, L. (2021). Acute stress leaves fear generalization in healthy individuals intact. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-021-00874-0

Keltner, D., & Kring, A. M. (1998). Emotion, Social Function, and Psychopathology. *Review of General Psychology*, *2*(3), 320–342. https://doi.org/10.1037/1089-2680.2.3.320

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens. *The Journal of Neuroscience*, *21*(16), RC159–RC159. https://doi.org/10.1523/JNEUROSCI.21-16-j0002.2001

Korn, C. W., Staib, M., Tzovara, A., Castegnetti, G., & Bach, D. R. (2017). A pupil size response model to assess fear learning. *Psychophysiology*, *54*(3), 330–343. https://doi.org/10.1111/psyp.12801

Korzybski, A. (1933). *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. International Non-Aristotelian Library Publishing Company.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. Retrieved March 14, 2022, from https://www.frontiersin.org/article/10.3389/neuro.06.004.2008

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lange, I., Goossens, L., Michielse, S., Bakker, J., Lissek, S., Papalini, S., Verhagen, S., Leibold, N., Marcelis, M., Wichers, M., Lieverse, R., van Os, J., van Amelsvoort, T., & Schruers, K. (2017). Behavioral pattern separation and its link to the neural mechanisms of fear generalization. *Social Cognitive and Affective Neuroscience*, *12*(11), 1720–1729. https://doi.org/10.1093/scan/nsx104

Lashley, K. S., & Wade, M. (1946). The Pavlovian theory of generalization. *Psychological Review*, *53*(2), 72–87. https://doi.org/10.1037/h0059999

Laufer, O., Israeli, D., & Paz, R. (2016). Behavioral and Neural Mechanisms of Overgeneralization in Anxiety. *Current Biology*, *26*(6), 713–722. https://doi.org/10.1016/j.cub.2016.01.023

Laufer, O., & Paz, R. (2012). Monetary Loss Alters Perceptual Thresholds and Compromises Future Decisions via Amygdala and Prefrontal Networks. *Journal of Neuroscience*, *32*(18), 6304–6311. https://doi.org/10.1523/JNEUROSCI.6281-11.2012

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, *23*(4-5), 727–738. https://doi.org/10.1023/a:1025048802629

LeDoux, J. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences*, *111*(8), 2871–2878. https://doi.org/10.1073/pnas.1400335111

Lee, J. C., Lovibond, P. F., & Hayes, B. K. (2019). Evidential diversity increases generalisation in predictive learning. *Quarterly Journal of Experimental Psychology*, *72*(11), 2647–2657. https://doi.org/10.1177/1747021819857065

Lee, J. C., Lovibond, P. F., Hayes, B. K., & Navarro, D. J. (2019). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General*, *148*, 289–303. https://doi.org/10.1037/xge0000496

Lee, M. D., & Pope, K. J. (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology*, *47*(1), 32–46. https://doi.org/10.1016/S0022-2496(02)00019-6

Lehnert, L., Littman, M. L., & Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLOS Computational Biology*, *16*(10), e1008317. https://doi.org/10.1371/journal.pcbi.1008317

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/j.neuron.2016.12.040

Lissek, S. (2012). Toward an Account of Clinical Anxiety Predicated on Basic, Neurally Mapped Mechanisms of Pavlovian Fear-Learning: The Case for Conditioned Overgeneralization. *Depression and Anxiety*, *29*(4), 257–263. https://doi.org/https://doi.org/10.1002/da.21922

Lissek, S., Biggs, A. L., Rabin, S. J., Cornwell, B. R., Alvarez, R. P., Pine, D. S., & Grillon, C. (2008). Generalization of Conditioned Fear-Potentiated Startle in Humans. *Behaviour research and therapy*, *46*(5), 678–687. https://doi.org/10.1016/j.brat.2008.02.005

Lissek, S., Bradford, D. E., Alvarez, R. P., Burton, P., Espensen-Sturges, T., Reynolds, R. C., & Grillon, C. (2014). Neural substrates of classically conditioned fear-generalization in humans: A parametric fMRI study. *Social Cognitive and Affective Neuroscience*, *9*(8), 1134–1142. https://doi.org/10.1093/scan/nst096

