From Clinical Neuroscience
Karolinska Institutet, Stockholm, Sweden

IT'S THE INTENTION THAT MATTERS: NEURAL REPRESENTATIONS OF LEARNING FROM INTENTIONAL HARM IN SOCIAL INTERACTIONS

Irem Undeger

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-863-2
Cover illustration: Life of Nichiren: Rock Suspended by the Power of Prayer. Ca. 1835, Utagawa Kuniyoshi.

It's the intention that matters: Neural representations of learning from intentional harm in social interactions

Thesis for Doctoral Degree (Ph.D.)

By

Irem Undeger

The thesis will be defended in public at Nobelsväg 9, Solna, 2022-12-09, 14:30

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

Co-supervisor(s):
Fredrik Åhs
Mittuniversitetet
Department of Psychology and Social Work

Armita Törngren Golkar
Stockholm University
Department of Psychology

Opponent:
Joseph Dunsmoor
University of Texas Austin
Department of Psychiatry & Behavioral Sciences

Examination Board:
Erika Jonsson Laukka
Karolinska Institutet
Department of Neurobiology, Care Science and Society

Janina Seubert
Karolinska Institutet
Department of Clinical Neuroscience
Division of Psychology

Arvid Erlandsson
Linköping University
Department of Behavioral Sciences and Learning
Division of Psychology

Dedicated to all who have the courage to seek answers to their questions

Popular science summary of the thesis

Imagine somebody bumping into you very forcefully on the street, as you get mad and feel a sharp pain in your shoulder you hear them say: "I didn't mean it", "It was an accident". These are responses we commonly hear from people when they are accused of being guilty. Not seldom are they defending themselves by claiming to be ignorant about what happened: they were not informed enough to know that their actions would lead to a negative outcome, they didn't mean to hurt anyone, or simply put, it was not intentional.

Why and how is intentionality so important to us? Why do we so quickly forgive and forget when someone harms us accidentally but not otherwise? Why and how do we vividly remember a childhood memory when a friend knowingly hurt and bullied us? These are all questions that have helped shape ideas that created this thesis. Here, we will try to understand the biological mechanisms that accompany the perception of intentional harmful actions, with many aspects we can easily observe in our daily lives.

Of essential importance to this thesis is social interactions, as they provide the opportunity to study intentionality, but also harm and threat. In this thesis, I am presenting three studies examining the role of intentionality on how we learn about others, and from their actions. I provide evidence that suggests intentional and harmful actions lead to a myriad of behavioral and neural changes. For example, people are angrier towards the person that knowingly and willingly harmed them, compared to a person that did it accidentally (Study I and II). Similarly, intentional harmful actions provoke feelings of revenge, leads to increased discomfort, and dislike of the individual causing the harm (Study I and II). In a social interaction where a harmful action is performed with intention, the person receiving the harm overestimates the number of harmful actions performed and has negative emotions towards the person intentionally hurting them. If the actions were performed unintentionally, without knowledge, these effects disappear (Study I and II).

In addition to the behavioral changes intentional harmful actions create, we investigated how intentionality is represented by activity in the brain. I present evidence for specific signatures for intentional harmful actions in the brain, meaning certain brain regions distinguish between an unintentional and intentional harmful action (Study I). Moreover, going through a social interaction where harm was received from others led to changes in connectivity between certain brain regions, which was associated with increases in how easy the intentional harm do-er was recognized (Study III). Remembering novel but similarly categorized images as the one that delivered harm in the first place was also related to an increase in connectivity between certain brain regions, showing that the intentional harmful actions generalize (Study III).

Better remembering negative events, sharpening of recognition of conspecifics, or generalizing the harmful event serves a key biological purpose: to protect the self from future danger. This mechanism has served adaptive functions across species, including our own. On the flip side, when this protective learning mechanism goes beyond its initial purpose and neutral events are perceived as dangerous, it can do more harm than good. Thus, it is essential to understand what makes certain things more threatening to us. I hope that the findings in this thesis can help future research in how threat is processed during social interactions, and psychopathological conditions related to it.

Abstract

As a social species, humans are not only driven by the pursuit of necessities such as food and shelter, but also complex processes such as social interactions. To navigate our everyday life, we use information gathered throughout a lifetime of social interactions in which we learn from others and their actions but also, and not less importantly, about others. To create a complete picture of a social interaction, we assess the individual we interact with, make judgements about them and their actions, and integrate what we know with the consequences of their actions. This way, we learn the relationship between events (e.g. others' actions) and environmental stimuli, such as other individuals that predict the actions. As we encounter more people and go through more interactions, we continuously update information stored in memory from previous experiences. A common task, for example, going through the busy corridor in our workplace in a hurry does not only include avoiding physical harm caused by bumping into the coffee machine with a sharp corner, but also avoiding a co-worker we are in a feud with, and whom we believed knowingly spilled hot coffee on another co-worker the week before.

How social information is processed is key in understanding rarer but more impactful events that can have lifelong impact on an individual's life. Interpersonal trauma, a type of trauma that is acquired from harm received from another individual, leads more often to post-traumatic stress disorder (PTSD) than non-socially related trauma, for example, a car crash (Kleim, Ehlers, & Glucksman, 2007). To understand why a specific social harm affect us negatively, it is crucial to study how the brain integrates social, as well as non-social (physical) information during the harmful event.

In Study I, II, and III we investigated how different streams of information (social and physical) are integrated during a social interaction. We were interested in how intentionality of an action that has direct aversive consequences on an individual can change the individuals' judgements of the action and the person performing it. Using a time-based neuroimaging approach, we investigated how the value of an action is integrated with that of the intention behind it. Study I revealed evidence that suggests that intentionality of a directly experienced aversive action is represented throughout the cortex in neural activity patterns that form over time. Study II highlighted the importance of timing and sample size in similar paradigms, and that neural pattern formation in response to aversive actions regardless of the intentions behind them are robustly replicated.

In Study III we asked questions about how these learned action outcomes and knowledge about the people performing the harmful action change neural connectivity, and how this translates into changes in perception and memory 24-hours later. We found an increased connectivity between the hippocampus and the amygdala, which correlated with generalized memory responses to images associated with shocks from an intentional

harm do-er, and increased connectivity between the FFA and the insula, as well as the FFA and the dorsomedial prefrontal cortex (dmPFC) correlated with facilitated recognition of the intentional harm do-er's face.

List of scientific papers

I. **Undeger, I.**, Visser, R. M., & Olsson, A. (2020). Neural Pattern Similarity Unveils the Integration of Social Information and Aversive Learning. *Cerebral Cortex*, 30(10), 5410–5419.

II. **Undeger, I.**, Visser, R. M., Becker, N., de Boer, L., Golkar, A., Olsson, A. Model-based representational similarity analysis of BOLD fMRI captures threat learning in social interactions. *R. Soc. Open Sci.* 8: 202116. https://doi.org/10.1098/rsos.202116

III. **Undeger, I.**, Vieira, J. B., Thompson, W., Olsson, A. Brain functional connectivity after social interaction predicts enhanced memory for intentional harm. *Manuscript*.

Contents

1 Introduction 1
2 Literature review 3
2.1 Social Cognition 3
2.1.1 It's the intention that matters 3
2.1.2 Biological responses to others' intentions 4
2.1.3 Long term impact of negative social interactions 6
2.2 Aversive Learning and Extinction 7
2.2.1 Acquisition of aversive responses 7
2.2.2 Aversive memories 9
2.2.3 The extinction of aversive responses 10
2.2.4 Integrating social cognition with aversive learning and memory 11
2.2.5 Multivariate pattern analysis 13
2.2.6 Resting state connectivity 14
3 Research aims 17
4 Materials and methods 19
4.1 Participants 19
4.2 Experimental Stimuli 20
4.3 Experimental Design and Procedure 20
4.3.1 Manipulation of intentionality 20
4.3.2 Aversive learning 21
4.3.3 Post-learning test 23
4.3.4 Extinction learning 23
4.3.5 Functional localizer task 23
4.3.6 Perception and memory test 23
4.3.7 Post-experimental questionnaires and behavioral measures 24
4.4 Eye tracking: data collection and analysis 24
4.5 Magnetic Resonance Imaging 25
4.5.1 Preprocessing 25
4.5.2 Regions of interest 26
4.5.3 Representational similarity analysis 26
4.5.4 Resting state functional MRI 27
4.6 Ethical considerations 27
4.6.1 Working with electrical stimulation and deception 27
4.6.2 Privacy 28
4.6.3 Neuroimaging 28
5 Results 29
5.1 Study I. Neural pattern similarity unveils the integration of social information and aversive learning 29

