From THE DEPARTMENT OF CLINICAL NEUROSCIENCE
THE DIVISION OF PSYCHOLOGY
Karolinska Institutet, Stockholm, Sweden

THE ROLE OF AVERSIVE LEARNING IN SOCIAL INTERACTIONS

Tanaz Molapour

Karolinska Institutet

Stockholm 2016

Front cover: The creative efforts of Dr. Christopher Berger, myself, and Georges Seurat (from the painting, “A Sunday Afternoon on La Grande Jatte”).

Published by Karolinska Institutet.
Printed by AJ E-print AB

ISBN 978-91-7676-410-7

The Role of Aversive Learning in Social Interactions

Thesis for Doctoral Degree (Ph.D.)

By
Tanaz Molapour

M.A.

Principal Supervisor:
Dr. Andreas Olsson
Department of Clinical Neuroscience
Division of Psychology
Karolinska Institutet

Co-supervisor:
Prof. Henrik Ehrsson
Department of Neuroscience
Karolinska Institutet

Opponent:
Dr. Grit Hein
Laboratory for Social and Neural Systems Research
Institute for Empirical Research in Economics
University of Zurich

Examination Board:
Prof. Håkan Fischer
Department of Psychology
Stockholm University

Dr. Fredrik Åhs
Department of Psychology
Uppsala University

Dr. Lisa Thorell
Department of Psychology
Karolinska University

“Some say they see poetry in my paintings, I see only science.”

Georges Seurat

ABSTRACT

It could be argued that our survival as humans hinges on our ability to interact socially with others. Our social interactions are influenced by evaluations of each other: we cooperate with those we like and avoid or are aggressive towards those we dislike or are afraid of. The aim of this thesis was to investigate how we come to learn to fear or dislike other individuals based on who they are; and how such learned evaluations influence actual social interactive behaviors. One elegant way to study how humans respond and react to threats in the environment is classical fear conditioning, where we can study how emotional values are created, upheld and changed. Research using classical fear conditioning has found that people are predisposed to develop stronger associations between threatening events and certain categories of stimuli (e.g., snakes, angry faces, and faces of individuals belonging to social out-groups). These biased aversions tend to persist even when circumstances change and the threat is no longer present. Though the fear system underlying this type of learning may be useful under some circumstances, it may also be at the root of some persistent social problems affecting modern societies (e.g., xenophobia). To address these questions experimentally, this thesis aimed to identify how we learn to associate threats to different social groups (e.g., racial and hierarchical) (**Study I & Study II**); whether learned aversions influence anti-social interactive behaviors (**Study III**); and to study the mechanisms of maladaptive reciprocal punishments in dyadic interactions (**Study IV**).

In **Study I**, we found that activity linked to both conditioned fear and perception of racial out-group members jointly contributed to the expression of race-based biases in learning and behavior. Importantly, we showed that brain activity in the fear-learning-bias network was related to participants' discriminatory interactions with new out-group members at a later time. In **Study II**, we investigated the interaction between learned social dominance and social out-group (i.e., ethnicity) threats to understand if dominance hierarchy knowledge (i.e., observation of threats) can change direct experience with out-group members. We found a dissociation between implicit and explicit measures of out-group biases, such that implicit measures (i.e., Implicit Association Task and skin-conductance responses) of the participants revealed out-group biases, whereas their explicit measures (i.e., modern racial prejudice scale and a social interactive task; the modified 'Cyberball' game) did not. In **Study III**, we found that learned aversions influenced future retaliation in a social context. Our results suggest that aggressive traits, when paired with aversive learning experiences, enhance the likelihood to act anti-socially toward others. In **Study IV**, we demonstrate that participants punish co-players, despite the cost of receiving punishment back. These findings describe a form of self-punitive behavior previously documented in animals. Participants' tendencies to administer shocks were exacerbated when the co-player initiated punishment, indicating that a small initial offense motivated punishing behavior over time. This finding suggests a simple experimental model of a vicious cycle of punishments. Together these findings highlight the role of aversive learning in social interactions.

LIST OF PUBLICATIONS

This thesis is based on the following publications, which are referred to in the text by their roman numerals (**Studies I-IV**):

I. **Molapour, T**., Golkar, A., Navarrete, C.D., Haaker, J., Olsson, A. (2015) Neural correlates of biased social fear learning and interaction in an intergroup context. *NeuroImage, 121*, 171-183.

II. **Molapour, T**., Haaker, J., Olsson, A. The relationship between social dominance threat and racial biases. *Manuscript*.

III. **Molapour, T**., Lindström, B., & Olsson, A. (2016) Aversive learning influences anti-social behavior. *Frontiers in Psychology*, 7:833.

IV. **Molapour,T**., Lindström, B., Bellander, M., Haaker, J., & Olsson, A. Reciprocal punishment in social dyadic interactions. *Manuscript*

LIST OF ADDITIONAL PUBLICATIONS

Publications by the author from the Department of Clinical Neuroscience, which are not included in this thesis[1]:

I. Haaker, J., **Molapour, T**., & Olsson, A. (2016) Conditioned social dominance threat: Observation of other's social dominance biases threat learning. *Social Cognitive and Affective Neuroscience*.

II. Lindström, B. R., Selbing, I., **Molapour, T**., & Olsson, A. (2014) Racial bias shapes social reinforcement learning. *Psychological Science*, 25(3), 711-719.

[1] For a complete list of publications by Tanaz Molapour visit: https://scholar.google.se/citations?user=WAqlae0AAAAJ&hl=en

CONTENTS

1 Introduction **1**

1.1 Social Aversive Learning *1*

1.2 Background on Classical Conditioning and Fear Conditioning *2*

1.2.1 Psychophysiological Responses in Fear Conditioning 3

1.2.2 The Neural Networks of Fear Conditioning 4

1.3 'Not All Stimuli Are Created Equal' *5*

1.3.1 Social Group Biases Based on Race 5

1.3.2 Fear Conditioning and Racial Biases 8

1.3.3 Fear Conditioning and Social Dominance Hierarchy 9

1.4 Aggressive Anti-Social Interactions *10*

1.4.1 Maladaptive Aggression 11

2 Aims **15**

2.1 Study I Aim *15*

2.2 Study II Aim *15*

2.3 Study III Aim *15*

2.4 Study IV Aim *15*

3 Methods **17**

3.1 Participants *17*

3.2 Experimental Stimuli *17*

3.2.1 Electric Shocks 18

3.3 Psychophysiological Measurements *19*

3.3.1 Skin Conductance Response 19

3.4 Functional Magnetic Resonance Imaging *19*

3.4.1 Parametric Analysis 20

3.4.2 Psychophysiological Interaction Connectivity Analysis 21

4 Overview of Studies **23**

4.1 Study I: Neural Correlates of Biased Social Fear Learning and Interaction in an Intergroup Context *23*

4.1.1 Study I Background and Rationale 23

4.1.2 Study I: Results and Conclusions 23

4.2 Study II: The Relationship Between Social Dominance Threat and Racial Biases *27*

4.2.1 Study II Background and Rationale 27

4.2.2 Study II Results and Conclusions 27

4.3 Study III: Aversive Learning Influences Anti-Social Behavior *28*

4.3.1 Study III Background and Rationale 28

4.3.2 Study III Results and Conclusion 29

4.4 Study IV: Reciprocal Punishment in Dyadic Social Interactions *31*

4.4.1 Study IV Background and Rationale 31

4.4.2 Study IV Results and Conclusions 31

5 Discussion **33**

6 Acknowledgments **43**

7 References **45**

LIST OF ABBREVIATIONS

ACC	Anterior Cingulate Cortex
ANOVA	Analysis of Variance
B	Basal nucleus
BLA	Basolateral amygdala
BOLD	Blood oxygen level dependent
CR	Conditioned response
CS	Conditioned stimulus
CS+	Conditioned stimulus coupled to US
CS-	Conditioned stimulus never coupled to US
FFA	Fusiform face area
fMRI	Functional magnetic resonance imaging
FPS	Fear-potentiated startle
GLM	General linear model
IAT	Implicit Association Test
KDEF	Karolinska directed emotional faces
MR	Magnetic resonance
MRI	Magnetic resonance imaging
PPI	Psycho-physiological interaction
PTSD	Post-traumatic stress disorder
RaFD	Radboud faces database
SEM	Standard error of the mean
SIT	Social interactive task
US	Unconditioned stimulus
vmPFC	Ventromedial prefrontal cortex

1 INTRODUCTION

Most of our social interactions depend on how we evaluate each other: we behave differently towards people we like compared to those we dislike. We tend to approach and help people we like, whereas we tend to avoid and aggress towards people we dislike. Aggression and avoidance behaviors can be adaptive if someone poses a threat to our well-being and survival. However, based on many real-life scenarios it seems as though people behave aggressively and anti-socially even when there is no real threat, and even when it sometimes comes at a cost. Our experiences from which we learn about threats or safety are not only isolated to specific circumstances, for example, our pre-conceived notions based on culturally learned information can change our experiences with someone. In studies contained within this thesis, I have examined the critical components of how we learn to associate aversive experiences to people based on who they are, and how these experiences are then reflected in social interactive behaviors. I have also examined how maladaptive anti-social behaviors (i.e., punishments) are initiated and maintained in a social context. In the following sections I will first outline the key concepts within the field of social aversive learning which are relevant to the studies contained within this thesis. I will then describe our methodological approach as well as some methodological considerations before summarizing the key findings from our studies. Lastly, I will attempt to situate our findings within the field of social aversive learning and interactive behavior, and I will also briefly describe some possible future directions and possible applications of our findings.

1.1 SOCIAL AVERSIVE LEARNING

If your neighbor behaves aggressively towards you, you may become scared and start to avoid him or her. Depending on the situation and your behavioral dispositions, you might instead choose to aggress back (i.e., reciprocate punishment). What role does classical conditioning—a simple associative learning process—play in social situations like these? Classical conditioning enables an organism to form associations between threatening events and preceding innocuous cues. Detecting and reacting to environmental threats clearly serves a critical evolutionary function, and fear learning offers an efficient way to transmit information about potential threats in the environment. While responses to naturally aversive events, such as painful stimuli (e.g., shocks) are hardwired reactions, organisms must learn to adapt these behaviors to predict and avoid many different types of potential threats. One critical question in fear learning is how to determine which stimuli do, and do not, pose a threat in the dynamic environment that we live in. The flexibility of fear learning allows the individual to generalize learned information in the past to make predictions about

the future. Although the ability to generalize fear responses to similar dangers in the future is adaptive behavior, it can also be maladaptive if learned experiences are generalized too broadly. Stimuli that are perceived as threats in one environment may not pose a threat in another environment. For example, encountering a snake on a hike should signal danger and result in an appropriate threat response; however, observing a snake in an enclosed space at a zoo does not require the same response. Moreover, in the case of social interactions, an inappropriate threat response could cause you to avoid or aggress towards an innocent neighbor simply because they resemble someone you had an aversive encounter with earlier that morning. Thus, the predictive values of certain signals for danger do not always lead to an appropriate response to these signals in the future. The maladaptive nature of overgeneralization of defensive behaviors can be exemplified by clinical anxiety disorders (e.g., post-traumatic stress disorder (PTSD), phobias, and panic disorders). Most importantly, in the studies within this thesis, I have examined the role of fear conditioning in various situations that involve social threat (involving other people). These findings are important for understanding the role of aversive learning in social interactions.

1.2 BACKGROUND ON CLASSICAL CONDITIONING AND FEAR CONDITIONING

In classical conditioning, learning is an association that is created between two stimuli presented close in time (Diamond & Rose, 1994; McLeod, 2007). For example, an initially neutral stimulus (i.e., the conditioned stimulus; CS), such as a sound, is presented together with rewarding or punishing stimulus (i.e., the unconditioned stimulus; US), that in turn evokes a behavioral response. As a result of the learned association between the CS and US, the organism shows an enhanced response to the CS alone, also referred to as a conditioned response (CR). This outcome indicates that an association has been created between the CS and US (Bouton & Moody, 2004; Diamond & Rose, 1994; Rescorla, 1988). *Fear conditioning* is specific form of classical conditioning used to investigate aversive learning in which the US is unpleasant. For example, fear conditioning experiments in rats use light or a sound as a neutral CS that does not evoke a fearful response. After repeated pairings between the CS and an aversive foot shock (US), the CS by itself can elicit a startle or freezing response (Davis, 1992; Phelps & LeDoux, 2005; Phelps, 2006a). That is, over time, the CS itself can come to evoke an anticipatory response, the conditioned response or CR in the rat. It is also important to understand how acquired fears change and diminish; for example, how one stops expressing conditioned responses to the sound by learning that the sound no longer poses any danger. This is done through extinction training. During

Extinction, all CSs are repeatedly presented again without the aversive US, allowing for the formation of new memory that the CS has become safe, which inhibits expression of the original fear memory (Lattal, Radulovic, & Lukowiak, 2006; Pavlov, 1927). At this point, the CR is considered extinguished (Davis, 1992; Fendt & Fanselow, 1999; Phelps, 2006a). Interestingly, extinguished fear can be recovered; it can re-emerge after extinction (i.e., spontaneous recovery), or as a result of a change in the context (i.e., the renewal effect), or from one or more unsignaled presentations of the US (i.e., the reinstatement effect). The re-emergence of the previously learned fear suggest that extinction is a new learning process, and the fear reduction results from inhibition rather than erasure of the original fear memory (Bouton, Westbrook, Corcoran, & Maren, 2006; Ji & Maren, 2007).

1.2.1 Psychophysiological Responses in Fear Conditioning

What happens when you feel threatened? Fear is characterized by both high arousal and negative valence (Lang, 1995), and the expression of fear is characterized, for example, through a common psychophysiological response pattern including potentiation of the startle reflex, increases in skin conductance response (SCR), blood pressure and heart rate acceleration (Globisch, Hamm, & Esteves, 1999). The startle reaction is a defensive reflex that is elicited in response to a sudden and intense stimulus such a loud noise or a light flash. In humans, the blink reflex is the component of the startle response commonly used as an index of CR in human fear conditioning preparations, and within this context, it is commonly referred to as the fear-potentiated startle (FPS) (Lang, Davis, & Öhman, 2000). The basis of the FPS is that the startle blink reflex is potentiated when the individual is in an aversive or fearful state and the magnitude of the startle reflex has been shown to be directly related to negative stimulus valence (Hamm & Weike, 2005). Another index that is commonly used is skin conductance response (SCR), also known as the electrodermal response. This is the measure I have used in **Studies I-III** of this thesis. SCRs reflect the phasic increase in skin conductance that occurs in response to physiologically arousing stimuli, such as negative or positive pictures, and are modulated both by stimulus novelty and intensity as well as by attentional processes (Öhman, 1979). In human fear conditioning, CR is commonly inferred from increased SCRs in the presence of a CS that is predictive of the US (the CS+) compared to SCRs to the control stimulus (CS-). The difference between these measures is referred to as differential SCR. There are many advantages in using SCRs; it is a non-intrusive measure, and interestingly, SCR changes have been linked to changes in amygdala activity during fear conditioning (e.g., Cheng,

Knight, Smith, & Helmstetter, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Wood, Ver Hoef, & Knight, 2014).