Lommen, M. J. J., Engelhard, I. M., & van den Hout, M. A. (2010). Neuroticism and avoidance of ambiguous stimuli: Better safe than sorry? *Personality and Individual Differences*, *49*, 1001–1006. https://doi.org/10.1016/j.paid.2010.08.012

Loose, L. S., Wisniewski, D., Rusconi, M., Goschke, T., & Haynes, J.-D. (2017). Switch-Independent Task Representations in Frontal and Parietal Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(33), 8033–8042. https://doi.org/10.1523/JNEUROSCI.3656-16.2017

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5–5. https://doi.org/10.1167/3.8.5

Markovic, J., Anderson, A. K., & Todd, R. M. (2014). Tuning to the significant: Neural and genetic processes underlying affective enhancement of visual perception and memory. *Behavioural Brain Research*, *259*, 229–241. https://doi.org/10.1016/j.bbr.2013.11.018

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. https://doi.org/10.1037/a0039400

McHugh, T. J., Jones, M. W., Quinn, J. J., Balthasar, N., Coppari, R., Elmquist, J. K., Lowell, B. B., Fanselow, M. S., Wilson, M. A., & Tonegawa, S. (2007). Dentate Gyrus NMDA Receptors Mediate Rapid Pattern Separation in the Hippocampal Network. *Science*, *317*(5834), 94–99. https://doi.org/10.1126/science.1140263

Menon, V. (2011). Large-scale brain networks and psychopathology: A unifying triple network model. *Trends in Cognitive Sciences*, *15*(10), 483–506. https://doi.org/10.1016/j.tics.2011.08.003

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Morey, R. A., Haswell, C. C., Stjepanović, D., Dunsmoor, J. E., & LaBar, K. S. (2020). Neural correlates of conceptual-level fear generalization in posttraumatic stress disorder. *Neuropsychopharmacology*, *45*(8), 1380–1389. https://doi.org/10.1038/s41386-020-0661-8

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling Assumptions in Inductive Generalization. *Cognitive Science*, *36*(2), 187–223. https://doi.org/10.1111/j.1551-6709.2011.01212.x

Navarro, D. J., Lee, M. D., Dry, M. J., & Schultz, B. (2008). Extending and Testing the Bayesian Theory of Generalization. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1746–1751.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Nonyane, B. A. S., & Theobald, C. M. (2006). Design Sequences for Sensory Studies: Achieving Balance for Carry-over and Position Effects.

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning (D. Lee, Ed.). *eLife*, *7*, e34779. https://doi.org/10.7554/eLife.34779

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522. https://doi.org/10.1037/0033-295x.108.3.483

Öhman, A. (2009). Of snakes and faces: An evolutionary perspective on the psychology of fear. *Scandinavian Journal of Psychology*, *50*(6), 543–552. https://doi.org/10.1111/j.1467-9450.2009.00784.x

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness"? *Journal of Personality and Social Psychology*, *36*(11), 1251–1258. https://doi.org/10.1037/0022-3514.36.11.1251

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466–478. https://doi.org/10.1037/0096-3445.130.3.466

Onat, S. (2018, October). *Towards a new understanding of fear generalization and its neural origin* (preprint). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27311v1

Onat, S., & Büchel, C. (2015). The neuronal basis of fear generalization in humans. *Nature Neuroscience*, *18*(12), 1811–1818. https://doi.org/10.1038/nn.4166

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orr, S. P., & Lanzetta, J. T. (1980). Facial expressions of emotion as conditioned stimuli for human autonomic responses. *Journal of Personality and Social Psychology*, *38*(2), 278–282. https://doi.org/10.1037/0022-3514.38.2.278

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford Univ. Press.