5.1.1 Study I results and conclusions 29
5.2 Study II. Model based representational similarity analysis of bold fMRI captures threat learning in social interactions 33
5.2.1 Study II results and conclusions 33
5.3 Study III. Memories of intentional harm: Changes in perception and neural connectivity following an aversive social interaction 35
5.3.1 Study III results and conclusions 35
6 Discussion 39
6.1 Intentional vs. unintentional harm 39
6.2 Intentionality is integrated with aversive learning throughout the cortex 40
6.3 Neural changes that regulate subsequent memory 42
7 Conclusions 45
8 Points of perspective 47
9 Acknowledgements 49
10 References 51

List of abbreviations

PTSD	Post-traumatic stress disorder
CS	Conditioned stimulus
CR	Conditioned response
US	Unconditioned stimulus
SCR	Skin conductance response
fMRI	Functional magnetic resonance imaging
rs-fMRI	Resting state fMRI
RSA	Representational similarity analysis
MVPA	Multivariate pattern analysis
ROI	Region of interest
RSFC	Resting state functional connectivity
SEM	Standard error of the mean
ACC	Anterior cingulate cortex
FFA	Fusiform face area
dmPFC	Dorsomedial prefrontal cortex
vmPFC	Ventromedial prefrontal cortex
IFG	Inferior frontal gyrus
TPJ	Temporoparietal junction
arSTS	Anterior superior temporal sulcus
prSTS	Posterior superior temporal sulcus

1 Introduction

Intentionality behind a person's actions changes both how we perceive the person and how we perceive their actions. An individual that intentionally acts is perceived to be in full responsibility of their actions and are thus judged accordingly. For instance, an individual that intentionally harms another person is found blameworthy, whereas an accidental harm is forgiven (Ames & Fiske, 2015; Cushman, 2008). In a judiciary setting, intentionality plays an important role in punishment decisions (Zimring, 2000). It is crucial to understand why intentionality behind an individual's actions alters our perception of the actor because this understanding can have important bearing on both our understanding of relevant aspects of our everyday lives, and certain clinical diagnoses, such as post-traumatic stress disorder (PTSD).

Every day, we learn from others, and learn about them (such as the intentionality behind their actions). We learn the relationship between events and the environmental stimuli that predict them, which is reflected in our reactions to situations in the present and the future. This means that our present and future reactions are shaped by our understanding of others' actions in the past. In the example of a social interaction where we receive harm, this would mean that an individual that intentionally harms us will be remembered in the future as a potential threat as we expect them to be capable of knowingly harming us. This can be useful in avoiding dangerous situations, but in others might lead to unwanted consequences when the perception of threat generalizes to neutral situations such as in the case of PTSD (American Psychiatric Association, 2013). The overall aim of this thesis is to investigate how we acquire, keep and update information we learn about and from others, and how this information modulates learning and memory of aversive events.

Humans have evolved to live in social groups (Dunbar & Shultz, 2007), allowing us to survive by helping each other get food, and care for each other in dire situations. In evolutionary terms, survival only occurs if a fine-balanced trade-off between danger and reward is met. For instance, a hunter gatherer might need to trespass territory that belongs to a lion to reach food. This trade-off between reward and punishment requires both learning and remembering details about the dangerous territory, and about the reward that awaits on the other side. Thus, when we think of the biological mechanisms that govern our perception of intentional harms, we need to consider both the systems that involve social living and threat and reward processing. In the scope of this thesis, I will concentrate on the threat aspect of this balance, in the hopes of understanding how we integrate intentionality with harm and how this affects future encounters both in behavior and in neural markers.

To this end, in Study I, we investigated how harmful actions from intentional individuals are learned as threatening per se. Here, we used functional magnetic resonance imaging (fMRI) to investigate the neural signatures that govern the integration of intentionality

information with that of threat. We found that harmful actions are represented by specific patterns of neural activity during learning. In addition, we found tentative evidence that throughout the cortex, intentional and harmful actions are represented by specific patterns of neural activity whereas the unintentional or non-harmful actions were not. Supporting the view that intentional actions are also perceived differently, we found that intentional actions cause more anger, dislike, and feelings of revenge in individuals. Finally, we found that participants felt more discomfort from intentional harm than unintentional ones and even overestimated the frequency of intentional harm. Study II aimed to replicate these results and add additional ways to investigate the nature of these neural patterns. In Study III, we investigated the changes that occur in connectivity between brain regions before and after learning. Furthermore, we investigated how these changes relate to the expression of memory 24 hours after learning.

2 Literature review

2.1 Social Cognition

Social cognition refers to the perception, interpretation and evaluation of others and the self (Amodio, 2018). Everything from social constructs, such as race, gender, age, religion, or political affiliation can be thought of as a part of social cognition. These concepts are reflected in how we evaluate others and their actions. In addition to what we know of them, our impressions of others are shaped by the consequences of their actions, as well as the intentions we ascribe to these actions. Through our interactions, we learn about other individuals -and their actions- to adaptively navigate our social environment.

2.1.1 It's the intention that matters

According to both every-day moral code and formal legal practice, harm to others is seen as a ground for punishment. From rhesus monkeys (Masserman, Wechkin, & Terris, 1964) and 6-month old human babies (Hamlin, 2012; Van de Vondervoort & Hamlin, 2016), many species recognize and respond to when a conspecific is harmed. Harmful acts such as murder, abuse or theft are condemned across cultures and are basis for laws. When judging the harmful actions of others, we consider *why* a certain action was taken (L. Young & Saxe, 2009). Two types of information play a crucial role in making judgments; the outcome of the action, and the intentionality behind the action (Parkinson & Byrne, 2018). The intentionality gives information about if the action is deliberately inflicted, with knowledge about the consequences of its outcomes. For example, consider a scenario where a driver runs over a person. To decide on the punishment, judges would have to collect enough information to decide if this act was intentional or accidental. The outcome would vary from getting a fine, to time in prison, the severity of punishment increasing if the driver is found to be deliberately driving over another person (Zimring, 2000). In the event of the driver being below the age of 18, or a mentally disabled person driving the car the punishment is altered as mental capacity, and thus the intentionality, comes into question. Even if the drive over did not lead to harm, deliberately driving towards someone would still warrant punishment. Many studies to date have shed light on how judgments are made given different combinations of information on social and perceptual domains: intentionality (i.e., intentional, and unintentional) and outcome valence (i.e., harmful, or safe). These studies have shown that the valence (harmful or safe) and the intentionality (intentional or unintentional) of the action interacted and lead to different judgments on both the individual performing the action and the action itself (S. Levine, Mikhail, & Leslie, 2018). For instance, when an intentional action leads to harm, it is deemed more blameworthy and negative than an unintentional harm (Ames & Fiske, 2015; Cushman, 2008), and even if the intentional action fails to deliver harm (i.e. a failed-attempt) the individual is still found blameworthy (Young & Saxe, 2009). An unintentional and harmful action, on the other hand, is forgiven (i.e. given less punishment) (Martin &

Cushman, 2016). Individuals perceive intentional harm worse both explicitly, i.e. when asked, but also implicitly, i.e. without necessarily being conscious about it (Kurdi, Krosch, & Ferguson, 2020).

A great majority of work about intentional and harmful actions are about judging events that happened to another individual. These studies used stories that happened to other people and asked the participants to judge the individuals that did the intentional harm, or their actions. Very often, however, the outcomes of others' actions can be experienced in first person. This raises the question: if we are ourselves intentionally hurt by another individual, does that also feel worse than it does if it were to be unintentional? Recent literature suggests that intentional harms that are received in first person are indeed worse than unintentional ones. Intentional shocks were perceived more painful than unintentional ones when the same voltage of electrical shock (Gray & Wegner, 2008b). Overall, intentionality plays a role in both altering judgements about others and their actions, and the perception of a physical aversive event. This means that intentional harms are processed differently than unintentional harms.

2.1.2 Biological responses to others' intentions

The act of inferring another individual's intentions falls into the category of 'mentalizing'. Mentalizing is defined as the capability to infer other people's mental states; beliefs, needs, desires or goals, based on their behavior (Frith & Frith, 2005). Through mentalizing an individual predicts other people's future actions, understands how their own behavior might influence the behavior of others, and forms impressions (Koster-Hale & Saxe, 2013; Mende-Siedlecki, Cai, & Todorov, 2013; Wu, Liu, Hagan, & Mobbs, 2020). A specific network of brain regions is recruited during such tasks, referred to as the 'mentalizing network' (Figure 1) (Redcay & Schilbach, 2019). This network includes most notably the temporoparietal junction (TPJ), the dorso- and ventro- medial prefrontal cortex (dmPFC and vmPFC), the inferior frontal gyrus (IFG), and the insula (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014).