1.2.2 The Neural Networks of Fear Conditioning

Researchers have made a lot of progress in mapping the neural circuitry and processes critical for understanding fear learning. The central structure in the circuitry for the expression of fear CRs is the Amygdala (Cheng et al., 2006; Delgado, Nearing, Ledoux, & Phelps, 2008; Phelps, 2006b; Sehlmeyer et al., 2009). Other brain regions including the hippocampus, the ventromedial prefrontal cortex (vmPFC), and anterior insula (AI) have also been found to be involved in fear conditioning and extinction. Below, I will briefly discuss the significance of these regions in fear conditioning (for thorough reviews, see Ji & Maren, 2007; Sehlmeyer et al., 2009). Much of what we know about the basic neurobiological mechanisms of fear learning stems from classical conditioning in non-human animals. Studies have demonstrated that lesions of the amygdala inhibit many measures of conditioned and unconditioned fear responses (e.g., freezing, increases in blood pressure, heart rate changes, or fear-potentiated startle) (Fanselow, 1994; Kapp, Whalen, Supple, & Pascoe, 1992; LaBar & LeDoux, 1996). Recent evidence suggests that different parts of the amygdala play different functional roles. Two sub-regions within the amygdala are particularly important for fear conditioning: the basolateral complex (BLA), which includes the lateral (LA) and basal (B) nuclei, and the central nucleus (CE). In brief, relevant projections within the amygdala include the formation of CS-US association in the LA, which serves as an input to the CE. The CE in turn controls the expression of conditioned responses (CR) by sending information to other regions; for example (1) to the hypothalamus, which has been found to be important for mediating autonomic responses and in humans can be indexed by skin conductance response, and (2) to the brainstem which has been implicated in the regulation of behavioral expressions of fear (Davis, 1992; Fendt & Fanselow, 1999; Ji & Maren, 2007). While the neurophysiological substrates of fear conditioning have mostly been investigated in rodents using invasive procedures, the role of particular brain regions in fear conditioning has more recently been explored in humans using noninvasive functional brain imaging techniques. Research in humans using functional magnetic imaging (fMRI) for example, has provided support for the notion that amygdala plays a critical role in the acquisition and expression of conditioned fear (e.g., Büchel, Morris, Dolan, & Friston, 1998; Fullana et al., 2016; Sehlmeyer et al., 2009). Specifically, fMRI studies have shown increased blood oxygenation level-dependent (BOLD) signal (a proxy for neural activity) in the amygdala after a neutral stimuli has been associated with an aversive event; the

magnitude of this amygdala activation has also been shown to be correlated with SCRs as an indication of arousal to the CS event which is one the most common ways to study CR in humans (Büchel, Morris, Dolan, & Friston, 1998; LaBar & Phelps, 1998; Phelps, 2006). All of these results provide evidence of the amygdala's involvement in fear conditioning in humans. Studies have also shown that the vmPFC and hippocampus are important in fear conditioning. More specifically, it has been reported that activation in both these regions are especially important in extinction learning (Lang et al., 2009; Milad et al., 2007). One critical aspect of extinction is that it involves formation of new memories rather than removal of old fear memories that had been acquired during acquisition (for review see, Bouton, 2004). The anterior insula (AI) has also been implicated in fear conditioning as response to anticipation of fear, pain, processing of emotionally relevant stimuli, and arousal (Craig, 2009; Critchley, Mathias, & Dolan, 2001; Lamm & Singer, 2010).

1.3 'NOT ALL STIMULI ARE CREATED EQUAL'

In the early 1970's, the evolutionary preparedness theory was proposed by, Seligman (1971). This theory relied on the assumption that there is a functional significance of maintaining learned fears to a certain class of so called 'prepared stimuli' in order to avoid imminent threats. Empirical studies both in humans (Öhman, Fredrikson, Hugdahl, & Rimmo, 1976) and non-human primates (Mineka, Davidson, Cook, & Keir, 1984) have supported this notion. In humans, conditioned fear responses to fear-relevant stimuli (e.g., snakes and spiders) have consistently been found to resist extinction compared to fear-irrelevant stimuli (e.g., flowers and butterflies) (Öhman & Mineka, 2001). The idea that humans, like other animals, have a tendency to preferentially learn and retain some associations more readily than others has important implications for understanding social behavior. It provides a potential mechanism to understand the emergence and maintenance of social biases. Critically, preferential learning has been extended to angry facial expression (Öhman & Dimberg, 1996), and later to other social groups (e.g., race) (Mallan, Sax, & Lipp, 2009; Olsson, Ebert, Banaji, & Phelps, 2005). These observations suggest several research avenues to examine biased learning occurring in social interactions, which is the main focus of the current thesis. However, in order to better understand the potential links between aversive learning and social-cognitive processes, it is important to consider research on social biases.

1.3.1 Social Group Biases Based on Race

Social group biases are central aspect of human behavior. People often categorize others according to their race, gender, age, religion, political affiliation, or any other salient

social category. It could be argued that evolution may have biologically prepared people to be able to quickly identify others as belonging to their in-group or out-group. Across many different scenarios, people tend to prefer, and be more helpful towards, their in-group members compared to their out-group members. Further, it has been demonstrated that emotional expressions from in-group members are encoded faster and more accurately; empathic responding is stronger to in-group members; in-group members' faults are downplayed more than out-group members; and trust and cooperation are extended to in-groups more than to out-groups (for review, see De Dreu & Kret, 2016). In a study by Hein and colleagues (2010), soccer team belongingness was used as a social group manipulation to investigate costly helping behavior. Participants witnessed a fan of their favorite team (in-group member) or of a rival team (out-group member) experience pain, and were later asked to choose to help the fans by enduring physical pain themselves to reduce the others' pain. Their results showed increased AI response to when participants witnessed an in-group member, as compared with an out-group member, suffering pain, in line with studies showing empathy modulation (Hein, Silani, Preuschoff, & Batson, 2010). Importantly, they found increased nucleus accumbens (NAcc) activity (known to be involved in reward processing) and decrease in activation in AI (known to be involved in aversive experiences and empathy for others' discomfort), when participants were witnessing an out-group member suffering pain (Grit Hein, Silani, Preuschoff, Batson, & Singer, 2010). These results suggest that it can be more aversive to see an in-group member suffer and more rewarding to watch an out-group member suffer.

The flip side of in-group favoritism is out-group bias; decades of research have shown that social categorization often elicits stereotypes, prejudice, and discrimination towards out-group members (Brewer, 1999). One social category that has received much attention in research recently is racism and racial prejudice. Within the field of social psychology, prejudice refers more specifically to evaluations (i.e., attitudes) and emotional responses towards a group and its members (Amodio, 2014). Despite globalization and increased diversity, prejudices continue to be a core problem in intergroup conflict and discrimination. Our social norms have become more egalitarian, thus prejudices have become more difficult to detect, study, and ultimately change (Amodio, 2014). It has been suggested that because of pro-social norms, people may sometimes engage in self-regulatory methods (e.g., cognitive control) to adjust their behaviors. However, these types of biases often operate implicitly, and they can be activated, and influence judgments and behaviors without conscious awareness

(Greenwald & Banaji, 1995; Olson & Fazio, 2006). For the past few decades, scientists interested in race and prejudice have increasingly used methods of electrophysiology, functional magnetic resonance imaging (fMRI) and other physiological measures in an effort to identify the processes through which these biases are formed. These methods increase our understanding of how people perceive and evaluate race and how these processes relate to a range of social behaviors that can have consequential effects in social interactions. A network of regions have been proposed to be involved in race perception and racial prejudice involving amygdala, fusiform face area, insula, striatum and ventromedial frontal cortices (see reviews, Amodio, 2014; Kubota et al., 2012). Many of these brain regions have also been implicated in fear conditioning (see section 1.2.2). Most research in the field of race perception has used passive viewing paradigms to understand these processes. It is well understood that faces are processed quickly, however, research has shown that there are fundamental differences in processing racial out-group compared to in-group faces.

There are primarily two different theories that account for these types of differences. One is the "expertise" account, where it is argued that frequent contact with one's own racial group members leads to better discrimination of facial information of one's own racial group (compared to one's racial out-group) (Rhodes, Brake, Taylor, & Tan, 1989). This has been supported by studies showing enhanced activity in the 'fusiform face area' (FFA), for same-race relative to other-race faces (Golby, Gabrieli, Chiao, & Eberhardt, 2001). In previous work, this brain region has been associated with processing and individuating faces (Kanwisher, McDermott, & Chun, 1997; Rhodes, Byatt, Michie, & Puce, 2004) and general perceptual expertise (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999). Another theory suggests that in-group members are considered motivationally relevant or important, and therefore, lead participants to attend more to them (Hugenberg & Bodenhausen, 2004). In line with this account, it has been demonstrated that people are better at discriminating between faces belonging to members of an out-group when they display cues signaling threat (e.g., angry facial expressions) (compared to faces belonging to members of one's in-group) (Ackerman et al., 2006; Hugenberg & Bodenhausen, 2003). Researchers have also found that negative emotions from other races are detected more easily, and are more likely to be correctly categorized as negative even if facial expression itself is ambiguous (Hugenberg & Bodenhausen, 2003, 2004; Liu, Lin, Xu, Zhang, & Luo, 2015).

The amygdala has also been implicated in social perception and evaluation of other-race faces. Several studies have found increased amygdala activity when viewing images of racial out-group compared to racial in-group members (Hart et al., 2000). Furthermore, it has been found that this difference in amygdala activity correlates with implicit measures of racial bias (Kubota et al., 2012). Although these findings have led researchers to argue that the differences in amygdala activation during intergroup perception could be evidence of negativity—including disgust and fear—toward stigmatized groups (e.g., Harris & Fiske, 2006; Krendl, Macrae, Kelley, Fugelsang, & Heatherton, 2006) others have proposed that the amygdala may play a role in processing motivationally-relevant stimuli in general (e.g., Krendl, Kensinger, & Ambady, 2012; Stillman, Van Bavel, & Cunningham, 2015). For example, if race is the most salient social category, the amygdala may be responsive to members of groups who are stereotypically perceived as more threatening; however, if there are other bases for categorization, the amygdala may be responsive to members of groups that are currently relevant, such as group affiliation (i.e., in-group vs. out-group)(Van Bavel & Cunningham, 2009). Many findings mentioned here (regarding racial social groups) concern the basic mechanisms of social cognition that, to different extents, underlie other forms of biases, such as those based on gender, sexual preference and nationality. Given that social groups are perceived differently, and that group-belongingness can influence behavior, fear conditioning provides a powerful paradigm for exploring how such in-group and out-group learning biases are acquired and changed in a laboratory setting.

1.3.2 Fear Conditioning and Racial Biases

Studies using fear conditioning have investigated how learned fears towards out-group members are acquired, expressed, and changed. In one study by Olsson and colleagues (2005) participants were presented with unfamiliar faces of two black and two white male individuals with neutral expressions. During fear acquisition, one stimulus (CS+) from each stimulus category was paired with a mild electric shock (US). The other stimulus from each category (CS–) was presented without shock. During the extinction phase that followed, all CSs were presented again but no shocks were administered. SCRs were measured during both acquisition and extinction. The conditioned fear response (CR) was assessed as the differential SCR (CS+ minus CS–) from the same stimulus category (Black or White)[2], The

[2] In this way, this manipulation controls for differences in the emotional salience of stimulus categories as a confounding variable.

results indicated that the faces from a racial out-group were more readily associated with an aversive stimulus and were more resistant to extinction than the faces from an in-group (Olsson et al., 2005). However, the neural networks of fear learning to racial in-group and out-group members are unknown. We sought out to address this gap in our understanding of the neural networks of fear learning to in- and out-group members in **Study I** of this thesis.

1.3.3 Fear Conditioning and Social Dominance Hierarchy

Learning about threats in a social environment is a core component of human behavior and helps protect a person from dangerous stimuli or from stimuli one has previously had a negative experience with (Seligman, 1971). Social groups of humans and non-human animals are organized along dominance hierarchies, which help regulate access to food and mates, and help determine who constitutes a physical threat to whom. As modern humans we live in relatively egalitarian groups compared to our ancestors, but many of our social environments still contain situations where social hierarchy can be experienced as a threat (Sapolsky, 2004). Learning about the relative dominance of others, and to adaptively respond to their threat value, is central to one's health and survival (Adler et al., 2008; Kaplan & Manuck, 1999; Sapolsky, 2005). Previous research suggests that being a member of a high-status group will increase prejudice towards out-groups (Sidanius, Pratto, Martin, & Stallworth, 1991), increase in-group favoritism (Guimond, Dif, & Aupy, 2002), even when status is randomly assigned (Bettencourt, Talley, Benjamin, & Valentine, 2006). Much of past research has relied on static images with facial features and body postures as dominance cues (Marsh, Blair, Jones, Soliman, & Blair, 2009; Oosterhof & Todorov, 2008). However, social dominance rank can also be inferred by the probability that one (oneself or someone else) would win or lose an agonistic confrontation against in-group conspecifics (Rowell, 1974). In a social setting, confrontations are not only relevant to the people involved in the confrontation, but to everyone that might be observing as well. Indeed, across species, learning about the value of stimuli and situations through observing interactions is common (Olsson & Phelps, 2007), and reduces the risk of injury for the observer. Thus, the evidence to date suggests that observing confrontations can transmit critical information about social status.