Ptak, R., Schnider, A., & Fellrath, J. (2017). The Dorsal Frontoparietal Network: A Core System for Emulated Action. *Trends in Cognitive Sciences*, *21*(8), 589–599. https://doi.org/10.1016/j.tics.2017.05.002

Radua, J., & Fullana, M. A. (2022). The amygdala and the nine circles of scientific hell - Confirmation bias in the brain correlates of psychopathy. *Neuroscience & Biobehavioral Reviews*, *143*, 104951. https://doi.org/10.1016/j.neubiorev.2022.104951

Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Resnik, J., Sobel, N., & Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. *Nature Neuroscience*, *14*(6), 791–796. https://doi.org/10.1038/nn.2802

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Rolls, E. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, *7*. Retrieved January 4, 2023, from https://www.frontiersin.org/articles/10.3389/fnsys.2013.00074

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative Valence Widens Generalization of Learning. *Journal of Neuroscience*, *30*(31), 10460–10464. https://doi.org/10.1523/JNEUROSCI.2377-10.2010

Schönemann, P. H., & Lazarte, A. (1987). Psychophysical maps for subadditive dissimilarity ratings. *Perception & Psychophysics*, *42*(4), 342–354. https://doi.org/10.3758/BF03203090

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Seligman, M. E. (1970). On the generality of the laws of learning. *Psychological Review*, *77*(5), 406–418. https://doi.org/10.1037/h0029790

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. https://doi.org/10.1007/BF02289630

Silva, L., & Zanella, G. (2022, September). *Robust leave-one-out cross-validation for high-dimensional Bayesian models* (tech. rep. No. arXiv:2209.09190). arXiv. Retrieved October 5, 2022, from http://arxiv.org/abs/2209.09190

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*, *121*(3), 526–558. https://doi.org/10.1037/a0037018

Soto, F. A., & Wasserman, E. A. (2010). Integrality/Separability of Stimulus Dimensions and Multidimensional Generalization in Pigeons. *Journal of experimental psychology. Animal behavior processes*, *36*(2), 194–205. https://doi.org/10.1037/a0016560

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*, 430–444. https://doi.org/10.1037/h0062885

Strang, G. (2021). *Introduction to linear algebra* (Fifth edition). Wellesley-Cambridge Press.

Struyf, D., Zaman, J., Hermans, D., & Vervliet, B. (2017). Gradients of fear: How perception influences fear generalization. *Behaviour Research and Therapy*, *93*, 116–122. https://doi.org/10.1016/j.brat.2017.04.001

Struyf, D., Zaman, J., Vervliet, B., & Van Diest, I. (2015). Perceptual discrimination in fear generalization: Mechanistic and clinical implications. *Neuroscience and Biobehavioral Reviews*, *59*, 201–207. https://doi.org/10.1016/j.neubiorev.2015.11.004

Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2021). Learning biases to angry and happy faces during Pavlovian aversive conditioning. *Emotion*, *21*(4), 742–756. https://doi.org/10.1037/emo0000733

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General*, *147*(6), 905–923. https://doi.org/10.1037/xge0000424

Summerfield, C., Luyckx, F., & Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, *184*, 101717. https://doi.org/10.1016/j.pneurobio.2019.101717

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Tenenbaum, J. B., & Griffiths, T. L. (2001b). Some specifics about generalization. *Behavioral and Brain Sciences*, *24*(4), 762–778. https://doi.org/10.1017/S0140525X01780089

Tomov, M. S., Dorfman, H. M., & Gershman, S. J. (2018). Neural computations underlying causal structure learning. *The Journal of Neuroscience*, *38*(32), 7143–7157. https://doi.org/10.1523/JNEUROSCI.3336-17.2018

Tuominen, L., Boeke, E., DeCross, S., Wolthusen, R. P., Nasr, S., Milad, M., Vangel, M., Tootell, R., & Holt, D. (2019). The relationship of perceptual discrimination to neural mechanisms of fear generalization. *NeuroImage*, *188*, 445–455. https://doi.org/10.1016/j.neuroimage.2018.12.034

Uddin, L. Q., Yeo, B. T., & Spreng, R. N. (2019). Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, *32*(6), 926–942. https://doi.org/10.1007/s10548-019-00744-6

Vaidya, A. R., Jones, H. M., Castillo, J., & Badre, D. (2021). Neural representation of abstract task structure during generalization (M. Liljeholm, R. B. Ivry, C. Ranganath, & S. Michelmann, Eds.). *eLife*, *10*, e63226. https://doi.org/10.7554/eLife.63226

van Meurs, B., Wiggert, N., Wicker, I., & Lissek, S. (2014). Maladaptive behavioral consequences of conditioned fear-generalization: A pronounced, yet sparsely studied, feature of anxiety pathology. *Behaviour Research and Therapy*, *57*, 29–37. https://doi.org/10.1016/j.brat.2014.03.009