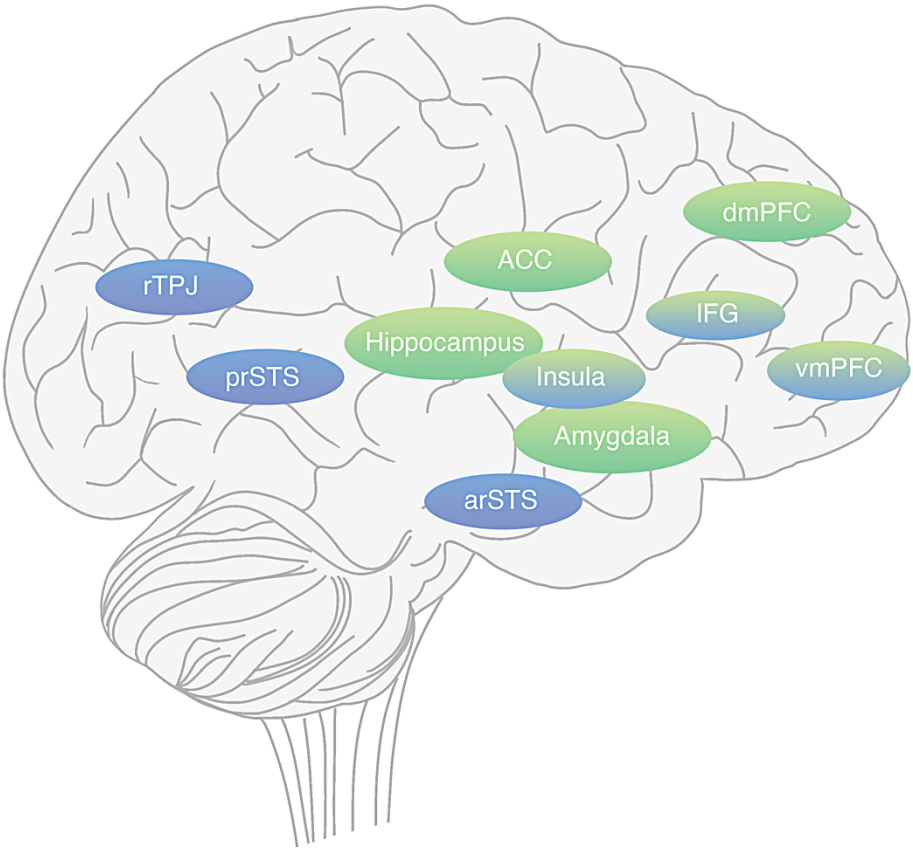

Figure 1. The mentalizing network. The key regions in the mentalizing (blue) and aversive learning (green) networks. Regions reported in both networks are shaded with both blue and green.

When experiencing aversive actions of others, the individual considers two components: i) the negative consequence of the action and ii) the intentionality behind the said action (Parkinson & Byrne, 2018). Previous research in intentionality has implemented both second-party and third-party tasks making it harder to understand to what extent personally experiencing the negative outcomes matter (Liljeholm, Dunne, & O'Doherty, 2014; Young & Saxe, 2008; Yu, Li, & Zhou, 2015). To understand a judiciary setting, third-party tasks are essential, as these tasks allow us to understand how intentionality of an action that has consequences on a third-party individual alters judgements. In this thesis, however, I will focus on the second-party scenarios where an individual directly experiences the negative outcomes of another individual's actions. These scenarios are more similar to what we usually experience during social interactions in our everyday, and situations that potentially lead to interpersonal trauma such as rape, or torture.

A recent study has aimed to understand the neural underpinnings of second-party intentional harm: using an interpersonal game, Liljeholm et al. (2014) has shown that insula was selectively activated for an intentional vs. an unintentional aversive experience. In another study, unintentionally caused aversive experiences (in this case, electrical

shocks), compared to intentional ones, were associated with activity in the right TPJ and the IFG. The intentional harmful choices that led to no shocks (failed-attempts) activated the insula, compared to the no-harm ones (Yu, Li, & Zhou, 2015). Together with the previous findings, this highlights the role of insula in processing negative experiences in social situations. The participants in these experiments formed their judgments through direct experience of the aversive event, as the confederates made choices. How these differences in perceiving unintentional and intentional actions evolved in time, as participants viewed decisions made by others, however, was left unexplored. Investigating the neural correlates of how social and perceptual information integrate over time is key in understanding how judgments about others and their actions are formed.

2.1.3 Long term impact of negative social interactions

Negative experiences serve us as a guideline to how we should navigate future experiences, in the case of a negative social interaction we can make decisions on who to befriend and who to avoid, as well as how to respond to others' actions. Without memory, none of the information we acquired in the past would be available to us. Memory requires learning, and negative events are in fact remembered better than others (M. M. Bradley, Greenwald, Petry, & Lang, 1992; Tambini, Rimmele, Phelps, & Davachi, 2017). As mentioned before, learning from negative events, i.e., aversive learning, did help our species survive when we lived in high-risk environments, but can have unwanted consequences on our mental health today.

Biological responses that aid us in face of danger can be detrimental to both mental and physical health when triggered out of context as in the case of psychopathological disorders. Post-traumatic stress disorder (PTSD) occurs in individuals that have experienced or witnessed a traumatic event and is described by vivid intrusions, or involuntary re-living of the traumatic event, avoidance, or avoiding certain situations at the expense of lowering life standards, and other negative cognitions and emotions (American Psychiatric Association, 2013). Traumatic events that lead to PTSD can be interpersonal, caused by an intentional action from another individual, or noninterpersonal, but not all traumatic events lead to PTSD. An interesting finding shows that experiencing interpersonal trauma leads more often to "complex PTSD", a more common and debilitating form of PTSD, than other kinds of traumas (Karatzias et al., 2019). Complex PTSD symptoms, in addition to PTSD symptoms listed above, are emotional dysregulation, negative self-cognitions and interpersonal hardship (Giourou et al., 2018). Simply the fact that the traumatic event was caused intentionally by another individual seems to alter how the traumatic event is processed, and greatly affect future behavioral and biological responses.

To sum up, regardless of the gravity of the outcome, how we perceive and respond to a negative action is altered when experienced as an intentional action. Thus, to study

aversive social interactions without accounting for intentionality, misses an essential part of such interactions. In an experimental setting, trauma can be studied by creating an experimental model of a situation in which harm is received, and the source of the harm is learned over time. Following this reasoning, we integrated *aversive learning* with a social interaction in Study **I**, **II** and **III**. As will be explained in the following section 2.2, aversive learning has well-established biological markers that can aid us in understanding what makes interpersonal traumatic events special.

2.2 Aversive Learning and Extinction

Learning from danger serves an essential role in survival. An animal that remembers sources of threat can avoid them in the future and increase chances of survival. If a baby hurts their hand by touching fire, they will not try again. As we interact with others, visit new places, and face novel information, sources of threat are not just items in the environment but include places, people, and contexts. To understand how we learn to associate certain individuals with threat, it is essential to understand what aversive learning is and how it takes place. As social information is integrated with physical harm in certain situations more readily than others (see section 2.1.1), it is important to understand how negative associations occur in the first place. Aversive learning paradigms allow us to understand how aversiveness of an event is acquired over time and can be followed by extinction learning paradigms that allow studying how a learned association can be updated to safety.

2.2.1 Acquisition of aversive responses

In most animals threat leads to responses that help the animal avoid danger by fighting, fleeing, or freezing (Bolles & Fanselow, 1980). Apart from these biological responses that are immediately apparent at the time of danger, if the animal learns from the threat long lasting changes in the brain occur (Thompson, 1986). Take an individual that gets assaulted in a certain street, for example. At the time of assault, the threat leads to biological responses that create signatures of specific events. If the individual learns about the street and the assailant as a threat, changes in the brain occur and the next time the individual passes by that street or sees the assailant, threat responses might occur regardless of if there is an actual danger or not.