As reviewed above (section 1.3.2), there are biases in threat responses towards social racial out-group members which tend to persist even in the absence of the threat (i.e., extinction) (Kubota et al., 2012; Olsson et al., 2005). In a recent study, this kind of learning was extended to faces associated with varying levels of experimentally induced social

dominance, showing that relatively higher levels of dominance of a target face was associated with more persistent threat learning (Haaker, Molapour, & Olsson, 2016). In this study, the researchers used a fear conditioning paradigm to investigate the acquisition and expression of threats associated with dominant and subordinate others learned through observation (Haaker et al., 2016). Participants first learned about the dominance rank of others by observing their dyadic confrontations. During subsequent fear learning, the dominant and subordinate others were both equally predictive of an aversive consequence (mild electric shock) to the participant. Importantly, the results from this study revealed that threats associated with a dominant person elicited stronger and more persistent learned threat responses as measured by physiological arousal and amygdala activity. These results suggest that aversive experiences associated with dominant others pose more of a threat and are more resist to extinction than those associated with subordinate others. It is well understood that there are learning biases based on race, and dominance, but it is highly relevant to understand how these racial and dominance based biases interact in modern heterogeneous societies. Building upon the example described above of how our direct experiences with someone else can alter our social interactions with our neighbor: how are our social interactions with the neighbor affected after observing an altercation they had with the mailman? Does whether they 'win' or 'lose' the altercation play a role in how we interact with them in the future? Furthermore, does the ethnicity of the neighbor play a more important role than their social status? We examined these questions in **Study II** of this thesis.

1.4 AGGRESSIVE ANTI-SOCIAL INTERACTIONS

Most of what I have covered so far highlights how we can use fear conditioning to understand how threats associated with members of different social groups are acquired, expressed and changed. One critical question that remains unknown is how the development of aversive learning is related to behavior in interactive contexts. As discussed in the previous sections of this thesis, there is strong evidence that there are differences in aversive learning based on whom (i.e., social group belongingness) we learn from, but the relationship between learned aversions and actual social behavior is still unknown. Although there are a few studies that suggest that our learned aversive experiences through conditioning can be related to anti-social behavior, there is not much research indicating the exact relationship between fear-conditioning and actual anti-social behavior. However, a few studies have shown that stimuli associated with aversions through conditioning can enhance aggression. For example, a study in rats has shown that conditioned stimuli increased the probability of fighting behavior when

given the opportunity (Hutchinson, Renfrew, & Young, 1971), and one study in humans (Fraczek, 1974) showed that painful experiences associated with a specific colored stimulus incited stronger aggression towards a peer.

With regards to aggression, historical records show that widespread aggressive and violent behaviors can be traced back to the earliest prehistoric times (DeWall, Anderson, & Bushman, 2011), suggesting that violence is an intrinsic part of human behavior. Aggressive behavior usually involves the intent to harm another person or a group; the violence can be in physical or non-physical form. Psychologists have been using laboratory experiments in which aggressive behaviors can be assessed in a safe and ethical fashion (Anderson & Bushman, 2002; Berkowitz, 1993). In these aggression paradigms the participants are typically allowed to act aggressively/anti-socially toward others. These paradigms have revealed that although people are reluctant to act aggressively towards others in general, some situational factors, such as competiveness, provocation, and aggressive cues (e.g., guns) can influence participants to aggress more often (Anderson and Bushman, 2002). Characteristics of the agent (e.g., personality traits, attitudes, genetic predispositions) are also important factors in contributing to aggressive behavior (Mischel & Shoda, 1995). Furthermore, aggression theories have also implicated that general arousal and aggressive cues combine to increase aggression (Anderson & Bushman, 2002; Bandura, 1973; Berkowitz & Lepage, 1967). That is to say, if an angered person's arousal is attributed to the frustrating or insulting source of the anger, subsequent aggression will increase (Rule & Nesdale, 1976). In contrast to how fear motivates escape, anger motivates aggression, which stops ongoing and/or deters future transgressions. Yet few, if any, studies have investigated how learning mechanisms known to underlie learned aversions through Pavlovian conditioning can help to explain interactive behavior (specifically anti-social behavior) in the normal population. Given the prevalence of anti-social behavior in everyday life (Cowie, Naylor, Rivers, Smith, & Pereira, 2002; Folger & Baron, 1996), understanding the underlying mechanisms is crucial to illuminate the dark side of human psychology, and to inform preventive strategies. Thus, we examined the role of fear conditioning in anti-social behaviors in **Study III**.

1.4.1 Maladaptive Aggression

Can a small initial offense cause a longstanding feud? If your neighbor takes your newspaper, will you punish them? Individuals have evolved adaptive responses to threats occurring in the environment, thereby enhancing the organism's probability of survival (see Ledoux, 2012). Certain environmental events that are perceived as threatening, frustrating, or provocative, by

someone can lead to responses such as avoidance or aggression —serving an adaptive function (Blanchard, Bassett, & Koshland, 1977). However, aggression can also be maladaptive—as a result of heightened reactivity of the threat response system (Blair, 2007). The brain's threat response system involves amygdala-hypothalamus-periaqueductal gray which is regulated by several other brain regions (e.g., frontal cortex) (Gregg & Siegel, 2001) an impaired ability to regulate this system can lead to an increased chance of initiating aggression rather than freezing or flight (Blair, 2004). Others have argued that acting out aggression can be rewarding; for example it has been demonstrated that aggressive behavior is associated with activity in the brain's reward system in response to aggressive social stimuli (de Quervain et al., 2004; Decety, Michalska, Yuko, & Lahey, 2009; Golden et al., 2016). For example, punishing defectors in an economic game activated the dorsal striatum, which has been implicated in the processing of rewards (de Quervain et al., 2004).

Importantly, studies have shown that maladaptive behavior can be self-reinforced; it has been demonstrated in several species including rats (Brown, Martin, & Morrow, 1964) monkeys (Kelleher & Morse, 1968; McKearney, 1969) and humans (Renner & Tinsley, 1976) that learned contingencies between a behavior and an outcome is difficult to unlearn (even if they are maladaptive). For example, studies in squirrel monkeys and rats have demonstrated that performing an action that results in the reception of painful electric shocks is counter-intuitively reinforced rather than inhibited after the behavior-outcome contingencies change, indicating what is called 'self-punitive behavior' (Melvin and Anson, 1969; Morse et al., 1967). This indicates that behaviors with aversive outcomes (i.e., maladaptive) can be self-reinforced (Seymour, Singer, & Dolan, 2007). We hypothesize that self-reinforcing punishment in social interactions can be a factor to why vendettas emerge. To our knowledge, no experimental research has shown how and why vendettas emerge in dyadic interactions. In **Study IV** we examined how people reciprocated punishments when doing so was maladaptive (i.e., always resulting in receiving same number of shocks back), in the absence of economic or other instructed incentives other than received punishment.

In summary we have discussed the role of aversive learning in everyday life, and how fear conditioning offers a way to study how we acquire and express fears towards others based on *who* we learn about. We also provide evidence for how social aversive learning could offer insight into social interactive behaviors. We also discussed, how behaviors with aversive outcomes (i.e., maladaptive) could sometimes be self-reinforced. In the different studies of this thesis we have made an attempt to unravel some of the core components of how aversive

learning based on who learn *from* and *about* can influence social interactive behaviors, and how.

2 AIMS

The general aim of this thesis was to investigate what role aversive learning has in social interactions. Specifically, we sought to examine whether aversive learning differs based on whom we learn from and about, and if so, whether these differences in aversive learning are reflected in interactive social behaviors. We also aimed at understanding how reciprocal punishments are initiated and upheld in social interactions.

2.1 STUDY I AIM

- To examine the neural activity associated with aversive learning to members of in-groups and out-groups (i.e., racial), and whether behavior and brain activity related to this learning would impact future interactions with new in-group and out-group members.

2.2 STUDY II AIM

- To examine the interaction between knowledge about individuals' social dominance rank and ethnic group belonging during aversive learning (i.e, fear conditioning).

2.3 STUDY III AIM

- To investigate whether learned aversions cause retaliatory behavior in a social interactive context.

2.4 STUDY IV AIM

- To examine whether maladaptive punishment (i.e., self-punitive) occurs in dyadic interactions, and how such behaviors are initiated and maintained.

3 METHODS

3.1 PARTICIPANTS

All participants (N=193) gave their written consent before participation and were naive to the purpose of the experiment. The procedures were executed in compliance with relevant laws and institutional guidelines, and were approved by the Regional Ethical Review Board of Stockholm. Participants were paid for their participation at the conclusion of the experiments.

3.2 EXPERIMENTAL STIMULI

For all of our experiments (except for **Study IV**) we used different faces as stimuli. For **Study I,** we used Black faces expressing a neutral facial expression from the NimStim facial database (Tottenham et al., 2009), and in **Studies I** and **III**, we used White faces from the Radboud Faces Database (RaFD) (Langner et al., 2010). We also used the Karolinska Directed Emotional Faces for the White faces (KDEF; Lundqvist, Flykt & Öhman, 1998) and RaFD for the Middle Eastern faces in both side-view and front-view in **Study II**. In **Study IV** we used schematic figures of faces.

For **Studies I** and **II** we created a Social Interactive Task (SIT) (modified from the original "cyberball" game; Williams & Jarvis, 2006), in order to measure social interactive behavior. We devised this task to simulate real social interactions. In **Study I**, target faces consisted of one Black face (NimStim) and one White face (Radboud), while in **Study II** the target faces consisted of one Middle Eastern face (RaFD) and one White face. We created three additional distractor faces by morphing one of each category (Black or Middle Eastern and White) using a morphing program (Squirlz Morph: www.xiberpix.com). The new distractor faces consisted of 75%, 50%, and 25% similarity to the out-group face (See Figure 1).

Figure 1. *Illustration of the interactive environment during the SIT in Study I.* Participants were presented with one Black and one White face and three distractor (racially-morphed) faces. Participants were asked to pass the ball to each one of the other players. This task was also used in **Study II** but with White and Middle Eastern faces.

In **Study I**, the visual display was presented via MR-compatible LCD video goggles [NordicNeuroLab (NNL), Bergen, Norway] connected to a PC running Presentation (Version 14, Neurobehavioral Systems, Inc., www.neurobs.com). In **Studies II-IV** the stimuli presentations were managed using E-Prime 2.0 (Schneider et al., 2002). Shock deliverance was controlled by monopolar DC-pulse electric stimulation (STM200; Biopac Systems Inc., www.biopac.com). In all studies, before the start of the experimental sessions, the intensity of the electric shock (US) delivered to the wrist was calibrated for each participant. In a standard 'work-up' procedure, shock intensity was gradually increased until participants appraised it as uncomfortable, but not painful. Before the procedure, the shock electrode was attached to the participants' right wrist and a conductive gel (Signa, Parker) between the electrodes and the skin.

3.2.1 Electric Shocks

In our experiments (**Studies I-II**) we used electric shock as an aversive stimulus (US), which is a common human fear conditioning procedure (Critchley, Mathias, & Dolan, 2002; Olsson et al., 2005; Phelps & LeDoux, 2005). We also used electric shocks to investigate aggressive behavior in a social context in **Study III** and the underlying mechanisms of retaliatory behaviors in **Study IV**. The term aggression refers to a wide spectrum of behaviors, in the psychological literature, it is defined as any behavior intended to harm another person that wants to avoid being harmed (For review, see Anderson & Bushman, 2002). Many

researchers have previously operationalized aggressive behavior by measuring the intensity of electric shocks administered to another individual (e.g., Bailey & Taylor, 1991; Bushman, 1995; Giancola & Zeichner, 1995). Others have used monetary or point penalties, verbal attacks, and negative evaluations as proxies for aggression (Bailey & Taylor, 1991; Check & Dyck, 1986; Dougherty, Bjork, Marsh, & Moeller, 2000). In **Studies III & IV** we used administration of shocks, because we were interested in studying the effects of learning based on physically aversive consequences. Additionally, shock administration was a convenient method in our experimental paradigm. Finally, the use of shocks enabled us to avoid adding motivational factors, such as monetary gains and losses, to the paradigm.

3.3 PSYCHOPHYSIOLOGICAL MEASUREMENTS

3.3.1 Skin Conductance Response

The skin conductance response (SCR), represents one of the oldest measurements in the history of psychophysiological research. SCRs reflect the phasic increase in skin conductance that occurs in response to physiologically arousing stimuli. Both negative and positive stimuli evoke SCRs, which is modulated both by stimulus novelty, intensity and by attentional processes (Öhman, 1979). The SCR reflects the change in electrodermal conductance on the palmar surface of the hands and the sole of the feet through eccrine sweat glands, which increases due to sympathetic arousal (Öhman, 1971). In humans, SCRs are usually measured by a pair of electrodes attached to the distal phalanges of the index and middle finger. In the context of fear conditioning, SCR is the most commonly used index of CR and is commonly inferred from increased SCRs in the presence of a CS that is predictive of the US (the CS+) as compared to SCRs to the control stimulus (CS-). The difference between these measures is referred to as differential SCR. There are many advantages in using SCRs; it is a non-intrusive measure, and interestingly, SCR changes have been linked to changes in amygdala activity during fear conditioning (e.g. Cheng, Knight, Smith, & Helmstetter, 2006; Knight, Nguyen, & Bandettini, 2005; LaBar et al., 1998).

3.4 FUNCTIONAL MAGNETIC RESONANCE IMAGING

Functional Magnetic Resonance Imaging (fMRI) is a neuroimaging technique with both clinical and research applications. There are many advantages to use fMRI relative to other functional neuroimaging techniques for research applications. One advantage of using contrast fMRI over forms of functional imaging, is that BOLD contrast fMRI is non-

invasive and does not expose the subject to radiation. In research we often use fMRI in awake humans engaged in a different kinds of behavioral and cognitive tasks. The activity measured in fMRI is the blood oxygen level dependent (BOLD) response. When particular region in the brain is in use (i.e., neurons firing), blood flow to that region also increases (Brown, Perthen, Liu, & Buxton, 2007; Logothetis & Wandell, 2004; Logothetis, 2008). One can then detect the changes in the BOLD signal over time, which constitutes an indirect measure of electrical neural activity (Goense et al. 2012, Logothetis & Wandell, 2004). Although fMRI has been very useful for examining the neural correlates of many different tasks, it also has limitations. One common methodological consideration in fMRI is head movement, which influences both the quality of the scans and the functional inferences that can be made from them (Khanna, Altmeyer, Zhuo, & Steven, 2015; Yendiki, Koldewyn, Kakunoori, Kanwisher, & Fischl, 2014). This is important because head movement causes the voxels to be mis-aligned with the corresponding brain tissue across the whole scanning session. In order to avoid this problem as much as possible, we asked our participants to keep their head as still as possible during the experiment, and we used extra padding in the radio frequency (RF) head coil to further help to keep their head in place. For the remaining unavoidable movements, we used SPM 8 (Statistical Parametric Mapping, Wellcome Center for Neuroimaging), which uses a motion correction algorithm as part of the preprocessing of the data. We did not need to exclude any participants due to excessive head movement in the fMRI study (**Study I**).