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2019). Pareto Smoothed Importance Sampling. *arXiv:1507.02646 [stat]*. Retrieved July 9, 2020, from http://arxiv.org/abs/1507.02646

Vervliet, B., Kindt, M., Vansteenwegen, D., & Hermans, D. (2010). Fear generalization in humans: Impact of verbal instructions. *Behaviour Research and Therapy*, *48*(1), 38–43. https://doi.org/10.1016/j.brat.2009.09.005

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of Human Fear Acquisition and Extinction within a Novel Arbitrary Stimulus Category. *PLOS ONE*, *9*(5), e96569. https://doi.org/10.1371/journal.pone.0096569

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25. https://doi.org/10.1016/j.cogpsych.2015.07.001

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585–594. https://doi.org/10.1016/j.tics.2005.10.011

Wang, X., Cheng, B., Luo, Q., Qiu, L., & Wang, S. (2018). Gray Matter Structural Alterations in Social Anxiety Disorder: A Voxel-Based Meta-Analysis. *Frontiers in Psychiatry*, *9*, 449. https://doi.org/10.3389/fpsyt.2018.00449

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, *53*(3), 63:1–63:34. https://doi.org/10.1145/3386252

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120. https://doi.org/10.3758/BF03202828

Webler, R. D., Berg, H., Fhong, K., Tuominen, L., Holt, D. J., Morey, R. A., Lange, I., Burton, P. C., Fullana, M. A., Radua, J., & Lissek, S. (2021). The neurobiology of human fear generalization: Meta-analysis and working neural model. *Neuroscience & Biobehavioral Reviews*, *128*, 421–436. https://doi.org/10.1016/j.neubiorev.2021.06.035

Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, *35*(7), 1092–1104. https://doi.org/10.1111/j.1460-9568.2012.08017.x

Witty, S., Lee, J. K., Tosch, E., Atrey, A., Clary, K., Littman, M. L., & Jensen, D. (2021). Measuring and characterizing generalization in deep reinforcement learning. *Applied AI Letters*, *2*(4), e45. https://doi.org/10.1002/ail2.45

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752. https://doi.org/10.1016/j.neuroimage.2010.04.035

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Yantis, S. (2008). The Neural Basis of Selective Attention: Cortical Sources and Targets of Attentional Modulation. *Current Directions in Psychological Science*, *17*(2), 86–90. https://doi.org/10.1111/j.1467-8721.2008.00554.x

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yuan, C., Zhu, H., Ren, Z., Yuan, M., Gao, M., Zhang, Y., Li, Y., Meng, Y., Gong, Q., Lui, S., Qiu, C., & Zhang, W. (2018). Precuneus-related regional and network functional deficits in social anxiety disorder: A resting-state functional MRI study. *Comprehensive Psychiatry*, *82*, 22–29. https://doi.org/10.1016/j.comppsych.2017.12.002

Zaman, J., Struyf, D., Ceulemans, E., Beckers, T., & Vervliet, B. (2019). Probing the role of perception in fear generalization. *Scientific Reports*, *9*(1), 10026. https://doi.org/10.1038/s41598-019-46176-x

Zaman, J., Struyf, D., Ceulemans, E., Vervliet, B., & Beckers, T. (2021). Perceptual errors are related to shifts in generalization of conditioned responding. *Psychological Research*, *85*(4), 1801–1813. https://doi.org/10.1007/s00426-020-01345-w

Zeidan, F., Lobanov, O. V., Kraft, R. A., & Coghill, R. C. (2015). Brain Mechanisms Supporting Violated Expectations of Pain. *Pain*, *156*(9), 1772–1785. https://doi.org/10.1097/j.pain.0000000000000231

Abstract

In a dynamic environment, the ability to adapt to changes is a key factor for survival. This cognitive ability is called generalization and the fact that humans excel at it is a major reason for our evolutionary success. Commensurate with this, the literature on human generalization is vast, but our understanding suffers from a lack of a unifying framework. Instead, generalization in associative learning, in reinforcement learning and inductive reasoning are implicitly treated as separate entities. Stimulus generalization, i.e. generalization of associative learning is most often studied in the context of fear conditioning and the explanatory focus is typically strongly concerned with the role of perception. Generalization in reinforcement learning is usually assumed to rely on abstractions of state spaces, e.g. via dimensionality reduction. Lastly, research on inductive reasoning proposes different Bayesian mechanisms.