The most studied form of threat learning is through *Pavlovian* or *classical conditioning*. Based on the work of Ivan Pavlov, classical fear conditioning studies aversive learning in controlled settings (Pavlov, 1927). Classical fear conditioning involves an innocuous environmental stimulus, called conditioned stimulus (CS), that predicts a noxious stimulus, called unconditioned stimulus (US). The learned response to the previously neutral CS, which became aversive through repeated pairings with the US, is called a conditioned response (CR). After the animal has been fear conditioned, the CS alone will evoke a CR. In mice, the US is often a foot shock, and the CS is a light or a sound (Davis,

1992; Delgado, Olsson, & Phelps, 2006; Phelps, 2006), while the CR is startle reflex or freezing. These reflexes (CR) that occur specifically in response to a CS that is paired with an aversive stimulus (US) allow us to quantify and study responses both during and after learning. In a typical experiment of aversive learning, one stimulus is paired with the US and the other is not. The stimulus that has been paired, or reinforced, is referred to as CS+, whereas the other is referred to as CS-. CR to CS+ are quantified to proxy subjective experience of threat and danger.

Conditioned responses in humans can be quantified via behavioral, psychophysiological, and neurobiological measures. For physiological measures, many studies use skin conductance responses (SCR) and pupillometry to measure arousal in response to CS+, in comparison to CS- (Boucsein, 2012; Korn, Staib, Tzovara, Castegnetti, & Bach, 2017; Phelps, Delgado, Nearing, & Ledoux, 2004; Visser, Haan, Scholte, & Kindt, 2016). Throughout an aversive learning paradigm, the CS+ stimuli that is paired with an aversive stimulus cause a differential response in SCR and pupil size as both increase in response to threat (Reinhard & Lachnit, 2002). Most common behavioral measures are questionnaires about US expectancy (i.e., how much does the participant expect to receive the US when CS is presented), and contingency (i.e., which CS led to the US). Thus, an individual's responses to threat both during and after a fear paradigm can be reliably quantified by physiological responses that occur automatically at the face of danger (i.e., pupil responses and skin conductance response) and from subjective reports of threat.

Translating a classical aversive learning paradigm into the real-life example above, the assailants' face, the assault weapon and the smells and sights that make the street are what we will refer to as conditioned stimuli (CS). As the individual walks through the street, these are simply neutral things that naturally occur in the environment. After the assault, which would be the US, the CS that are paired with the assault become a CS+. Now the face of the assailant or the street is no longer neutral but are associated with an aversive event.

The assault leads to physiological responses, such as an increased pupil size. This response can be quantified both during the assault and later when the individual faces the assailant again, to investigate if successful aversive learning occurred. If learning did occur, we expect to see threat responses even in the absence of real danger at a later timepoint.

Neural mechanisms behind classical fear conditioning are well-established. The predominant brain circuitry consists of the amygdala, the hippocampus, the insula, and the prefrontal cortex (PFC) (among many other regions and sub-regions, see Fullana et al., 2020). At the center of this circuitry is the amygdala (for arguments against this see the meta-analysis Fullana et al., 2016). At the face of threat the amygdala starts a signaling cascade which leads to the physiological effects such as increased blood flow to large

muscles , increased heart rate (Cacioppo, Tassinary, & Berntson, 2007), and increased pupil dilation (Cacioppo et al., 2007; Olsson & Undeger, 2017). Apart from its involvement in expressions of threat, the amygdala works in concert with other brain regions in the acquisition (i.e. learning) and storage of associations of threat (Ledoux, 2000). The amygdala is highly connected to the rest of the brain (M. P. Young, 1993) The expression of learned threat associations are mediated by connectivity within and between the amygdala and the hippocampus (de Voogd, Fernández, & Hermans, 2016; Hermans et al., 2016; Tambini et al., 2017). The hippocampus is an essential brain region in the formation and expression of episodic memories (Chadwick, Bonnici, & Maguire, 2012). Disrupting the connections between the amygdala and the hippocampus leads to emotional memory enhancement (Roesler, Roozendaal, & McGaugh, 2002) and stimuli associated with aversive stimulation (i.e. CS+) showed increased activity in the anterior cingulate cortex (ACC) to during fear conditioning, compared to a safe one (Dunsmoor, Ahs, Zielinski, & LaBar, 2014; Visser, de Haan, et al., 2016; Visser, Scholte, Beemsterboer, & Kindt, 2013). ACC was also activated after a post learning reminder, along with vmPFC (Feng, Zheng, & Feng, 2015). Lesions in the ACC led to attenuated skin conductance response (SCR) (Tranel & Damasio, 1994), and an increase in activity in the ACC, the insula, which correlated with pupil response magnitude (Leuchs, Schneider, Czisch, & Spoormaker, 2017). Furthermore, increased connectivity within the insula have been related to increased arousal and emotion-related activity (Touroutoglou, Hollenbeck, Dickerson, & Feldman Barrett, 2012). A recent study reported that the activity in the insula, the amygdala and the vmPFC strongly predicted SCR to the CS+, but less so subjective fear ratings of the same stimuli (Taschereau-Dumouchel, Kawato, & Lau, 2020). To sum up, the role of the insula, the amygdala and the vmPFC seem to be tied together with that of physiological measures used to quantify the subjective experience of threat. Together, the amygdala, the hippocampus, the ACC, the insula, and parts of the PFC are considered to be a part of the *aversive learning network*, an interconnected hub in the brain that is involved in processing threat (Figure 1).

2.2.2 Aversive memories

Memories can be divided into many categories depending on how they are acquired, how long they are stored and what kind information they store (Norman, 1970). Memories that represent our personally experienced events, such as the one in the example when the person who gets assaulted on the street, are called episodic of memories (Tulving, 1972). Such memories include several components related to the event: where, when, what the event was (Kitamura, 2017), and who it was performed by (Hitti & Siegelbaum, 2014) and thus are based on complicated memory systems that include many subsystems. For example, such memories are associated with emotions (i.e. lead to emotional memories) (LeDoux, 1993), and are not explicitly controlled (i.e. implicit memories) (Schacter, 1990).

In the scope of this thesis, we will mainly focus on the episodic and emotional memory components.

Events that lead to emotional arousal are better remembered than neutral ones (Reisberg & Heuer, 1992). One of the most prevalent ways to frame emotion is in terms of two components: *valence*, which varies in a scale of unpleasant to pleasant, and *arousal*, which varies from quiet to active (Lang, Greenwald, Bradley, & Hamm, 1993; Wundt, 1912). Measures such as pupillometry and SCR are commonly used to measure the arousal of an experimental subject (Lonsdorf et al., 2017). When trying to understand emotional memories, we need to consider both the processes that govern memory *encoding* (i.e. acquisition) and *recall* (i.e. remembering) (Hamann, 2001). The process that leads to the storage of a learned experience is referred to as *consolidation*. Briefly, an experience is *encoded*, and if this encoding leads to memory formation, then the memory is *consolidated* (i.e., stored). Consolidation can be tested by *recall*. As mentioned briefly in the previous section, emotional memory encoding is governed chiefly by the connectivity between the hippocampus and the amygdala (Figure 2) (Hamann, Cahill, McGaugh, & Squire, 1997). Plus, throughout the acquisition of aversive memories, synchronization of neural activity between the hippocampus and the amygdala increased linearly, and the persistence of connectivity between the hippocampus and the amygdala predicted the long term expression of fear (Hermans et al., 2016).

It is worth noting that what is referred to as fear memories (i.e., memories acquired from aversive learning) have certain unique features. For aversive learning, context (e.g. where the association was made, like the street the assault happened) plays a great role (see Yonelinas & Ritchey, 2015 for an overview). Thus, it is interesting to see the effects of a social presence, such as in the case of social interactions, as it also sets a "social context" in which aversive learning occurs.

In the case of traumatic events, the mechanisms that allow such strong encoding play a key role in understanding psychopathologies such as PTSD. As PTSD occurs more readily upon socially related trauma (Kleim, Ehlers, & Glucksman, 2007), and presents itself far into the future of an individual, it is essential to study the effects of social cognition on future memory, as it is to study them during learning. In **Study III** of this thesis, we will investigate how intentional harm alters connectivity between certain brain regions before and after learning, and how this translates into changes in perception and judgments of intentional harm doers and their actions 24 hours later.

2.2.3 The extinction of aversive responses

In our daily lives, we constantly update information we acquire. For instance, a colleague we have had bad experiences with can become a friend in the next months. Without understanding *how* these associations are formed over time, we cannot fully understand *why* some are harder to update. *Aversive learning* paradigms (section 2.2.1) allow us to

understand how aversiveness of an event is acquired over time, whereas an *extinction learning* paradigm allows studying how a learned threat association can be inhibited or updated by creating new memories of safety.