3.4.1 Parametric Analysis

Based on studies in both in fear conditioning (Büchel et al., 1998; LaBar et al., 1998) and race perception (Hart et al., 2000; Kubota et al., 2012) showing important time-dependent effects, we wanted to examine possible changes of activity in a given region. For example, previous studies have found temporally graded amygdala responsivity in both animal and human populations (Quirk et al., 1997). Therefore, in addition to categorical conditioned responses (CRs) (i.e., overall activity), we also examined differences in parametric responses linearly changing over time. The parametric modulation allowed us to examine possible interactions between stimulus and time that are absent in categorical analyses of the mean responses. Both categorical and parametric effects were analyzed separately on group level in a 2 × 2 full factorial design including the parameter estimates of each CS separated on two factors: CS type (CS+ and CS−) and race (Black and White). We defined the interaction contrast from the 2 × 2 factorial design as (Black CS+ minus Black CS−) >

(White CS+ minus White CS−), thus significant voxels containing neuronal populations that are specifically involved in fear learning to Black faces as compared to White faces.

3.4.2 Psychophysiological Interaction Connectivity Analysis

In **Study I**, we evaluated changes in the degree of neural connectivity using psychophysiological interaction (PPI) analysis (Friston et al., 1997). A PPI measures the strength of connectivity between a seed region and other brain regions changing with the experimental variable. Thus, a significant PPI indicates that the connectivity between one region and another increases significantly with the experimental or psychological variable (Friston et al., 1997). We specifically used a generalized PPI toolbox (gPPI; http://www.nitrc.org/projects/gppi), which compared to standard PPIs in SPM, allows for the interaction of more than two task conditions in the same PPI model and improves model fit, specificity to true-negative findings, and sensitivity to true-positive findings (McLaren et al., 2012). In our analysis we examined connectivity changes between a pre-defined seed region (i.e., amygdala) and the rest of the brain. The interaction terms in the PPIs were computed by multiplying the time series from the psychological regressors with the physiological variable. To examine the effect of the interaction terms, activity within the amygdala was regressed on a voxel-wise basis against the interaction with the physiological and psychological variables serving as regressors of interest. In our experiment, the individual CR Black > CR White contrast images were entered into separate second-level 2 (CS) × 2 (Race) ANOVAs for the left and right amygdala to determine whether there were any CS × Race interactions on functional connectivity. The resulting activation maps from this analysis reflect the functional connectivity between amygdala and other brain regions that was significant for fear learning to out-group and in-group faces. As such, this connectivity analysis allowed us to examine the connectivity between amygdala and other brain regions during learning and expression of learned fear to out-group and in-group faces.

4 OVERVIEW OF STUDIES

In the following section, I will briefly summarize the main findings and conclusions of the individual studies. For details, I refer the reader to the full manuscripts attached at the end of this thesis.

4.1 STUDY I: NEURAL CORRELATES OF BIASED SOCIAL FEAR LEARNING AND INTERACTION IN AN INTERGROUP CONTEXT

4.1.1 Study I Background and Rationale

In light of previous research which has demonstrated that a learned association between a fearful experience and a member of a social out-group (i.e., ethnic) is more resistant to change than a learned association between a fearful experience and a member of an in-group (Olsson et al., 2005; Kubota et al., 2012), we wanted to investigate the neural activity underlying these types of aversive learning biases. Additionally, we wanted to examine whether the neural activity was predictive of future interactive behavior to new in and out-group members. Previous research has identified a network of regions involved in the acquisition and expression of conditioned fear (LaBar and LeDoux, 1996; Phelps and LeDoux, 2005). In addition to the amygdala–a key region involved in the acquisition of fear—the vmPFC, hippocampal complex, and insula have also been implicated in fear conditioning. However, a different line of research has identified how racial out-groups are perceived in passive viewing paradigms, and has revealed a network of brain regions with many overlapping regions similar to the fear conditioning neural network (e.g., amygdala, and insula). We hypothesized that brain regions linked to both fear conditioning and race perception jointly contribute to aversive learning to in- and out-group members. We were specifically interested in time-dependent changes that were associated with the CR to in-and out-group faces. Moreover, we also investigated the connectivity between the amygdala and the rest of the brain. We predicted that a learning bias would involve changes in activity over time. Based on previous research on threatening stimuli (Anderson and Phelps, 2001, Hariri et al., 2003 and Morris, 1998), we expected increased connectivity between the amygdala and the visual cortex during perception of conditioned out-group faces.

4.1.2 Study I: Results and Conclusions

We demonstrated that activity in brain regions previously linked to conditioned fear, and perception of individuals belonging to racial or stigmatized out-groups, jointly contribute to differential brain activity and biased behavior based on race. During Acquisition of CRs, we found increased activity in the amygdala, AI, and ACC to both Black and White faces.

During Extinction we found enhanced activity in the dorsal AI for CR to Black vs. White faces (see Figure 2), which is indicative of aversive subjective experiences (Craig, 2009) and processing of stigmatized individuals (Harris and Fiske, 2006). Further, this activity might reflect an attempt to down-regulate aversive experiences during confrontation with conditioned out-group faces. Our analysis over time during Extinction revealed activity increasing over time in left amygdala, bilateral fusiform gyrus, and right hippocampus to Black compared to White faces (i.e., across CS + and CS −). This is in line with previous passive viewing paradigms (Kubota et al., 2012). Our connectivity analysis revealed an enhanced coupling between the amygdala and the fusiform gyrus during the acquisition and expression of learned fear to Black faces. The enhanced connectivity between the amygdala and the FFA in our results is consistent with the notion that the amygdala guides the visual system to prioritize encoding of visual information that best predict aversive events or threats (Anderson and Phelps, 2001). Importantly, both the amygdala and AI predicted interactive behavior. Specifically, individual variability in the preferential passing to the White vs. Black co-player, was predicted by an anti-Black learning bias observed in the dorsal AI (See Figure 3).

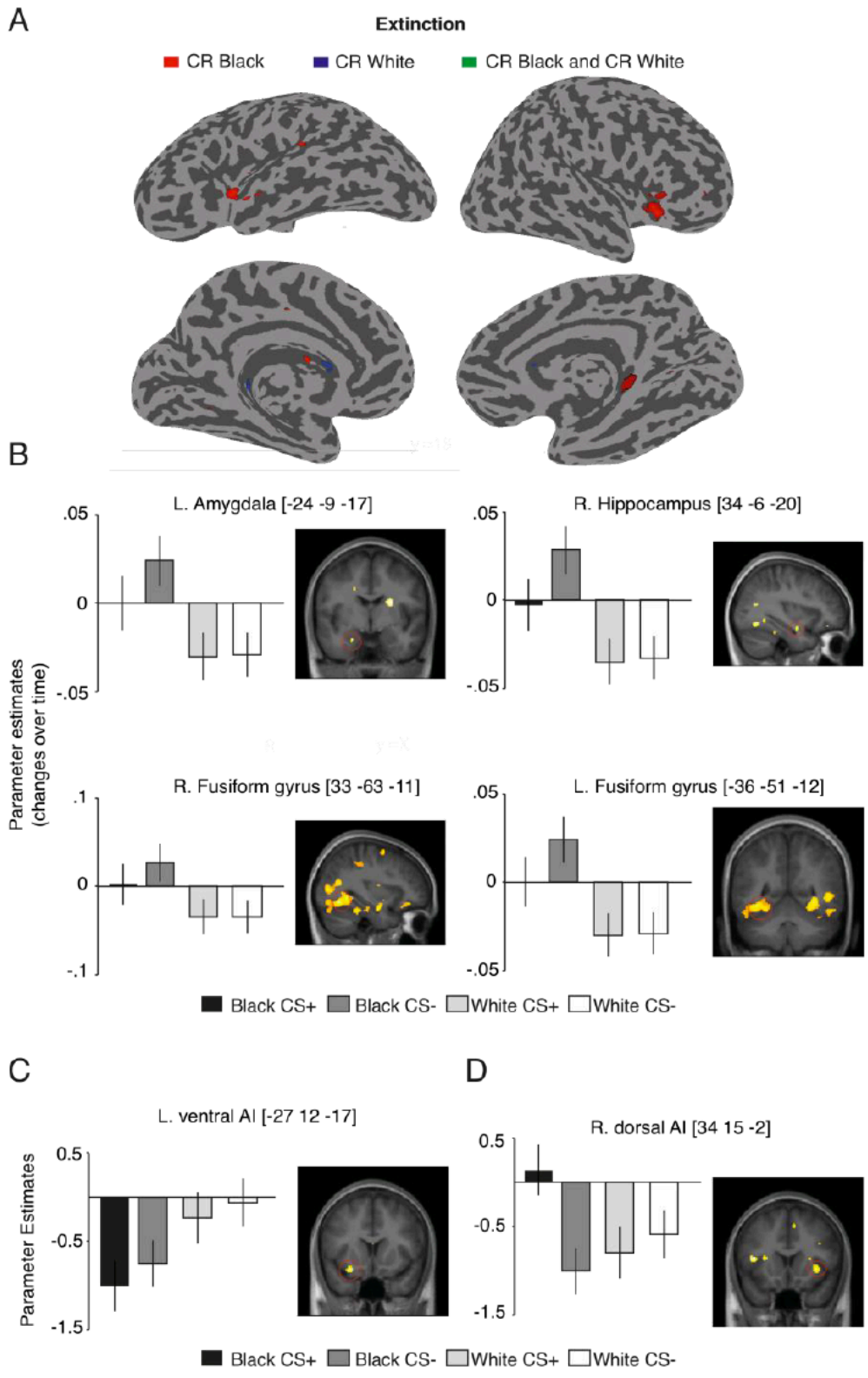

Figure 2. *Brain activity to CR Black vs. CR White faces during Extinction.* (A) Overview of the brain regions during Extinction that are significant for CR to Black faces (in red), and CR to White faces (in blue), and conjunction for both CR Black and CR White faces (in green). For display purposes only, the activation map was displayed at a threshold of $p < .001$ (uncorrected for multiple comparisons). (B) Bar plots shows activation in the left amygdala, right hippocampus, and right and left fusiform gyrus for the contrast (Black > White) during Extinction stage indicating changes in activity over

time. (C) Bar plot shows overall activation in the left ventral AI for the contrast (Black > White) during Extinction stage. (D) Bar plots shows overall activation in the right dorsal AI for the contrast (CR Black > CR White) during Extinction stage. The reported coordinates are in the MNI space. Error bars denote ± SEM, and activation maps are displayed at $p_{\text{uncorrected}} < .01$ for display purposes only.

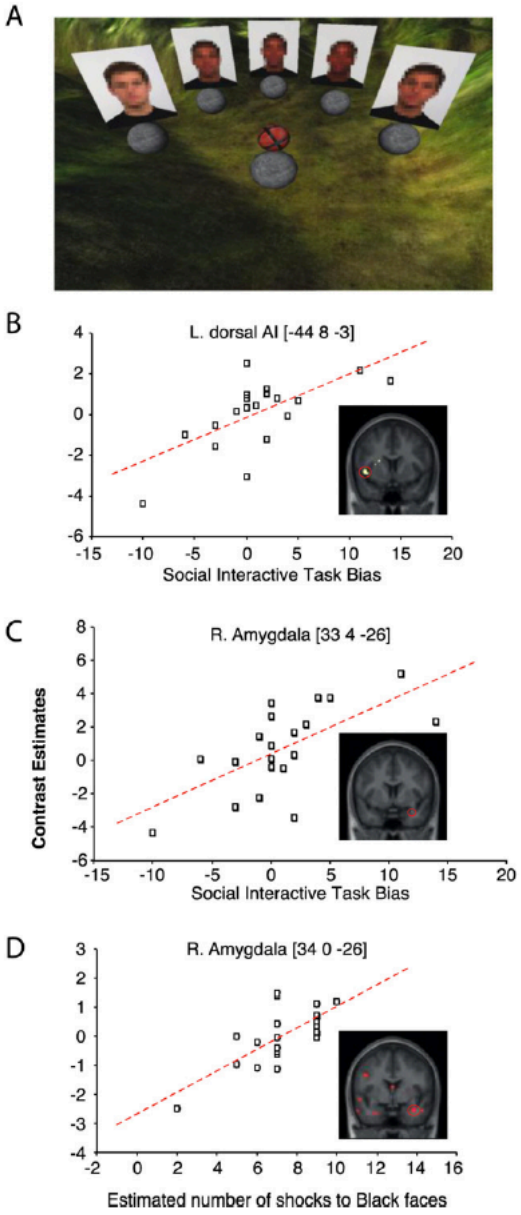

Figure 3. *Brain activity predicts behavior.* (A) Illustration of the interactive environment during the SIT. Participants were presented with one Black and one White face and three distractor (racially-morphed) faces. Participants were asked to pass the ball to each one of the other players. (B) Significant relationship between activity in left dorsal AI for CR Black > CR White during the Extinction stage, and the strength of anti-Black SIT bias (i.e., passing less often to the Black faces). (C) Significant relationship between activity in right amygdala in response to shock to Black faces, and the strength of anti-Black SIT bias. This relationship was not observed for shock to White faces. (D) Significant relationship between activity in right amygdala for CR Black > CR White during Acquisition and the number of estimated passes to Black faces. For illustration purposes, results are displayed at uncorrected significance ($p < .001$) thresholds.