To arrive at a more general mechanism of generalization, I propose a Bayesian model of generalization that integrates dimensionality reduction into a probabilistic framework and is applicable to probabilistic reinforcement and therefore the typical study designs in stimulus generalization. To test the predictions of the model and to find common ground between stimulus generalization and abstraction in reinforcement learning with respect to their neural signature, I conducted a series of experiments. Importantly, I used face stimuli that differed on facial identity and facial expression, which allowed me to investigate dimensional preferences and their relationship with prior knowledge. In addition, the study design contained time-resolved ratings to characterize the temporal dynamics. Using the proposed model I could then make specific predictions for the data that I expected to arise from those studies.

Behavioral ratings closely followed the predictions of the model in all three studies, independently of the value of the reinforcement. Initial ratings were dependent on the value and strength of the emotional expression but became increasingly dependent on the perceptual similarity to the reinforced stimulus. The latter effect was stronger along the emotion dimension in all studies. Model comparison confirmed that the data followed all of the predictions of the model.

Using fMRI data from one of the studies, I found positive correlations with behavioral generalization in the frontoparietal attention network and the salience network and negative correlations in the default mode network. Importantly, generalization along the emotion dimension was only associated with the frontoparietal attention network and the salience network, which mirrors results from reinforcement learning. In a last step, I found that

representations in the middle frontal gyrus mirrored the behavioral relevance of the different dimensions.

Taken together, I present coherent theoretical considerations and empirical evidence for a common mechanism of generalization that can be well explained as a Bayesian model and suggest that low-dimensional representations of stimuli are one key neural mechanism underlying generalization.

Zusammenfassung

In einer dynamischen Umwelt ist die Fähigkeit sich anzupassen ein essentieller Prädiktor für das Überleben. Diese kognitive Fähigkeit nennen wir Generalisierung und die Tatsache, dass Menschen darin unübertroffen sind ist ein wichtiger Grund für unseren evolutionären Erfolg. Dementsprechend ist die Literatur zu Generalisierung in Menschen sehr umfangreich. Allerdings leidet unser Verständnis darunter, dass es keinen einheitlichen erklärenden Rahmen gibt. Stattdessen werden Generalierung in assoziativem Lernen, Verstärkungslernen und induktive Schlussfolgerung implizit als separate Entitäten behandelt. Stimulusgeneralisierung, d.h. die Generalisierung von assoziativem Lernen wird meistens im Kontext von Furchtkonditioning erforscht und der Fokus der Erklärungsansätze liegt häufig auf der Rolle der Wahrnehmung. Generalisierung im Verstärkungslernen wird meistens als Abstraktion von *state spaces*, z.B. durch Dimensionsreduktion, erklärt. Und letztens, die Forschung zur induktiven Schlussfolgerung schlägt verschiedene Bayesianische Mechanismen vor.

Um zu einem generelleren Verständnis von Generalisierung zu gelangen, schlage ich ein Bayesianisches Modell der Generalisierung vor, dass Dimensionsreduktion in einen probabilistischen Rahmen integriert und auf probabilistische Verstärkungen und damit auf die typischen Forschungsdesigns in der Forschung zur Stimulusgeneralisierung anwendbar ist. Um die Vorhersagen des Modells zu testen und Gemeinsamkeiten zwischen Stimulusgeneralisierung und Abstraktionen im Kontext von Verstärkungslernen mit Bezug auf deren neuronale Signatur zu finden, habe ich eine Reihe von Experimenten durchgeführt. Hierbei ist relevant, dass ich als Stimuli Gesichter verwendet habe, die sich in Bezug auf die Identität und die Art und Stärke des emotionalen Ausdrucks unterscheiden. Außerdem enthielt das Studiendesign mehrere Einschätzungen der Probanden mit zeitlicher Auflösung, die es mir ermöglichten, die zeitliche Dynamik zu charakterisieren. Mithilfe des vorgeschlagenen Modells konnte ich dann spezifische Vorhersagen für die Daten machen, die ich als Ergebnis der Experimente erwartet habe.