Extinction learning paradigms include repeated exposures to CS+ without its pairing to the US. This procedure extinguishes the CR to the CS+, resulting in the reversal of the differential responses to the CS+ and the CS- (Dunsmoor, Ahs, et al., 2014; Kindt & Soeter, 2013). However, this process does not "erase" the previous aversive memory trace, but rather creates a new memory trace for the association of CS+ with safety (for a review, see Dunsmoor & Murphy, 2015). This new association can be studied similarly to the arousal responses used during aversive learning, with measuring SCR, pupil responses, brain activity and self-reported questionnaires. The neural correlates of human extinction learning are still up for debate. For instance, although it is considered a key region in extinction learning, (Phelps et al., 2004) a recent meta-analysis (Fullana et al., 2018) was unable to capture vmPFC activity during extinction learning tasks throughout different studies. The extinction process is particularly susceptible to context, such as the experimental setup. Despite the debates in brain imaging, extinction learning paradigms are suitable to study how a previously acquired aversive experience is expressed later.

When aiming to study social aversive learning, extinction learning paradigms allow us to understand how the aversive interaction generalizes to naturalistic settings and gives a way to quantify learning related changes. Taking the example of the assault on the street, the victim would firstly respond with fear to facing the assailant later, but if the future interactions were always safe and without any harmful outcomes the victim might get more and more comfortable and thus less fearful.

2.2.4 Integrating social cognition with aversive learning and memory

Much progress has been made to uncover the neural and behavioral mechanisms of social cognition. How these processes affect learning, however, has rarely been a topic of interest. This is interesting because previous research shows that information about others, as well as information acquired from others, affect emotional responses. For instance, perceived level of pain intensity can be enhanced simply through making the recipient believe that the pain is caused by a person intending it to happen (Gray & Wegner, 2008a). In this thesis we will focus on social cognition's integration with aversive learning to shed light on processes that govern the interaction of our judgments about others and how we acquire threat information.

The idea that certain stimuli lead to greater CR than others is not novel. In the 1970s Seligman suggested that there has to be a certain class of stimuli to which the animal maintains fears in order to avoid imminent threats (Seligman, 1971). These stimuli are commonly related to phobias. These so-called 'prepared stimuli' would aid a species to survive in the future, as learning to avoid it would be faster. In fact, pictures of fear-related

stimuli, such as snakes and spiders, were more readily associated with an aversive stimulus (Tomarken, Mineka, & Cook, 1989), led to greater CR, and have shown a 'resistance' to extinction learning (Öhman, 1986). This resistance is quantified as a high CR response throughout extinction trials, as non-fear-related stimuli would lead to less CR in time during extinction learning (see section 2.2.2). This means that once a relationship between fear-related stimuli and an aversive association has been learned, it is harder to create a new extinction learning memory that signals safety. The theory of prepared stimuli has not, however, gone unchallenged. A recent study has shown that most studies fail to replicate the resistance to extinction findings (Åhs et al., 2018). Here, the authors argue that preferential aversive learning of certain stimuli is due to genetic factors and thus explaining the psychopathological overlaps such as phobia. In **Study II** of this thesis, I will present findings about the responses to intentional aversive stimuli during extinction learning.

Prepared stimuli are not contained to commonly known dangerous animals such as snakes, but also certain social stimuli such as another individual's face. Faces with angry expressions, or faces of individuals from other racial ethnicity (Molapour, Golkar, Navarrete, Haaker, & Olsson, 2015; Olsson, Ebert, Banaji, & Phelps, 2005) seem to be processed in a similar way to the traditional prepared stimuli. This suggests that also social information, such as what race we perceive another individual to be or the emotions we think they carry can alter our perception of a neutral stimulus, and thus modulate aversive learning. In this thesis we will investigate if intentionality attributed to an aversive action can have similar implications and cause a 'prepared stimulus' effect. Some questions we will address are: Does intentionality of an aversive action make it easier to associate it with an aversive response? Will memories of intentional harm resist an update to safety? These questions have important implications for understanding how we navigate our social environment now and, in the future, as social cues are one of the most salient cues in our daily lives. Furthermore, these questions will aid us in understanding psychopathologies, such as PTSD, which stems from trauma that cannot be updated to safety.

To this end, here, I will marry methodologies from two research traditions, social cognition, and aversive learning, in order to delineate the neural and behavioral mechanisms underlying the *integration* of social information into aversive learning. In **Study I** and **II,** we will take advantage of time-sensitive analysis methods (see Methods section) on neuroimaging data to understand the neural mechanisms that govern the integration of intentionality of an action with its aversive value during a social interaction. In **Study III**, we will investigate the changes in neural connectivity before and after the integration of social cognition and aversive learning, and how these changes predict perception, and memory.

2.2.5 Multivariate pattern analysis

Functional magnetic resonance imaging, or fMRI, data is commonly used in both social cognition and learning research. fMRI is non-invasive and provides good spatial resolution. Depending on the conditions of the research, the brain can be observed as divided into 1x1x1 mm cubes – namely *voxels*. This is similar to pixels in a digital image, the smaller the voxel the more resolution a brain image has. Each of these voxels sum up the activity of around a billion neurons. While it is possible to detect regional activity in the brain using traditional analyses, novel analysis methods allow us to capture fine grained information represented in the activation of individual voxels. The activity of neuronal populations can be observed as patterns of activity from voxels in a certain brain region, coining the term multivariate pattern analysis, MVPA. Where traditional analysis asks the question *where* in the brain a certain condition is represented, multivariate analyses ask *how* a certain condition is represented in the brain. A classic example is that of V1 neurons that are sensitive to the orientation of a stimulus. While it is not possible to decode the orientation of a stimulus from traditional analysis, MVPA can robustly capture brain signatures that represent the orientation of a visual object (Haynes & Rees, 2005). An intuitive way to understand MVPA would be to imagine a black and white image and the value of each of these pixels that create it. Each pixel would range from 0 (white) to 1 (black) in value, forming a greyscale. Put together, each of these pixels will form an image that we recognize as an object, say a square photograph of a face. In traditional analysis this image would be represented as an average value of all the pixels that make up the image, which would result in a homogenous gray square. Similarly, using MVPA, certain neural activity patterns can be seen as signatures of a specific event and thus makes it possible to decode the state of the brain from this activity.

At times, it is informative to compare and contrast brain activity. For example, how similarly are monkey and human faces represented in the human brain? This question can't be answered by traditional analysis, as it is only possible to contrast average activity over regions of the brain. Applying MVPA to the visual cortex, Kriegeskorte et al. (2008) compared the different patterns that form in the human visual cortex while viewing images of human and non-human animals. This comparison was made by creating a matrix, which consists of the "dissimilarity" (calculated by subtracting the similarity from 1, i.e., 1-correlation) between the neural patterns that are activated in response to each image. As this analysis unveils the activation pattern similarity, it is called representational similarity analysis (RSA). Importantly, the correlation matrix has shown that there is a categorical representation of objects, faces and bodies in the visual cortex, represented by similar patterns for category exemplars but less similar patterns for outside category exemplars.

Researchers studying aversive learning have implemented the developments in this methodology and found elegant applications of these methods. As learning happens over

time, average activity over time offers valuable but limited insight. Neural activity can be examined by recording from a single neuron in animal research, but the equivalent is rarely possible in humans due to ethical reasons, apart from clinical cases. Time and location sensitive methods of recording neural activity is important since aversive learning research has the potential to allow understanding of pathologies such as PTSD which rely heavily on the moment of exposure and what follows and localization of potential targets in the brain would help future efforts in helping patients. Time-based RSA offers a solution to the issues of low spatial and temporal resolution and offers a way to understand how brain patterns evolve over time, during aversive learning. During classical fear conditioning, RSA shows a sharpening in the patterns of neural activity to reinforced CS's as a function of their common threat value (Visser, Scholte, & Kindt, 2011), creating a "category" representation of CS+. In particular, participants that retained memories of the aversive experience specifically formed patterns of activity in brain regions involved in aversive learning such as ACC, insula, and vmPFC (Visser, Kunze, Westhoff, Scholte, & Kindt, 2015; Visser et al., 2013).

In this thesis, we took advantage of the power of time-based RSA and its ability to unveil brain activity pattern changes over time. We implemented time-based RSA in an aversive learning paradigm where confederates inflicted electrical shocks to the participant either intentionally, or unintentionally. This means that the participants learned the aversive contingencies of the confederates' actions over time but knew about the intentionality of these actions throughout the experiment. This allowed us to investigate how attributions of intentionality are integrated with aversive learning, and its neural correlates.

2.2.6 Resting state connectivity

Resting state connectivity analysis uses fMRI to quantify changes in brain activity during rest. Compared to task-based fMRI, for example an aversive learning paradigm where participants are instructed to view images, no task is performed during resting state fMRI (rs-fMRI). Resting state functional connectivity (RSFC) is commonly measured during rs-fMRI, for which the correlation of signal fluctuations in the brain during rs-fMRI are calculated. This means that brain regions from which physiological signals are temporally correlated, are recruited for the same "resting" task.