4.2 STUDY II: THE RELATIONSHIP BETWEEN SOCIAL DOMINANCE THREAT AND RACIAL BIASES

4.2.1 Study II Background and Rationale

Humans and non-human animals alike respond especially strong to certain threats. For example, research has shown that learned fear towards particular classes of stimuli, such as snakes, angry faces, and faces of individuals belonging to social out-groups is more persistent (resists extinction) compared to neutral stimuli (e.g., Olsson, Ebert, Banaji, & Phelps, 2005). Recently, similar response biases have been extended to faces associated with a high relative level of social dominance (Haaker et al., 2016). Social groups are organized along dominance hierarchies, which regulate access to food and mates, and help determine who poses a physical threat to whom. Therefore, learning about other individuals' relative dominance and adaptively responding to their threat value is central to one's health and survival (Adler et al., 2008; Kaplan & Manuck, 1999;Sapolsky, 2005). We investigated the interaction between learned dominance and social out-group (ethnicity) threats to understand if dominance hierarchy knowledge changes out-group biases. In our paradigm, participants learned the social hierarchy of two North-European (NE) and two Middle Eastern (ME) individuals by watching their dyadic face-to-face confrontations. Later, participants were conditioned to fear the same faces. We were specifically interested in investigating whether indirectly learned relative dominance of others influences direct aversive learning (fear conditioning) towards the same individuals. We hypothesized that in-group members would be considered more dominant and threatening when the participants were passively observing confrontations and threats were not directed towards them per se. However, during Pavlovian conditioning, we hypothesized that we would find increased skin conductance responses (SCRs) resisting extinction to dominant out-group members, as they were directly threatening and relevant to the participant.

4.2.2 Study II Results and Conclusions

Firstly, we show that the participants correctly learned the relative dominance of the individual faces regardless of ethnicity. This is consistent with previous research, demonstrating that a confrontation between others is sufficient to learn their social hierarchy dominance rank (Jones et al., 2011). Secondly, participants were biased in choosing NE faces to be more dominant than ME faces during the observation of the hierarchy phase. This is in line with the account that in-group members are more relevant and, possibly, more attended to. For example studies have shown that emotions from in-group members are encoded faster and more accurately, and empathic responding is stronger to in-group members (in-group

members are liked better, their faults are downplayed more, and trust and cooperation are extended to in-groups more than to out-groups) (For review, see De Dreu & Kret, 2016). Importantly, during direct fear learning, ethnic out-group faces elicited stronger and more persistent learned threat responses as measured by physiological arousal. This in line with studies showing for example that angry out-group faces are more threatening (Ackerman et al., 2006b; Fox et al., 2000). Interestingly, our findings also suggest that measures of explicit of biases do not reveal any racial biases, but that implicit measures do. It could be that people are driven by egalitarian beliefs and political correctness, but implicit measures reveal underlying racial biases. Accordingly, we did not find that learning biases influenced behavior in our social interactive task, participants passed the ball equally to all of the co-players regardless of ethnic background. Our results improve our understanding of how learning through observing conflicts between others produces learning about their relative dominance, and how this learning is impacted by social group belongingness. Finally, we demonstrate how this observational learning subsequently affects learning through direct, aversive, experiences (Pavlovian conditioning).

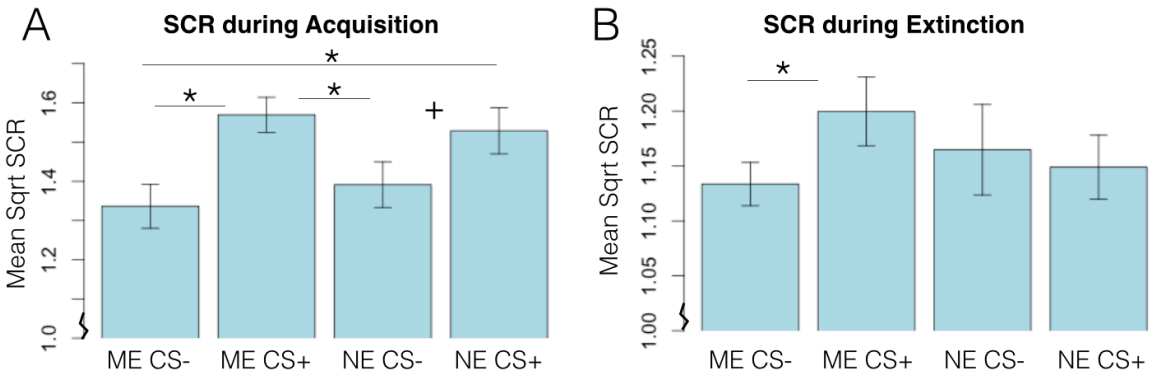

Figure 4. *Skin conductance results.* The amplitude of SCRs is shown in microsiemens. Fear elicited enhanced SCRs to CSs during (A) Acquisition and again during (B) Extinction. Error bars indicate standard deviation (SEM). Asterisks indicates a statistically significant difference $p < .05$. Plus sign indicates a marginally statistically significant difference $p = .066$.

4.3 STUDY III: AVERSIVE LEARNING INFLUENCES ANTI-SOCIAL BEHAVIOR

4.3.1 Study III Background and Rationale

We wanted to investigate the role of aversive learning in retaliatory behavior in social interactive context. One of the oldest ideas in psychology is that human behavior is governed by likes and dislikes formed about others. Indeed, people prefer to spend time with individuals they learned to like, and try to avoid, or aggress toward those they learned to be afraid of or dislike. Although adaptive in some situations, anti-social behaviors, such

as avoiding, aggressing, or punishing others, can be detrimental in social interactions. In two experiments (n = 35, n = 34), we used a modified fear-conditioning paradigm to investigate the role of aversive learning in retaliatory behavior in social context. Participants first completed an initial aversive learning phase in which the pairing of a neutral conditioned stimulus (CS; i.e., neutral face) with a naturally aversive unconditioned stimulus (US; i.e., electric shock) was learned. Then they were given an opportunity to interact with the same faces again, during a Test phase, with the possibility to administer shocks (i.e., administer 0–2 shocks). We hypothesized that if learned aversions would influence retaliatory behavior, participants would administer more shocks to the faces that were previously paired with most number of shocks (i.e., CS++, and CS+).

4.3.2 Study III Results and Conclusion

In this study we examined the role of aversive learning in retaliatory behavior. We used classical fear conditioning with an added Test phase allowing for social interactive behavior. This provided an opportunity to examine the strength of aversive learning about specific individuals and its influence on subsequent social interactive behavior with the same individuals. In two separate experiments, we demonstrate how previously learned aversions influence future retaliatory behavior. In both experiments, we found that participants showed the largest SCRs to the faces paired with one or two shocks during Acquisition, indicating successful aversive learning. These findings were corroborated with US Expectancy ratings in Experiment 2. Most importantly, we demonstrated that participants administered more shocks to the individuals delivering the most number of shocks when the opportunity was given during the subsequent Test phase. Our findings are consistent with results on evaluative conditioning, showing that repeated pairings of CSs and USs influence subsequent evaluative *judgments* of the CSs (De Houwer et al., 2001; Baeyens et al., 2005; Walther et al., 2005), and classical fear conditioning studies showing that learned fear associations influence behavior (e.g., pathogenesis of anxiety disorders; Mineka and Zinbarg, 2006). However, our results go beyond these findings and show that pairings of CS and aversive US enhanced retaliatory *behavior* toward another person in a social context.

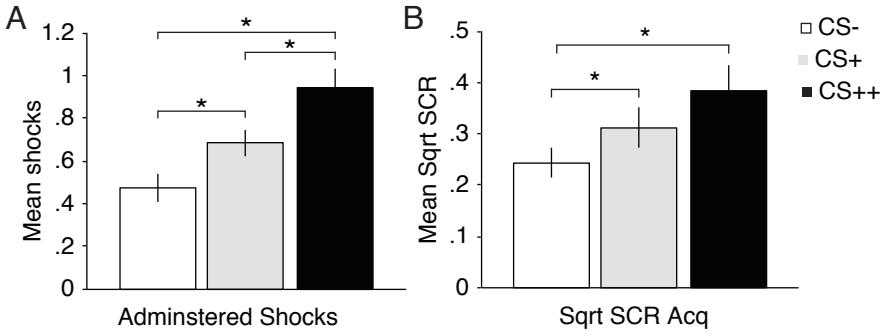

Figure 5. *Administration of shocks and SCR results Experiment 1*. (A) Average number of administered shock to each CS during the test phase, showing a linear increase in punishing behavior as a function of received shocks. (B) The amplitude of SCRs in microSiemens, showing strongest SCRs to CS+ relative CS- during Acquisition. Error bars indicate standard deviation (SEM). Asterisks indicates a statistically significant difference $p < .05$.

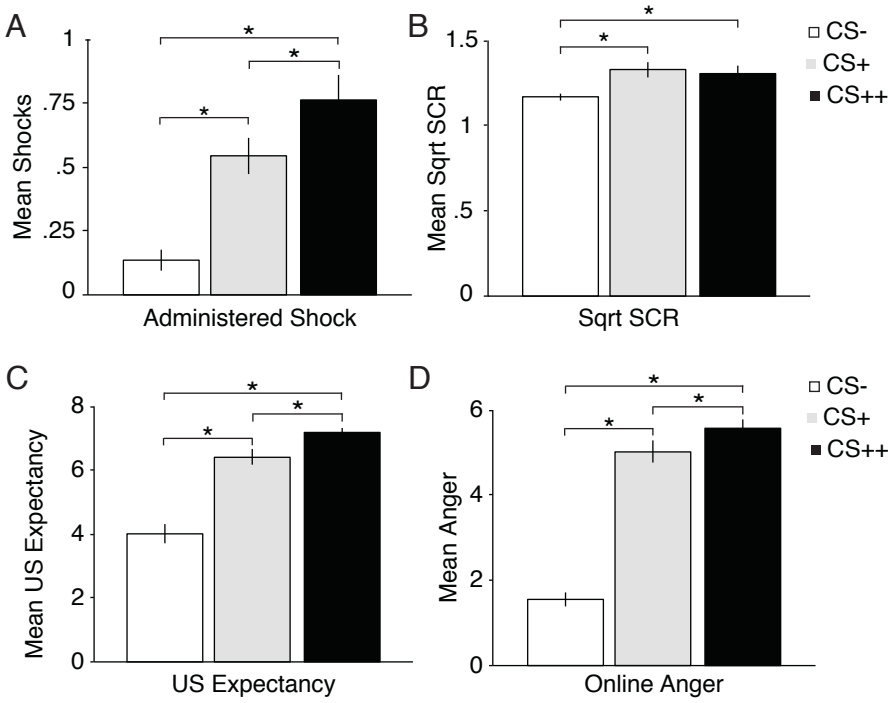

Figure 6. *Administration of Shocks, SCR, US Expectancy and Anger results Experiment 2.* (A) Average number of administered shocks to each CS during the Test phase, showing a linear increase in punishing behavior as a function of received shocks. (B) The amplitude of SCRs is shown in microSiemens, showing stronger SCRs CS++, CS+ relative to CS- during Acquisition. (C) Mean US Expectancy ratings to CS++, CS+ and CS- during Acquisition. (C) Mean online Anger ratings to CS++, CS+ and CS- during Acquisition. Error bars indicate standard error of the mean (SEM). Asterisks indicate a statistically significant differences $p < .05$.

4.4 STUDY IV: RECIPROCAL PUNISHMENT IN DYADIC SOCIAL INTERACTIONS

4.4.1 Study IV Background and Rationale

In this study, we wanted to examine the basic mechanisms underlying retaliatory behaviors as can be seen in varying forms of conflicts in dyadic social interactions which can lead to vendettas and vicious cycles of aggression. Certain environmental events that are perceived as threatening, frustrating, or provocative, by someone can lead to responses such as avoidance or aggression—which are usually adaptive (Blanchard et al., 1977). However, aggression can also be maladaptive—as a result of heightened reactivity of the threat response system (Blair, 2007). Importantly, studies have shown that maladaptive behavior can be self-reinforced; it has been demonstrated in several species including rats (J. S. Brown et al., 1964) monkeys (Kelleher & Morse, 1968; McKearney, 1969) and humans (Renner & Tinsley, 1976) that learned contingencies between a behavior and an outcome are difficult to unlearn (even if they are maladaptive). This indicates that behaviors with aversive outcomes (i.e., maladaptive) can be self-reinforced (Seymour et al., 2007). We hypothesize that self-reinforcing punishment in social interactions can be a factor in why vendettas emerge. To investigate the role of these potential mechanisms in retaliatory exchanges experimentally, we devised a new dyadic interactive paradigm in which participants could chose whether to administer one, many, or no shocks to an alleged co-participant. Separately, as an application of our experimental results to a real world setting, we examined online behavior of commenters in dyadic interactions in an Internet forum. We expected that people would engage in vicious cycles of verbal exchanges online resulting in verbal feuds analogous to what we observe in our experimental paradigms.

4.4.2 Study IV Results and Conclusions

This study reveals that individuals punish an anonymous co-player despite the fact that their punishing behavior resulted in receiving exactly the same punishment back from the 'co-player' and lead to a potentially long-standing vendetta. These findings describe a form of self-punitive behavior previously documented in animals. The participants' tendencies to administer shocks were exacerbated when the co-player initiated punishment, indicating that a small initial offense motivated punishing behavior over time. The best predictor for administration of shocks was the number of shocks participants received on a previous trial. The Internet data corroborated the experimental data; in online dyadic interactions the level of aggressive comments previously received predicted the level of aggressive comment sent back. The results from this study improve our understanding of the processes underlying

common destructive social phenomena where people punish others even at a cost to themselves.

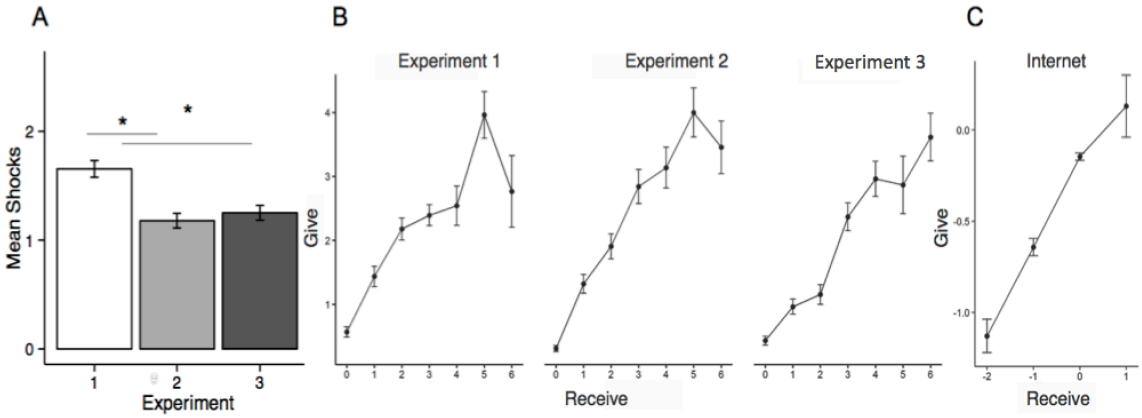

Figure 7. *Reciprocation of punishment.* (A) Mean number of administered shocks in Experiments 1, 2 and 3. (B) The mean number of shocks participants received on a previous trial predicted how many shocks they gave to the co-player in Experiments 1-3. Error bars represent the SEM. *$p < 0.05$. (C) Internet data: The degree of aggressive comments received predicted the degree of aggressive comments reciprocated.