In allen Studien folgten die Einschätzungen der Probanden den Vorhersagen des Modells ziemlich exakt. Dies war unabhängig von der Wertigkeit des Verstärkers. Anfänglich hingen Einschätzungen primär von Wert und Stärke des Emotionsausdrucks ab, aber mit der Zeit wurde die perzeptuelle Ähnlichkeit zum verstärkten Stimulus relevanter. Dieser Effekt war in allen Studien stärker entlang der Emotionsdimension. Modellvergleiche bestätigten dass die Daten allen Vorhersagen des Modells entsprachen.

Mithilfe von fMRT Daten aus einer der Studien konnte ich zeigen, dass Aktivität im frontoparietalen Aufmerksamkeitsnetzwerk und im Salienznetzwerk positiv und Aktivität im

Ruhezustandsnetzwerk negativ mit der Generalisierung im Verhalten korreliert war. Besonders relevant war, dass die Generalisierung entlang der Emotionsdimension nut mit dem frontoparietalen Aufmerksamkeitsnetzwerk und dem Salienznetzwerk assoziiert war, was sich mit Ergebnissen aus der Literatur zum Verstärkungslernen überschneidet. In einem letzten Schritt fand ich, dass Repräsentationen im mittleren frontalen Gyrus die Verhaltensrelevanz der verschiedenen Dimensionen widerspiegelten.

Zusammengenommen präsentiere ich kohärente theoretische Überlegungen und empirische Evidenz für einen gemeinsamen Mechanismus der Generalisierung, der gut als Bayesianisches Modell erklärt werden kann und schlage vor, dass niedrigdimensionale Repräsentationen von Stimuli ein entscheidender neuronaler Mechanismus sind, der der Generalisierung zugrunde liegt.

Acknowledgements

Although my name is the only one on the title page, a lot of people have contributed directly or indirectly to the work presented in this thesis. I would like to thank all of them for their help and support during this not always easy period of my life.

First of all, Prof. Dr. Christian Büchel, for his guidance and encouragement. For always having an open door, trusting in my abilities and for giving me room to grow scientifically. The colleagues that I had the pleasure of working with for providing me with a friendly and stimulating environment. In particular, I would like to mention Lea Kampermann, Selim Onat, Björn Horing, Alexandra Tinnermann, Alina Panzel, Noah Hipp and Tobias Sommer. Some of you became friends that I do not want to miss in my life. Waldemar Schwarz, Katrin Bergholz, Kathrin Wendt for handling the scary big magnet, Jenny Rohwer for supporting me with behavioral data collection and Estefanía Orozco Rodríguez for handling the organization of the fMRI data collection and making the long hours in front of the scanner much more bearable. Heike Path and Carolina Dlugosch for understanding chaotic scientists while staying organized themselves and Matthias Pietsch for showing me the magic of bash scripts.

Guido van Rossum, Linus Torvalds, Donald Knuth, Bob Carpenter, Bram Moolenaar and countless others for providing me with tools that made my life easier. Thanks to you I did not have to use Matlab, Windows or Word any longer than absolutely necessary.

Mauro Pulin and Mattia Chini for their companionship while working out the body and the brain, respectively. Madita Standke-Erdmann for being the best friend one could wish for. My parents, Johannes and Barbara, my brother David and my sister Eva for their unconditional support and for providing me with a safe haven that I know I can always return to. And for providing me with a loving and stimulating upbringing that made me who I am today. I am well aware of how privileged I am to have such a family.

And finally, Marie, my beautiful and smart girlfriend. For being an amazingly intelligent colleague and a wonderful partner, all at the same time. I am very lucky to have you by my side.

Curriculum Vitae

Entfällt aus datenschutzrechtlichen Gründen.

Entfällt aus datenschutzrechtlichen Gründen.

Eidesstattliche Erklärung

Ich versichere ausdrücklich, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzen Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe. Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe. Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zu Erkennung von Plagiaten überprüft werden kann.