Many different functional connectivity networks have been identified to date, thanks to RSFC studies. For instance, the default mode network (DMN) has been identified as a network of regions that are active during rest, compared to during task (Buckner, Andrews-Hanna, & Schacter, 2008). This network is active by "default", in the absence of any instruction or external stimuli. Rest immediately following social interaction is investigated less commonly in the literature, however research shows a network of regions related to mentalizing that overlap with that of DMN (Mars et al., 2012). The dorsomedial subsystem of the DMN includes dmPFC, and TPJ, which are essential parts

of the mentalizing network (Figure 1). A recent meta-analysis highlighted the importance of these regions in mentalizing tasks, compared to the other counterparts of the DMN (Spreng & Andrews-Hanna, 2015). Social dynamics, such as being excluded during a social interaction by friends or inclusion during a social interaction modulates activity in the brain regions that are associated with mentalizing (Mars et al., 2012). Connectivity between regions associated with mentalizing, such as the TPJ and the insula have been linked to negative messages in a social media platform (H. Zhang & Mo, 2016). A recent study investigated if the mentalizing network is involved in consolidating social information (Meyer, Davachi, Ochsner, & Lieberman, 2018). Here, participants were introduced to information about individuals and rs-fMRI was performed before and after the encoding (i.e., learning) period. At the end of the experiment, participants performed a surprise memory test to assess their memory performance. An increase in neural activity was observed in the rTPJ, l TPJ and mPFC during social encoding. Additionally, RSFC analyses revealed an increase between rTPJ-lTPJ, and rTPJ-mPFC connectivity after social encoding. Finally, connectivity between the hippocampus and mPFC predicted better social memory. These findings highlight the importance of these regions in processing social information both during a socially related task and in the aftermath of such social events, as well as the value rs-fMRI brings.

Understanding neural activity during rest is essential to understand learning and memory formation, as these so-called "offline periods" are key for processes such as aversive learning (Pape & Pare, 2010; Paré, 2003). Rodent research has shown that neural synchrony between the amygdala and the hippocampus during post-learning rest increase during and after acquisition of fear (Popa, Duvarci, Popescu, Léna, & Paré, 2010; Seidenbecher, Laxmi, Stork, & Pape, 2003). In humans, RSFC revealed a replay of recently performed tasks (Albert, Robertson, & Miall, 2009), and that connectivity between the hippocampus and neocortical regions persist during rest that follows memory encoding (Tambini & D'Esposito, 2020; Tambini, Ketz, & Davachi, 2010; Tambini, Rimmele, Phelps, & Davachi, 2016). Plus, aversive learning increased connectivity between the amygdala and the hippocampus, which predicted stronger recovery of fear 24 hours later (Hermans et al., 2016). These findings highlight the importance of studying awake rest following social interactions, to understand under which conditions social emotional memories are consolidated.

Findings about brain mechanisms that are related to aversive learning, memory and the interplay with social information is important for the development of preventative and treatment strategies in PTSD. Indeed, PTSD is characterized by re-experiencing trauma related memories, as well as an overgeneralization of fear responses (S. A. Joshi, Duval, Kubat, & Liberzon, 2020). In **Study III**, we investigated the neural correlates before and after a social interaction that led to aversive learning. We aimed to understand the

generalization of fear responses in relation to these neural correlates, and thus used a face perception task and a memory test that can capture fear generalization 24-hours after.

3 Research aims

The overall aim of this thesis is to investigate how social interactions can lead to aversive learning, and how this learning is shaped by information about the interaction partner. To this end, the following objectives were specified:

- To investigate how social information (i.e., intentionality) is integrated with aversiveness of an action outcome (Study I).
- To replicate previous findings and apply different methodologies that can unveil the mechanisms behind intentional harmful action processing (Study II).
- To understand changes in connectivity between brain regions during rest following an aversive social encounter, and how these changes are associated with subsequent behavior 24 hours later after an intentionally harmful social interaction (Study III).

4 Materials and methods

4.1 Participants

All participants signed an informed consent form prior to participation and were compensated for their participation after the experiment. All participants complied with any restrictions that are a part of participating in an MRI study. All participants were healthy, had normal-to-corrected vision, had no psychological disorders, used no prescription medication, were right-handed and reported normal hearing.

Study I and **Study III** are from the same data collection, thus both experimental stimuli and setting, and participant sample recruited are identical. For these studies, 48 participants were recruited. Four participants out of this sample failed to understand the instructions and were removed from further analyses, and with seven participants the scanning session was stopped early due to technical difficulties. For these participants, we continued scanning as post-learning resting state scans as scheduled as these scans would not be affected. Out of these seven, three participants did not understand the task and were also removed from the results of **Study III.**

In **Study I,** we used a pool of 40 participants out of the 48 recruited, due to technical problems as explained above. From the 40, pupillometry data comes from a sample of 35 due to data loss (n = 5). Because of excessive motion, six participants' neuroimaging data were removed from the sample, and one due to not understanding the task. Neuroimaging results are thus reported for 33 individuals in this study. Behavioral measures are reported for the same sample. Apart from the one participant that failed to understand the task, the participants removed from the pupillometry, and neural analysis do not overlap.

For **Study II,** 31 participants were recruited. Pupillometry data is reported for a sample of 23 participants. For the neuroimaging results, participants were excluded due to excessive motion (n = 4), and technical difficulties (n = 1) resulting in a sample size of 26 (mean age: 24.54).

In **Study III**, we used data from 42 individuals for the RSFC analyses, 2 individuals showed excessive motion during scanning (>0.5 mm mean displacement). Five participants have shown motion during scanning for the functional localizer task, and we could not locate FFA from the face localizer task for certain individuals (n = 6). Thus, we used 23 individuals for analyses involving individual FFA masks. The behavioral results are reported for 34 individuals for the face recognition and CS memory tests. This is the number of participants that are overlapping between the ones that showed up to the behavioral test, that have RSFC data available and that understood the task and have properly answered the questions (n = 3 answered every question with a "Yes").

4.2 Experimental Stimuli

For **Study I** and **Study II**, the face stimuli used during aversive learning were photographs of the confederates used in the study. The images that served as CS during aversive learning for **Study I–III** were drawn from four categories: tools, fruits, animals, and buildings. One image from each category served as a CS, and the four images used in both **Study I/III** and **II** were identical. The images were chosen from publicly available resources on the internet.

Participants were presented with more exemplars from the four categories that served as CS during the functional localizer task reported in **Study III**. These images were chosen from publicly available resources on the internet.

CS images used in the memory task reported in **Study III** were acquired from the same resource. For the perceptual face recognition task, photographs of the confederates used in the study were morphed into novel faces using an image morphing software (Squirlz Morph: www.xiberpix.com). Novel face images were acquired from the Radboud faces database (Langner et al., 2010).

All images used in **Study I–III** were equalized to match luminance using the SHINE Toolbox (Willenbockel et al., 2010). This step ensures that the pupillometry measures we collected during the experiments were not affected by changes in luminosity (see section 4.4).

4.3 Experimental Design and Procedure

4.3.1 Manipulation of intentionality

To study intentionality, we had to create an environment in which the participants were exposed to intentional and unintentional actions. In the literature, this has been done either using vignettes (S. Levine et al., 2018; Martin & Cushman, 2016; Parkinson & Byrne, 2018; Yang et al., 2019; L. Young & Saxe, 2009), and more recently using social interaction paradigms (Liljeholm et al., 2014; Yu et al., 2015). In our studies (**I–III**) we chose to create a socio-interactive environment to create a naturalistic task. This means that the participant not only observed the intentional actions but also experienced them.

In the studies reported in this thesis, the participant was made to believe that the confederates were recruited as participants for the same experiment. To create the experience that one of the confederates was intentionally and the other unintentionally delivered electrical stimuli, we used a script to introduce the experiment to the participant and the confederates while they were waiting in the same room. Following the instructions, the participant and the confederates were asked to pick a paper from a lottery bag that would decide who would be in the MR scanner for the experiment. This bag only contained MR scanner papers but when the experimenter asked who picked the MR scanner paper, the confederates would state that their papers stated “outside”.