5 DISCUSSION

An improved grasp on the impact of aversive experiences in social contexts is critical for the understanding of human interactive behavior. The overall aim of the studies presented in this thesis was to investigate how aversive experiences influence behavioral and brain responses in social contexts. Our general approach was to develop simple experimental paradigms to capture the underlying mechanisms of aversive experiences and social interactions. In **Study I**, we examined the neural basis of fear conditioning to racial in-and out-group members, and found that differences in brain responses were related to differences in actual interactive behavior at a later time. In **Study II**, we found that when participants observed others' interactions, they learned the social dominance hierarchy of racial in-group members better than racial out-group members. However, during direct aversive encounters (Pavlovian conditioning) with the same individuals, participants showed stronger and more persistent learned threat responses to racial out-group members. **Study III** showed that aversive learning through fear conditioning influenced future retaliatory behaviors, and **Study IV** found that individuals punished others even when maladaptive (i.e., received the same number of shocks back), which led to a vicious cycle of aggression. The participants' tendencies to administer shocks were exacerbated when the co-player initiated the punishment, indicating that a small initial offense motivated punishing behavior over time. We also examined a real-world example of this behavior using comment exchanges from an Internet forum, which corroborated our experimental results. Here, I will briefly discuss some general theoretical and methodological questions and conclusions gleaned from these studies.

Examining the neural correlates of fear learning of racial in- and out-group members in **Study I**, revealed activity in two overlapping networks of brain regions, one which has previously been linked to conditioned fear, and the other to the perception of racial out-group members. We observed that differential brain activity in these two networks was related to biased behavior based on race. In accordance with our predictions we found enhanced activity in the dorsal AI for CR to Black vs. White faces during Extinction, which is indicative of aversive subjective experiences (Craig, 2009) and processing of stigmatized individuals (Harris & Fiske, 2006). This area has also been functionally connected to the brain's cognitive control network (Dosenbach et al., 2007) that is implicated in monitoring and control of conflicts between emotional responses and egalitarian motives (Amodio, Devine, & Harmon-Jones, 2008). In line with previous studies on passive viewing of racial out-group faces, we also found enhanced activity in the amygdala, fusiform gyrus, and hippocampus to out-group faces during Extinction. The enhanced activity in these regions to

out-group faces, is consistent with research showing enhanced activity related to processing emotionally salient stimuli (e.g., unpleasant) (Kober, Barrett, & Joseph, 2008; Sabatinelli, Bradley, Fitzsimmons, & Lang, 2005; Straube, Mentzel, & Miltner, 2006), and threatening faces (Vuilleumier, Armony, Driver, & Dolan, 2003). It is also important to note that facial stimuli in our study reflected greater threat value resulting from direct aversive learning experiences to the faces, compared to passive viewing paradigms.

Although most of our predictions were based on differential CRs to out-group faces during Extinction in line with previous studies (measuring SCRs) (e.g., Olsson et. al., 2005), our results also demonstrated race dependent differences of brain activity during the *acquisition* of conditioned fear. Specifically, we found a greater time-dependent CR effect in the amygdala for White (vs. Black) faces during Acquisition. The amygdala activity increased over time to the Black CS− and White CS+ faces, whereas responses to the Black CS+ and White CS– did not change over time. We believe that the relatively stronger differentiation of White faces during the acquisition could reflect a general in-group advantage in individuating and recognizing faces (Golby et al., 2001; Malpass & Kravitz, 1969). Along the same reasoning, a weaker individuation between the two Black out-group faces might have contributed to a greater generalization of fear response to the unsafe Black CS+ and the safe Black CS− (Dunsmoor, White, & LaBar, 2011; Vervliet, Kindt, Vansteenwegen, & Hermans, 2010). An alternative explanation of these results is that the safe Black (CS−) and the unsafe White (CS+) stimuli both violated the race stereotype (Blair, Ma, & Lenton, 2001), resulting in enhanced amygdala responses over time.

Additionally, we examined connectivity changes between a pre-defined seed region (i.e., amygdala) and the rest of the brain. Our connectivity analysis with the amygdala revealed functional connectivity between the amygdala and the fusiform gyrus associated with CRs to out-group faces. Interestingly, this enhanced connectivity (i.e., between the amygdala and the fusiform gyrus) during the learning and expression of learned fear to Black faces occurred both during Acquisition and Extinction stages. This enhanced amygdala—FFA coupling in our results is consistent with the idea that the amygdala helps guide the visual system to prioritize encoding of visual information that can predict aversive events or threats (Anderson and Phelps, 2001). Further, it is also possible that the threat value affected the coding of Black and White faces differently after pairing with an aversive event (i.e., an electric shock), which is reflected in the enhanced connectivity between the amygdala and the fusiform gyrus for the CR to Black faces in our results. This is in support of previous studies showing that out- vs. in-group faces are better remembered when they are potentially threatening (Ackerman et al., 2006).

Another important contribution was that we showed that brain activity in the racial-fear-learning-bias network was related to the participants' discriminatory interactions with novel out-group members on a later day. Specifically, we found that the AI activity observed for CRs to Black vs. White faces during Extinction predicted subsequent social interactions with unfamiliar Black and White individuals, in our modified version of Cyberball (i.e., Social interactive task; SIT). Individual variability in preferential passing to the White vs. Black co-player, was predicted by an anti-Black learning bias observed in the dorsal AI. The link between the AI and a discriminatory bias is indicative of research describing the AI as important in the processing of stigmatized individuals (Harris and Fiske, 2006), and decision making during uncertainty (Lamm & Singer, 2010; Singer et al., 2009).

The link between the participants' brain activity (reflecting biased learning and extinction responses to out-groups) and their subsequent interactive behavior indicated that those who showed a learning bias towards Black individuals also tended to display more discriminatory behaviors. This observation might be characteristic of the individuals, could reflect the aversive learning experience itself, or might reflect a combination of both. Unfortunately, with our current data we cannot differentiate between these alternative explanations. Nevertheless, our findings are the first to identify the neural mechanism of fear learning biases towards out-group members, and its relationship to interactive behavior. Moreover, our findings provide important clues towards understanding the mechanisms underlying biases between social groups.

In **Study II**, we followed up the study of group biases seen in **Study I** in a more dynamic and interactive social context. We examined the interaction between knowledge about individuals' social dominance rank and ethnic group belonging during aversive learning (i.e., fear conditioning). We wanted to understand how the interaction of others' relative social dominance and ethnicity impacted learning and behavior. Critically, we showed that the participants correctly learned the relative dominance of the individual faces (i.e., dominant and subordinate), regardless of ethnicity. Specifically, our findings indicate that despite the fact that there was a hierarchy learning bias to *in-group* faces during observational hierarchy learning, during direct Pavlovian fear learning, the participants showed stronger and more persistent learned threat responses to *out-group* faces. Our results suggest that different theoretical accounts could help to explain why dominance and ethnicity interact in different ways during indirect learning of dominance and direct learning of threat. We propose two different possible accounts based on the 'relevancy' and 'additive' threat value of the facial stimuli. According to the 'relevancy' account, in-group members that are more likely to be competing over resources in day-to-day situations (Flinn, Geary, & Ward, 2005) are

perceived as a greater threat than out-group individuals. The 'additive' account, suggests that the threat value of posing aggressively (i.e., high in relative dominance in our paradigm) is added to the threat value of out-group members in accordance with previous studies (e.g. Olsson et al., 2005). The additive account finds further support in findings showing that although people often find it more difficult to distinguish between ethnic out-group members (compared to ethnic in-group members) (Anthony, Copper, & Mullen, 1992; Ostrom & Sedikides, 1992). However, the opposite is true when out-group members display threatening facial expressions, suggesting that out-group members do not become relevant or 'important' unless they pose a threat to the person (Ackerman et al., 2006b).

Interestingly, the participants that had a high IAT score (indicating an implicit bias against Middle Eastern people) also showed higher SCRs to Middle Eastern faces compared to North European faces during acquisition. However, the participants did not display any explicit ethnic biases (e.g., self-reports and interactive ball tossing) even though their learning was clearly biased. That is, dominance learning was facilitated for in-group faces, but when faces were paired with a direct aversive consequence (during Pavlovian conditioning) dominant out-group faces acquired a stronger threat value as evidenced in more persistent threat responses. The lack of explicit biases likely reflects egalitarian beliefs and political correctness.

In both **Studies I** and **II** we show that individuals associate racial out-group members more easily with an aversive stimulus, but it remains to explain why such biases exist. Previous research have suggested that prepared fear learning is a biologically evolved learning mechanisms that regard certain natural categories of stimuli as prepared to be associated with an aversive outcome (Öhman & Dimberg, 1996; Seligman, 1970). From an evolutionary perspective, it means that there is an advantage to have a prepared learning bias towards threat-relevant stimuli such as snakes and spiders, angry faces and socially dominant others because it appropriately signals threat.

However, understanding threats towards racial out-groups is unlikely to have the same evolutionary basis. Unlike the social sensitivity to dynamic dominance hierarchies seen across taxa, social categorizations based on racial differences among humans occurred relatively recently in human evolutionary history (Molnar, 1998); therefore it is unlikely that responses to racial differences have a strong genetic bases (Cosmides, Tooby, Fiddick, & Bryant, 2005). Race in itself is unlikely to explain the types of biases that we observe in our studies. However, even if a genetic predisposition to preferentially learn from in-group individuals do exist, this is likely to be influenced by socially and culturally acquired attitudes towards racial groups (Fiske & Taylor, 2008). Further, humans might have evolved a more

general preparedness to fear others who were dissimilar to them and not belonging to their social group because such individuals were more likely to pose a threat (Hamilton, 1964; Manson & Wrangham, 1991). If a general preparedness to fear out-groups did indeed evolve, then today's out-group members (as categorized by physical differences) could merely reflect a general out-group bias and therefore produce the conditioning effects we observe in our studies.

In **Study I**, we found that individual variability in preferential ball passing to the White vs. Black co-player was predicted by an anti-Black learning bias observed in the dorsal AI; which indicated that implicit measures can reveal biases in future behaviors. However, examining the relationship between our implicit measures (i.e., SCRs) and future social interactive behavior in **Study II** did not reveal a significant relationship. Although threat-elicited amygdala response has commonly been found to be correlated with threat-elicited SCRs (Cheng et al., 2006; Dunsmoor, Martin, & LaBar, 2012; Wood et al., 2014), the amygdala appears to mediate important aspects of the peripheral emotional response to threat that SCRs do not (Wood et al., 2014). Perhaps using fMRI in **Study II** would have revealed brain responses predicting social interactive behaviors that were not observed with SCRs. Further, the two studies were designed differently, one critical difference was that in **Study II** we had an observation of dominance learning phase, which could have influenced the learning and expression of fear responses during Pavlovian conditioning.

An important limitation of **Studies I** and **II** is that we only included White participants, which limits the generalizability of our conclusions to other social out-groups. Although previous behavioral studies have shown similar results for other categories of social out-groups (Bavel, Packer, & Cunningham, 2011; Navarrete et al., 2012), further research needs to examine the neural mechanisms of learning biases to other out-groups to better understand the generalizability of these biases. Further, in both **Studies I** and **II**, male and female participants were included in our experimental samples, whereas only male faces served as CSs. Therefore, female participants belonged to an additional out-group (i.e., gender), which could have influenced the results. However, we carefully examined the role of gender in our results, which did not reveal any differences based on gender, which is in agreement with previous findings on a race related learning bias (Golkar, Björnstjerna, & Olsson, 2015; Navarrete et al., 2012; Olsson et al., 2005).

In **Study III**, in two experiments, we demonstrated how previously learned aversions influenced future retaliatory behaviors. Participants showed largest SCRs to the faces paired with one or two shocks during Acquisition, which demonstrated successful aversive learning. Specifically, we showed that the participants administered more shocks to the individuals delivering the most number of shocks when the opportunity was given during the subsequent Test phase. However, we did not find a significant relationship between physiological arousal and retaliatory behavior. This finding is in contrast to previous research showing that general arousal and aggressive cues combine to increase aggression (Bandura, 1973; Berkowitz & Lepage, 1967). Interestingly, our results show that anger toward CSs paired with shocks and general aggressive traits increased the likelihood of administering shocks to the aversively reinforced faces (i.e., CS+), but not to the non-aversively reinforced face (i.e, CS−). This is in accordance with studies that show that anger is a key motivator of aggression (e.g., assaulting, attacking, kicking; Cuddy, Fiske, & Glick, 2007; Frijda; Roseman, Wiest, & Swartz, 1994; Rule & Nesdale, 1976). Examining of the relationship between aggressive trait, trial by-trial anger, and the administration of shocks revealed that trait aggression mediated the impact of trial-by-trial anger towards, and the administration of shocks given to, co-players. These findings are in line with theories of aggression describing an interaction between person and situational factors influencing aggressive behaviors (Anderson and Bushman, 2001). There are several possible explanations for why the participants' retaliatory behaviors increased toward the aversively reinforced CSs. The participants might have felt unfairly/unjustly treated, and that punishing the other person would 'balance out' the situation (i.e., retribution). Another explanation is that the participants might have attempted to alter the future behavior of the norm transgressor (co-player) by teaching them that acting unfairly does not pay (i.e., deterrence). Both retribution and deterrence are known motivators of human punishment (Crockett, Özdemir, & Fehr, 2014). The participants also could have punished the co-player simply to harm them out of spite (de Quervain et al., 2004), because anger-induced punishment can give pleasure to the punisher (Berkowitz, 1993).

In **Study IV**, we investigated retaliatory exchanges experimentally in a novel dyadic interactive paradigm in which the participants chose whether to administer one, many, or no shocks to an alleged co-participant. The participants' decisions to administer shocks resulted in receiving exactly the same number of shocks back from the alleged co-player. In this way, the optimal decision (i.e., avoid *getting* shocks) was to avoid *giving* shocks to the co-player. In contrast to **Study III**, where punishing someone based on previous aversive experiences could be regarded as adaptive (e.g., deterrence and retribution), in **Study IV** we wanted to

examine punishing behavior when it was maladaptive. In three experiments we showed that the participants administered shocks to a co-player, even when resulting in receiving same number of shocks back (maladaptive) and under the direct (but implicit) control of the participant.