Thus, the participant came to believe that in the following parts of the experiment, the two co-players (confederates) would perform their own tasks while the participant would watch their computer screens through an online connection from inside the scanner. At the beginning of the learning part of the experiment, the participant watched confederates making a choice that either signifies that they want to be able to deliver electrical shocks to the participant during the experiment or not. Here, one confederate said "Yes", thus becoming the intentional co-player and the other confederate said "No" and thus becoming unintentional. The participant is then informed that even if the unintentional co-player doesn't want to give shocks, the system will in fact deliver shocks when one of the images is chosen. Thus, the unintentional participant would "accidentally" deliver these shocks, without knowledge of delivering them. This setup was identical for both experiments used in **Study I-III**.

4.3.2 Aversive learning

In **Study I-III** we used a modified Pavlovian fear conditioning paradigm, which involves delivering electric shocks (US). This allowed us to investigate the effects of intentionality on aversive actions, as compared to neutral ones. For all the studies, the same aversive learning paradigm was used. Participants were informed that confederates would make choices between two images throughout the experiment, and one of these choices would result in electrical shocks. By assigning one choice to a shock outcome, and the other to no shock outcome, we mimicked a classical fear conditioning paradigm in a socio-interactive setting.

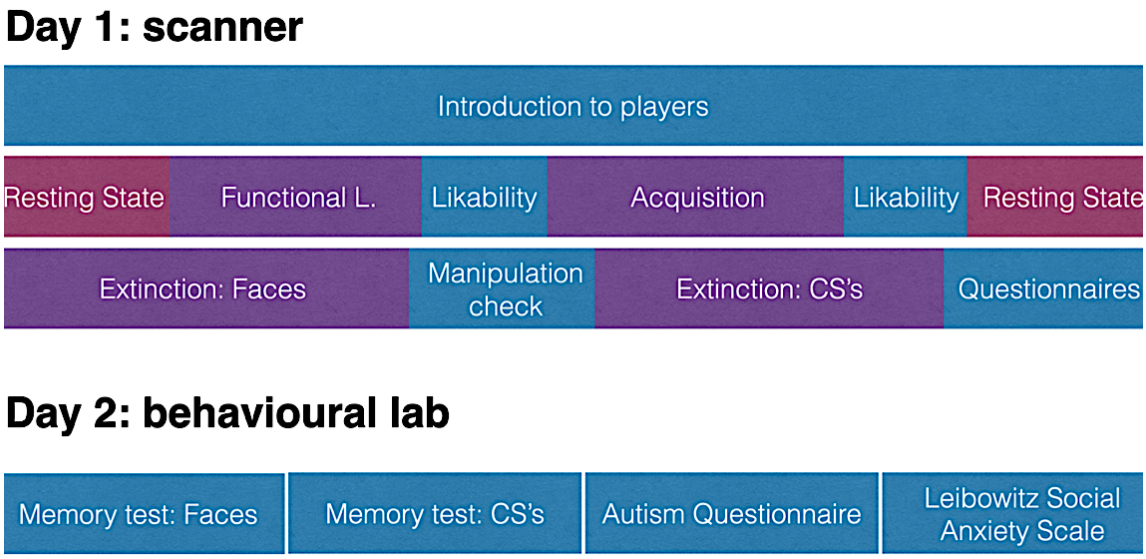

Figure 2. Experimental overview of Study I and III.

The aversive learning task consisted of 26 choices for each confederate. Confederates both chose the image that is associated with the shock outcome (i.e., CS+) 50% of the time (13 trials). Out of these 13, 6 trials were reinforced with the shocks and 7 were not. Each trial consisted of three phases: an early anticipation phase where a photograph of the confederate is presented (3s), followed by a choice phase where the confederate is making a choice (1s), and finally the chosen option phase where the choice made by the co-participant remained on the screen, but the unchosen option disappeared (3s) (Figure 3 A). Between each trial a fixation cross was presented for 13s. The onset of each trial was triggered by an fMRI pulse. In **Study I**, at the event of a reinforced CS+, a 200ms shock was presented during the choice task. In **Study II, a** 200 ms shock was presented during the chosen option phase, at 2.8 s.

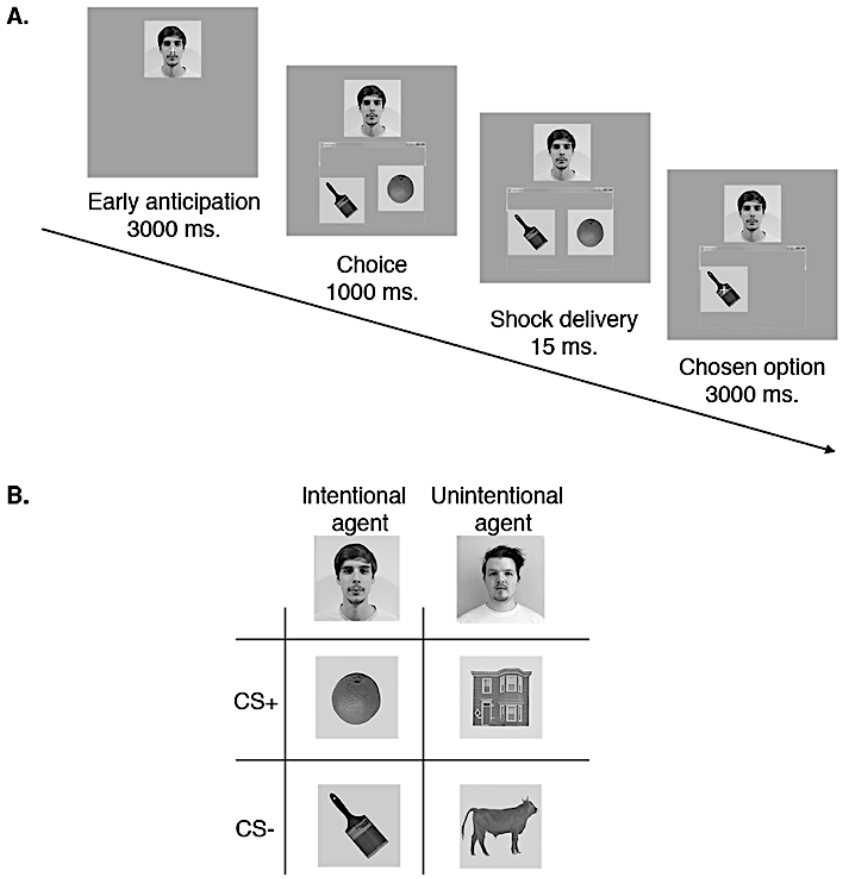

Figure 3. A) The breakdown of a trial in the learning paradigm used in Study I and III. In Study II, the timing of the shock was changed and was presented 2800 ms after the chosen option was visible. B) The 2x2 experimental design for the learning task in Study I, II and III.

4.3.3 Post-learning test

In the study design that is reported here as **Study I** and **III**, we implemented a post-learning task (Figure 2) in which participants were exposed to the face (i.e., face test) and CS stimuli (CS test). For the CS test, we used the same stimuli as during learning, the photographs of the confederates. Each image was presented for 3 seconds, 10 times per face and in a randomized order. The CS test was identical in design but included the CS images that were used during learning instead of the face photographs. This data was not used for any of the work presented in this thesis.

4.3.4 Extinction learning

The extinction learning task in **Study II** was identical to the aversive learning task, except that the participant did not receive any electrical shocks.

4.3.5 Functional localizer task

Functional localizer tasks are used to capture specific neural activity to certain stimuli. The most famous example is the functional localizer task used to distinguish faces (N Kanwisher, McDermott, & Chun, 1997) where participants were presented with images of faces and common objects. Upon retrieving the brain activity related to viewing faces, compared to the objects, the fusiform face area (FFA) was discovered. This area in the brain robustly responds to faces, and between faces has selectivity for familiar faces (Weibert & Andrews, 2015).

We implemented a functional localizer task in the design of **Study I** and **III**, to capture brain regions that are involved in processing faces, in order to investigate the responses to the face stimuli (i.e., the photographs of the confederates) we used during the experiment. The functional localizer task included exemplars from the four categories we used as CS images and faces. This task was performed before the learning scanning session (Figure 2) and is reported in **Study III**.

4.3.6 Perception and memory test

The perception and memory test reported in **Study III** was aimed to capture behavioral responses to the confederates and CS images that were presented to the participant 24-hours following learning. Here, we used photographs of confederates that were morphed into novel faces at varying degrees (10%-90%) to capture the generalization of facial perception. Individuals with interpersonal trauma can mistake unrelated individuals for their perpetrator and go through a triggering of the traumatic event, showing generalization of physical characteristics (Ehlers & Clark, 2000). Thus, we investigated if the intentional confederate would be perceived at a different percentage than the unintentional one.