Why do people engage in maladaptive reciprocal punishments? There are several potential explanations for why people might engage in maladaptive punitive behavior, which can lead to extended vicious cycles of aggression and possible vendettas. Our results lead us to three possible explanations: First, people might feel justified to punish the other person simply on the basis of a 'the other person started it' standpoint (Stillwell, Baumeister, & Del Priore, 2008). Exploring the difference between punishing behavior when the participants were provoked (i.e., Experiment 1) vs. unprovoked (i.e., Experiment 2), we found that, although provocation significantly increased overall punishment initially, in the long-run punishing behavior did not seem to differ. This indicates that the initial discomfort and other aversive emotions that initially provoked the participant might have dissipated over time. The participants might have based their behavior on a moment-to-moment basis rather than an overall estimation of what had happened from the beginning. Indeed, our analysis investigating the relationship between the number of received shocks on a given trial and the number of shocks administered on the previous trial strengthens this argument: decisions to administer shocks were directly influenced by how many shocks they had received from the co-player on the previous trial. This is in accordance with studies that show 'eye-for-an-eye' counter-punishments to restore equity, where high punishments are reciprocated with high punishments (Stillwell et al., 2008). In line with the analytic strategy used in **Study III**, we pooled data across the two experiments of **Study IV**. Interestingly, this also revealed that trait aggression and trait anxiety predicted the administration of shocks in general, strengthening the conclusion that personality factors influence aggression (Anderson, Buckley, & Carnagey, 2008). We believe this paradigm has advantages over previous ones examining maladaptive aggression and vendettas, mainly because our paradigm is devoid of explicitly instructed motivations to punish and behave aggressively, uses dyadic interactions, and examines the self-reinforcing nature of vendettas. Economists have used public goods games to explain possible mechanisms of punishment and the emergence of vendettas (Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis & Normann, 2008). Only a few studies have examined punishment in dyadic repeated interactions (compared to group interactions) and have shown that vendettas are rare in this context (Dreber, Rand, Fudenberg, & Nowak, 2008; Fehl, Sommerfeld, Semmann, Krambeck, & Milinski, 2012; Fehl, van der Post, & Semmann, 2011). However, as we know from many instances in life vendettas do occur in

dyadic interactions (e.g., bullying). Further, these economic games use monetary gains and losses, suggesting that vendettas occur within the realm of economic interests. However, economic public goods games do not help understanding conflicts that emerge in absence of economic factors in dyadic interactions. In social psychological paradigms that investigate punishment there are instructed motivations to punish others, such as requirements to punish a peer for incorrect answers, to improve others' task performance, or competitiveness to avoid loss or receiving shocks (Baron & Eggleston, 1972;Rule & Nesdale, 1974; Taylor, Gammon, & Capasso, 1976). Cover stories are usually used in these paradigms to ensure that there is a specific purpose to punish or act aggressively towards another person (in order to measure the participants' propensity to do so). For these reasons, past research is limited where the basic processes of reciprocal punishment that lead to vicious cycles of aggression and vendettas in social dyadic interactions is concerned. The current models of reciprocal aggression and vendettas do not show that maladaptive vendettas occur nor explain how these types of aggressive acts persist and are reciprocated between strangers online (Dreber et al., 2008; Fehl et al., 2012, 2011).

One possible limitation in **Studies III** and **IV**, and other social psychology paradigms examining aggression in social contexts, is that the participants might feel the need to comply with what they believe is the appropriate behavior in the laboratory (Zizzo, 2010). However, to offset this problem, the participants had the option to give zero shocks in both studies (**Study III** and **IV**), and the possibility to avoid seeing the face (by removing it) in **Study III**, which are both non-aggressive options. If the participants decided to administer shocks, the explanation is less likely to be related to experimental demand characteristics, as giving no shocks is the more socially desirable behavior. To support this, in the post-experimental interviews, participants reported retributive motives for administering shocks, and none of the participants reported that they chose to administer shocks in order to adhere to the demands the experiment. However, we cannot fully rule out this possibility, and future studies should take this into account.

As with most experimental paradigms specifically investigating aggression/anti-social behaviors, and experimental models of complex phenomena in general, the external validity of such experiments (i.e., probability to generalize to situations outside the laboratory) is limited (see Tedeschi and Bond, 2001). One caveat to our experimental paradigms is that giving and receiving shocks in a laboratory setting is different from real-life aggressive behaviors. However, others have argued that there is direct and indirect support for aggression paradigms in many research domains (e.g., Anderson and Bushman, 2001).

Furthermore, in **Study IV**, we used interactions from an Internet discussion forum (resembling dyadic interactions in our paradigm) where we showed that the basic mechanisms for maladaptive reciprocal punishments could explain similar behaviors outside of the laboratory. Although we demonstrate that people engage in maladaptive punishment, we cannot really say why such behaviors should have evolved, and what the basis for these types of punishments are. Future studies should examine other real-world scenarios to obtain a more comprehensive understanding of why irrational punishments occur. Additionally, the results of **Study III** in this thesis indicated that the participants reciprocated punishments to those who were associated with aversive events when given the possibility. Future research can make use of this paradigm to examine whether this is true in real-life scenarios where direct punishment is not always a suitable option (e.g., workplace).

The fear-conditioning paradigm that is used in this thesis to investigate the role of social aversive learning in social contexts extends previous findings showing that our aversive experiences are processed differently based on whom we learn from. There is a need for more research on fear conditioning and extinction processes using other social groups (i.e., other than racial) to gain additional understanding of the neural, physiological, and behavioral mechanisms as well as the generalizability of our results. One critical question is whether racial biases that we observe in our experiments can be inhibited. That is, are there any intervention strategies that can diminish the chances that aversive experiences have dysfunctional consequences? A few studies have investigated intervention strategies in the laboratory by using tasks in which images of racial out-group members are repeatedly paired with positive images and appetitive responses or with counter-stereotypical concepts. Others have suggested more control-based strategies, where people learn to attend to specific cues (in interactions with out-group members) in order to down-regulate their biases. Although these interventions have been effective to some degree, whether the positive changes that result from them are maintained outside of the laboratory where racial prejudices and stereotypes are often reinforced by, for example, media exposure remains unknown. Although these strategies might not eliminate prejudice altogether, they may help reduce some behaviors that may be detrimental to victims of prejudice (i.e., out-group members).

In conclusion, it is clear that aversive learning plays an important role in social interactions. Specifically, we find that aversive learning in social context differs (e.g., as indicated by neural and psychophysiological differences) based on whom we learn about; and aversive learning also influences anti-social behaviors (e.g., behavior in a ball tossing game, or punishment behavior). Further, it is also evident that people punish others even when it

comes at a direct physical cost to themselves, indicating a self-reinforcing punishing behavior that can lead to vicious cycles of aggression.

6 ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my main supervisor **Andreas Olsson**, and co-supervisor **Henrik Ehrsson**. **Andreas Olsson**, special thanks for the continuous support during my doctoral studies. Thank you for encouraging me to explore new ideas and for your invaluable guidance.

I am also thankful to my main co-authors, **Armita Golkar, Björn Lindström**, and **Jan Haaker** whose work has been essential for the studies that are included in this thesis. Thank you to the many members, past and present, of the **Emotion Lab** with whom I have had the pleasure to work with; **Ida Selbing**, **Irem Undeger**, **Philip Pärnamets, Jonathan Yi**, **Lisa Espinosa, Simon Jaangard, Olof Hjorth** and my summer students **Laura Kress**, **Tatjana Michel**, and **Fiammetta Dede Brugo**. Thank you for the stimulating discussions and making going to work exciting every day. Importantly, thank you for listening to 'We all live in a social world' throughout these five years, for the *many* laughs, headstands in the office, and for your friendship and support.

I would like to express my gratitude to the staff at the MR-Research Center in Solna, especially to **Rouslan Sitnikov** and **Jonathan Berrebi**, for technical support with the fMRI experiment.

I would also like to thank my family: my parents **Mojdeh Rastin** and **Jamshid Molapour**, my sister **Shabnam Molapour,** my brother **Ario Molapour** and my **Mamani (Mehri Sedighe)** for your love and support, always.

I additionally want to express my gratitude to colleagues, friends, and family, who have been extending their love, friendship, and support in many different ways; here is a list in no particular order: **Malin Lundahl**, **Hesho Khalaf**, **Ivo Todorov**, **Björn van der Hoort**, **Giovanni Gentile, April Johnston**, **Sara Dalai**, **Martin Bellander, Danya Porada**, **Christina Regenbogen**, **Phydilla** and **Bill Gimbal**, **Bob** and **Judy Berger**, **Nick** and **Sarah Berger**, **Joe**, **Chuck**, **Toby**, **Constance**, **Alison** and **Patrick Christiana**.

A very special thanks goes out to **Dr. Ezequiel Morsella**, without whose motivation and encouragement I would not have considered a graduate career in research.

Dr. Christopher Berger, sorry for not telling you more about Sweden before I brought you here (not really though). Thank you for being an inspiration and for being perfect. **Maebee Barksdale**, you make me happy everyday.

7 REFERENCES

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., Schaller, M. (2006b). They all look the same to me (unless they're angry): from out-group homogeneity to out-group heterogeneity. *Psychological Science*, *17*(10), 836–40.

Adler, N., Singh-Manoux, A., Schwartz, J., Stewart, J., Matthews, K., & Marmot, M. G. (2008). Social status and health: A comparison of British civil servants in Whitehall-II with European- and African-Americans in CARDIA. *Social Science & Medicine*, *66*(5), 1034–1045.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670–682.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: the role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology*, *94*, 60–74.

Anderson, and C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology*, *53*, 27–51.

Anderson, C. a, & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology*, *53*, 27–51.

Anderson, C. A., Buckley, K. E., & Carnagey, N. L. (2008). Creating Your Own Hostile Environment: A Laboratory Examination of Trait Aggressiveness and the Violence Escalation Cycle. *Personality and Social Psychology Bulletin*, *34*(4), 462–473.

Anthony, T., Copper, C., & Mullen, B. (1992). Cross-Racial Facial Identification: A Social Cognitive Integration. *Personality and Social Psychology Bulletin*, *18*(3), 296–301.

Bailey, D. S., & Taylor, S. P. (1991). Effects of alcohol and aggressive disposition on human physical aggression. *Journal of Research in Personality*, *25*(3), 334–342.

Bandura, A. (1973). *Aggression: A Social Learning Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Baron, R. a., & Eggleston, R. J. (1972). Performance on the "aggression machine": Motivation to help or harm? *Psychonomic Science*, *26*(6), 321–322.

Bavel, J. J. Van, Packer, D. J., & Cunningham, W. A. (2011). Modulation of the Fusiform Face Area following Minimal Exposure to Motivationally Relevant Faces: Evidence of In-group Enhancement (Not Out-group Disregard).

Berkowitz, L. (1993). Pain and aggression: Some findings and implications. *Motivation and Emotion*, *17*(3), 277–293.

Berkowitz, L., & Lepage, A. (1967). Weapons As Aggression-Eliciting Stimuli. *Journal of Personality and Social Psychology*, *7*, 202–207.

Bettencourt, B. A., Talley, A., Benjamin, A. J., & Valentine, J. (2006). Personality and

aggressive behavior under provoking and neutral conditions: A meta-analytic review. *Psychological Bulletin*, *132*(5), 751–777.

Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain and Cognition*, *55*(1), 198–208.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, *11*(9), 387–392.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841.

Blanchard, E. B., Bassett, J. E., & Koshland, E. (1977). Psychopathy and Delay of Gratification. *Criminal Justice and Behavior*, *4*(3), 265–271.

Bouton, M. (2004). Context and behavioral processes in extinction. *Learning & Memory*.

Bouton, M. E., & Moody, E. W. (2004). Memory processes in classical conditioning. In *Neuroscience and Biobehavioral Reviews* (Vol. 28, pp. 663–674).

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and Temporal Modulation of Extinction: Behavioral and Biological Mechanisms. *Biological Psychiatry*.

Brewer, M. B. (1999). The Psychology of Prejudice: Ingroup Love and Outgroup Hate? *Journal of Social Issues*, *55*(3), 429–444.

Brown, G. G., Perthen, J. E., Liu, T. T., & Buxton, R. B. (2007). A primer on functional magnetic resonance imaging. *Neuropsychology Review*.

Brown, J. S., Martin, R. C., & Morrow, M. W. (1964). Self-punitive behavior in the rat: Facilitative effects of punishment on resistance to extinction. *Journal of Comparative Physiological Psychology*, *57*(1), 127–33.

Büchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998). Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron*, *20*(5), 947–57.

Bushman, B. J. (1995). Moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of Personality and Social Psychology*, *69*(5), 950–960.

Check, J. V. P., & Dyck, D. G. (1986). Hostile aggression and type A behavior. *Personality and Individual Differences*, *7*(6), 819–827.

Cheng, D. T., Knight, D. C., Smith, C. N., & Helmstetter, F. J. (2006). Human amygdala activity during the expression of fear responses. *Behavioral Neuroscience*, *120*(6), 1187–1195.

Cosmides, L., Tooby, J., Fiddick, L., & Bryant, G. A. (2005). Detecting cheaters. *Trends in Cognitive Sciences*, *9*(11), 505–506.

Cowie, H., Naylor, P., Rivers, I., Smith, P. K., & Pereira, B. (2002). Measuring workplace

bullying. *Aggression and Violent Behavior*, *7*(1), 33–51.

Craig, A. D. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural Activity in the Human Brain Relating to Uncertainty and Arousal during Anticipation. *Neuron*, *29*(2), 537–545.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2002). Fear conditioning in humans: The influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron*, *33*(4), 653–663.

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology. General*, *143*(6), 2279–86.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The bias map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience*, *15*, 353–375.

De Dreu, C. K. W., & Kret, M. E. (2016). Oxytocin Conditions Intergroup Relations Through Upregulated In-Group Empathy, Cooperation, Conformity, and Defense. *Biological Psychiatry*.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science (New York, N.Y.)*, *305*(5688), 1254–8.

Decety, J., Michalska, K. J., Yuko, A., & Lahey, B. B. (2009). Atypical empathic responses in adolescents with aggressive conduct disorder: a functional MRI investigation. *Biological Psychology*, *80*(2), 203.

Delgado, M. R., Nearing, K. I., Ledoux, J. E., & Phelps, E. a. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron*, *59*(5), 829–38.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*(1), 145–167.

DeWall, C. N., Anderson, C. A., & Bushman, B. J. (2011). The general aggression model: Theoretical extensions to violence. *Psychology of Violence*, *1*(3), 245–258.

Diamond, D. M., & Rose, G. M. (1994). Does associative LTP underlie classical conditioning. *Psychobiology*, *22*(4), 263–269.

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., … Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(26), 11073–8.

Dougherty, D. M., Bjork, J. M., Marsh, D. M., & Moeller, F. G. (2000). A comparison between adults with conduct disorder and normal controls on a Continuous Performance Test: Differences in impulsive response characteristics. . *The Psychological Record* .

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–51.

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology*, *89*(2), 300–5.

Dunsmoor, J. E., White, A. J., & LaBar, K. S. (2011). Conceptual similarity promotes generalization of higher order fear learning. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *18*(3), 156–60.

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, *1*(4), 429–38.

Fehl, K., Sommerfeld, R. D., Semmann, D., Krambeck, H.-J., & Milinski, M. (2012). I dare you to punish me-vendettas in games of cooperation. *PloS One*, *7*(9), e45093.

Fehl, K., van der Post, D. J., & Semmann, D. (2011). Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters*, *14*(6), 546–551.

Fendt, M., & Fanselow, M. S. (1999). The neuroanatomical and neurochemical basis of conditioned fear. *Neuroscience and Biobehavioral Reviews*, *23*(5), 743–60.

Fiske, S., & Taylor, S. (2008). *Social cognition: from brains to culture*. New York, NY: McGraw-Hill.

Flinn, M. V., Geary, D. C., & Ward, C. V. (2005). Ecological dominance, social competition, and coalitionary arms races: Why humans evolved extraordinary intelligence. *Evolution and Human Behavior*.

Folger, R., & Baron, R. (1996). Violence and hostility at work: A model of reactions to perceived injustice. *American Psychological Association, Washington, DC*, 51–85.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial Expressions of Emotion: Are Angry Faces Detected More Efficiently? *Cognition & Emotion*, *14*(1), 61–92.

Fraczek, A. (1974). Informational role of situation as a determinant of aggressive behavior. *Determinants and Origins of Aggressive Behavior (The Hague, The Netherlands: Mouton)*, (25-30).

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage*, *6*(3), 218–29.

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.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*(6), 568–73.

Giancola, P. R., & Zeichner, A. (1995). Construct validity of a competitive reaction-time aggression paradigm. *Aggressive Behavior*, *21*(3), 199–204.

Globisch, J., Hamm, A., & Esteves, F. (1999). Fear appears fast: temporal xourse of startle reflex potentiation in animal fearful subjects. *Psychophysiology*, *36*, 66–75.

Golby, A. J., Gabrieli, J. D., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–50.

Golden, S. A., Heshmati, M., Flanigan, M., Christoffel, D. J., Guise, K., Pfau, M. L., … Russo, S. J. (2016). Basal forebrain projections to the lateral habenula modulate aggression reward. *Nature*, *534*(7609), 688–692.

Golkar, A., Björnstjerna, M., & Olsson, A. (2015). Learned fear to social out-group members are determined by ethnicity and prior exposure. *Frontiers in Psychology*, *6*.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27.

Gregg, T. R., & Siegel, A. (2001). Brain structures and neurotansmitters regulating aggression in cats: implications for human aggression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *25*(1), 91–140.

Guimond, S., Dif, S., & Aupy, A. (2002). Social identity, relative group status and intergroup attitudes: When favourable outcomes change intergroup relations...for the worse. *European Journal of Social Psychology*, *32*(6), 739–760.

Haaker, J., Molapour, T., & Olsson, A. (2016). Conditioned social dominance threat: Observation of others' social dominance biases threat learning. *Social Cognitive and Affective Neuroscience*.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*(1), 1–16.

Hamm, A. O., & Weike, A. I. (2005). The neuropsychology of fear learning and fear regulation. *International Journal of Psychophysiology*.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. *Psychological Science*, *17*(10), 847–53.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly

helping. *Neuron*, *68*(1), 149–60.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat. *Psychological Science*, *14*(6), 640–643.

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, *15*(5), 342–345.

Hutchinson, R. R., Renfrew, J. W., & Young, G. A. (1971). Effects of long-term shock and associated stimuli on aggressive and manual responses. *Journal of the Experimental Analysis of Behavior*, *15*(2), 141–166.

Ji, J., & Maren, S. (2007). Hippocampal involvement in contextual modulation of fear extinction. *Hippocampus*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–11.

Kaplan, J. R., & Manuck, S. B. (1999). Status, stress, and atherosclerosis: the role of environment and individual behavior. *Annals of the New York Academy of Sciences*, *896*, 145–161.

Kapp, B. S., Whalen, P. J., Supple, W. F., & Pascoe, J. P. (1992). Amygdaloid contributions to conditioned arousal and sensory information processing. In *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction.* (pp. 229–254).

Kelleher, R. T., & Morse, W. H. (1968). Stimuli. maintained shocks'. *Journal of the Experimental Analysis of Behavior*, *11*(6), 819–838.

Khanna, N., Altmeyer, W., Zhuo, J., & Steven, A. (2015). Functional Neuroimaging: Fundamental Principles and Clinical Applications. *The Neuroradiology Journal*, *28*(2), 87–96.

Kleiman, T., Hassin, R. R., & Trope, Y. (2014). The control-freak mind: stereotypical biases are eliminated following conflict-activated cognitive control. *Journal of Experimental Psychology. General*, *143*(2), 498–503.

Kober, H., Barrett, L., & Joseph, J. (2008). Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage*.

Krendl, A. C., Kensinger, E. A., & Ambady, N. (2012). How does the brain regulate negative bias to stigma? *Social Cognitive and Affective Neuroscience*, *7*(6), 715–26.

Krendl, A. C., Macrae, C. N., Kelley, W. M., Fugelsang, J. A., & Heatherton, T. F. (2006). The good, the bad, and the ugly: an fMRI investigation of the functional anatomic correlates of stigma. *Social Neuroscience*, *1*(1), 5–15.

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*(7), 940–8.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human Amygdala Activation during Conditioned Fear Acquisition and Extinction: a Mixed-Trial fMRI Study. *Neuron*, *20*(5), 937–945.

LaBar, K. S., & LeDoux, J. E. (1996). Partial disruption of fear conditioning in rats with unilateral amygdala damage: correspondence with unilateral temporal lobectomy in humans. *Behavioral Neuroscience*, *110*, 991–997.

Lamm, C., & Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Structure & Function*, *214*(5-6), 579–91.

Lang, P. J. (1995). The Emotion Probe - Studies of Motivation and Attention. *American Psychologist Association*, *50*(5), 372–385.

Lang, P. J., Davis, M., & Öhman, A. (2000). Fear and anxiety: Animal models and human cognitive psychophysiology. In *Journal of Affective Disorders* (Vol. 61, pp. 137–159).

Lang, S., Kroll, A., Lipinski, S. J., Wessa, M., Ridder, S., Christmann, C., … Flor, H. (2009). Context conditioning and extinction in humans: differential contribution of the hippocampus, amygdala and prefrontal cortex. *The European Journal of Neuroscience*, *29*(4), 823–32.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*(8), 1377–1388.

Lattal, K. M., Radulovic, J., & Lukowiak, K. (2006). Extinction: [corrected] does it or doesn't it? The requirement of altered gene activity and new protein synthesis. *Biological Psychiatry*, *60*(4), 344–51.

Ledoux, J. (2012). NIH Public Access. *Neuron*, *73*(4), 653–676.

Liu, Y., Lin, W., Xu, P., Zhang, D., & Luo, Y. (2015). Neural basis of disgust perception in racial prejudice. *Human Brain Mapping*, *36*(12), 5275–5286.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–78.

Logothetis, N. K., & Wandell, B. a. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, *66*, 735–69.

Mallan, K. M., Sax, J., & Lipp, O. V. (2009). Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology*, *45*(6), 1303–1307.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*, *13*(4), 330–334.

Manson, J. H., & Wrangham, R. W. (1991). Intergroup aggression in chimpanzees and humans. *Current Anthropology*, *32*, 369–377.

Marsh, A. A., Blair, K. S., Jones, M. M., Soliman, N., & Blair, R. J. R. (2009). Dominance

and submission: the ventrolateral prefrontal cortex and responses to status cues. *Journal of Cognitive Neuroscience*, *21*, 713–724.

McKearney, J. W. (1969). Fixed-interval schedules of electric shock presentation: extinction and recovery of performance under different shock intensities and fixed-interval durations. *Journal of the Experimental Analysis of Behavior*, *12*(2), 301–313.

McLeod, S. A. (2007). Skinner - Operant conditioning. *Simply Psychology*, *1*(1), 2.

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry*, *62*(5), 446–54.

Mineka, S., Davidson, M., Cook, M., & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology*, *93*(4), 355–372.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268.

Molapour, T., Golkar, A., Navarrete, C. D., Haaker, J., & Olsson, A. (2015). Neural correlates of biased social fear learning and interaction in an intergroup context. *NeuroImage*, *121*, 171–183.

Molnar, S. (1998). *Human variation: races, types, and ethnic groups*. (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: on the development and operation of cues for control. *Journal of Personality and Social Psychology*, *83*(5), 1029–1050.

Navarrete, C. D., McDonald, M. M., Asher, B. D., Kerr, N. L., Yokota, K., Olsson, A., & Sidanius, J. (2012). Fear is readily associated with an out-group face in a minimal group context. *Evolution and Human Behavior*, *33*(5), 590–593.

Nikiforakis, N., & Normann, H. T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, *11*(4), 358–369.

Öhman, A., & Dimberg, U. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*.

Ohman, A., Fredrikson, M., Hugdahl, K., & Rimmo, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General*, *105*(4), 313–337.

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality & Social Psychology Bulletin*,

32(4), 421–33.

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. a. (2005). The role of social groups in the persistence of learned fear. *Science (New York, N.Y.)*, *309*(5735), 785–7.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, *10*(9), 1095–102.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–92.

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin*, *112*(3), 536–552.

Pavlov, I. (1927). *Conditioned Reflexes*. Oxford: Oxford University Press.

Phelps, E. A. (2006a). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.

Phelps, E. A. (2006b). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron*, *48*(2), 175–87.

Renner, K. E., & Tinsley, J. B. (1976). Self-Punitive Behavior. In G. H. B. B. T.-P. of L. and Motivation (Ed.), (Vol. Volume 10, pp. 155–198). Academic Press.

Rescorla, R. a. (1988). Pavlovian conditioning. It's not what you think it is. *The American Psychologist*, *43*(3), 151–160.

Rhodes, G., Brake, S., Taylor, K., & Tan, S. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology*, *80*(3), 313–331.

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2004). Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience*, *16*(2), 189–203.

Roseman, I. J., Wiest, C., & Swartz, T. S. (1994). Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology*.

Rowell, T. E. (1974). The concept of social dominance. *Behavioral Biology*, *11*(2), 131–154.

Rule, B. G., & Nesdale, A. R. (1974). Differing functions of aggression. *Journal of Personality*, *42*(3), 467–481.

Rule, B., & Nesdale, A. (1976). Emotional arousal and aggressive behavior. *Psychological Bulletin*, *83*, 851–863.

Sabatinelli, D., Bradley, M. M., Fitzsimmons, J. R., & Lang, P. J. (2005). Parallel amygdala and inferotemporal activation reflect emotional intensity and fear relevance.

NeuroImage, *24*, 1265–1270.

Sapolsky, R. M. (2004). Social Status and Health in Humans and Other Animals. *Annual Review of Anthropology*, *33*(1), 393–418.

Sapolsky, R. M. (2005). The influence of social hierarchy on primate health. *Science*, *308*, 648–652.

Sehlmeyer, C., Schöning, S., Zwitserlood, P., Pfleiderer, B., Kircher, T., Arolt, V., & Konrad, C. (2009). Human fear conditioning and extinction in neuroimaging: A systematic review. *PLoS ONE*.

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review*, *77*(5), 406–418.

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy*, *2*(3), 307–320.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews. Neuroscience*, *8*(4), 300–11.

Sidanius, J., Pratto, F., Martin, M., & Stallworth, L. M. (1991). Consensual Racism and Career Track: Some Implications of Social Dominance Theory. *Political Psychology*, *12*(4), 691–721.

Singer, T., Seymour, B., Doherty, J. P. O., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2009). UKPMC Funders Group Empathic neural responses are modulated by the perceived fairness of others, *439*(7075), 466–469.

Stillman, P. E., Van Bavel, J. J., & Cunningham, W. A. (2015). Valence Asymmetries in the Human Amygdala: Task Relevance Modulates Amygdala Responses to Positive More than Negative Affective Cues. *Journal of Cognitive Neuroscience*, *27*(4), 842–851.

Stillwell, A. M., Baumeister, R. F., & Del Priore, R. E. (2008). We're All Victims Here: Toward a Psychology of Revenge. *Basic and Applied Social Psychology*, *30*(3), 253–263.

Straube, T., Mentzel, H.-J., & Miltner, W. H. R. (2006). Neural mechanisms of automatic and direct processing of phobogenic stimuli in specific phobia. *Biological Psychiatry*, *59*(2), 162–70.

Taylor, S. P., Gammon, C. B., & Capasso, D. R. (1976). Aggression as a function of the interaction of alcohol and threat. *Journal of Personality and Social Psychology*, *34*(5), 938–941.

Van Bavel, J. J., & Cunningham, W. A. (2009). Self-categorization with a novel mixed-race group moderates automatic social and racial biases. *Personality and Social Psychology Bulletin*, *35*(3), 321–335.

Vervliet, B., Kindt, M., Vansteenwegen, D., & Hermans, D. (2010). Fear generalization in humans: Impact of prior non-fearful experiences. *Behaviour Research and Therapy*, *48*(11), 1078–84.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nature Neuroscience*, *6*, 624–631.

Williams, K. D., & Jarvis, B. (2006). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods*, *38*(1), 174–180.

Wood, K. H., Ver Hoef, L. W., & Knight, D. C. (2014). The amygdala mediates the emotional modulation of threat-elicited skin conductance response. *Emotion (Washington, D.C.)*, *14*(4), 693–700.

Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., & Fischl, B. (2014). Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage*, *88*, 79–90.