Similarly, we presented the participant with multiple exemplars from the same category of each CS stimuli, along with the actual CS stimuli that were presented during the aversive learning task. Research suggests that when the CS+ and US contingency is learned, the arousal response can generalize to other stimuli in the same category as the CS+ in following tests, showing higher arousal to the CS+ category but not the CS- (Dunsmoor, Kragel, Martin, & LaBar, 2014; Dymond, Dunsmoor, Vervliet, Roche, & Hermans, 2015; S. M. Levine, Kumpf, Rupprecht, & Schwarzbach, 2020; Onat & Büchel, 2015). Thus, we investigated if the same generalization principles would hold with the learning paradigm, we implemented in **Study I**, and how and if intentionality would affect generalization principles. To this end, participants were asked if the images were identical to the ones they saw the day before. We hypothesized that participants would generalize the intentional CS+ image (i.e., answering yes to exemplars) but not the others.

4.3.7 Post-experimental questionnaires and behavioral measures

To get an initial baseline of how participants felt towards the confederates, participants were asked "how likable" they found each interaction partner at the before the learning task and after in **Study I** and **Study II**.

After the scanning was completed, participants filled out post-experimental questionnaires for **Study I** and **Study II**, inquiring about their experience during learning, and their emotional state afterwards (Figure 2). The first questionnaire they received was the contingency questionnaire, where each CS image was presented along with the questions: i) Did you receive any shocks when this picture was chosen? If yes, how many? ii) How many times was this picture chosen overall? And iii) On a scale of 1-5, how much did you expect to receive shocks when this picture was chosen? With these questions we were able to assess how participants associated each image with the aversive outcome.

The following questionnaire included questions regarding the confederates: i) how uncomfortable the shocks were coming from each confederate, ii) how many shocks they received from each confederate, iii) what the motivation of the intentional confederate was, iv) how they felt when there was a decision to deliver shocks, v) how the shocks made the participant feel when received, vi) if they would like to deliver shocks to the confederates, if yes how many, vii) if they paid any attention to the intentionality of the confederates throughout the experiment, viii) how angry they felt towards each of the confederates.

4.4 Eye tracking: data collection and analysis

Eye-tracking equipment collects information about the size and location of the pupil by emitting infrared light, which reflects from the pupil of the eye. The diameter of the pupil changes in response to changes in light, but also changes in cognitive and emotional

states. Past research has shown that the pupil size increases (i.e. dilates) in response to emotional stimuli (Bradley, Miccoli, Escrig, & Lang, 2008; Partala & Surakka, 2003; Siegle, Steinhauer, Carter, Ramel, & Thase, 2003), during mental cognitive tasks (S. Joshi & Gold, 2020). As pupils dilate at the face of emotional arousal, pupil size can be used as a read out for fear related processes. We used pupillometry in **Study I** and **II** in this thesis to index aversive learning and to index extinction learning in **Study II**. Pupil responses were recorded using an MR-compatible eye-tracker.

4.5 Magnetic Resonance Imaging

The goal of task-based fMRI experiments is to map responses of the brain to the perceptual, motor, or cognitive manipulations of the task. For **Study I**, **II,** and **III** we collected both T1-weighted anatomical images and functional images.

4.5.1 Preprocessing

The preprocessing of fMRI data was identical for **Study I** and **Study II**. The preprocessing was performed using the software FSL version 5.0 (Oxford Centre for Functional MRI of the Brain (FMRIB) Software Library, http://fsl.fmri b.ox.ac.uk/fsl/fslwiki/), and with the function FEAT (FMRI Expert Analysis Tool). During an fMRI scanning session, individual slices of brain images are recorded. Thus, this means that there is a delay between slices that make up the whole brain. To circumvent this issue, we performed slice-time correction. To correct for head movements of the test subjects, we used motion correction. This step is necessary because displacements of the head distort the homogeneity of the magnetic field, which means that the assumed anatomical location of brain activity would be changed. Changes that are too great cannot be compensated with correction, and thus individuals with excessive motion were excluded from the studies. To know the anatomical location of the activity detected in functional brain scans, each functional image needs to be matched with the same individual's anatomical brain scan. Thus, we co-registered structural images to the functional images. After this, to a group comparison all brain images need to be in the same brain *space*. This is possible by transforming all co-registered brain images from the participant sample to the same template. Therefore, we transformed the brain images to MNI (Montreal Neurological Institute) space using the FNIRT (FMRIB's Non-linear Image Registration Tool). The normalization parameters acquired from this process were applied to the functional images. Additionally, we used temporal high-pass filtering (SIGMA = 100s) to clean the signal from physiological noise, and pre-whitening to remove temporal autocorrelation.

In 2016, Eklund and colleagues analyzed data from 499 healthy controls and 3 million tasks, using the most common software packages. To their surprise, they faced a false positive rate of 70% (Eklund, Nichols, & Knutsson, 2016). The preprocessing of fMRI data in **Study III** was performed using fMRIprep (Esteban et al., 2020). We chose to use fMRIprep since it is a novel protocol that aims to minimize the methodological variability

between fMRI studies. Albeit its common collection in cognitive neuroscience experiments, the analysis of fMRI data and preprocessing differs widely between research groups, and even individuals within the groups. This difference resulted in failure to replicate findings using different software packages, affecting reproducibility of fMRI studies (Bowring, Maumet, & Nichols, 2019).

4.5.2 Regions of interest

In all the studies included in this thesis we used regions of interest. Using pre-defined regions when conducting fMRI analysis reduces the number of Type I errors, as it limits the number of statistical tests. This way, instead of correcting for the number of voxels present in the brain, the correction is made for the number of ROI used (Poldrack, 2007).

In **Study I** and **II** regions of interest (ROI) were chosen based on their relevance in the literature (Figure 1). The regions that we included due to their role in aversive learning (Visser, Haan, et al., 2016; Visser et al., 2013, 2011) included: the amygdala, the hippocampus, the insula, the ACC, and the vmPFC. We included the regions left and right TPJ, dmPFC, the anterior and superior parts of the superior temporal sulcus (arSTS, prSTS), and the IFG for their relevance in social cognition (Liljeholm et al., 2014; Y. Zhang, Yu, Yin, & Zhou, 2016). ROI masks for the right and left TPJ, anterior and posterior STS, and the dmPFC were acquired by creating a 5mm spherical mask on coordinates reported on the website Neurosynth when searching for the term 'intention' (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). We then intersected these masks with an anatomically derived mask that is proximal to the coordinate, to have an anatomically correct fit of the spherical mask. The remainder of the ROI masks were created from the Harvard-Juelich atlas, based on their anatomical locations.

In **Study I**, we supplemented the findings in our *a priori* defined ROI with other ROI present in the Harvard-Juelich atlas to explore the whole brain.

ROI can be defined anatomically, or functionally. To understand how a set of voxels respond to a specific manipulation, functional localizers are used. Functional localizer scans provide subject-specific activity in response to the stimuli used during the task and are often used to analyze tasks the same individual performs (Nancy Kanwisher, 2010). In **Study III**, we used a similar functional localizer approach and derived ROI for faces and captured the fusiform face area (FFA) activity in all subjects (N Kanwisher et al., 1997).

4.5.3 Representational similarity analysis

In the studies presented in this thesis we aimed to investigate how social information (i.e., the intentionality behind an action) and information about the aversive properties of an action (i.e., the relationship between chosen images and shock delivery) are integrated.

To understand the neural mechanisms that underlie this integration, it is required to look at what happens during learning, as the aversiveness of the actions are learned.

To this end, we used a trial-by-trial RSA, in which we calculated the similarity between brain patterns between each trial. In **Study I** we used this methodology during the learning task and investigated which (if any) brain regions represent intentionality of an aversive action, and if the interaction partners' representation in neural patterns change over time as learning unfolds.

In **Study II**, we added an extinction learning phase where we investigated how and if extinction learning also shows modulation of trial-by-trial neural pattern correlations due to the intentionality behind the actions.

4.5.4 Resting state functional MRI

Resting state functional MRI provides information about changes that occur in the brain, but not while the participant is not involved in a task. Rather, the participant is in rest and the changes in brain activity can be informative of changes in brain structures in relationship with another. In **Study III**, we investigated the changes in connectivity before and after learning, and how they related to recognition of the confederates' face 24 hours following the learning task, as well as how and if participants remember the imagery that was used as the choice images (i.e., CS+ and CS-).

4.6 Ethical considerations

4.6.1 Working with electrical stimulation and deception

To study aversive associations, participants had to undergo an aversive experience. Based on the previous literature (Fullana et al., 2020; Olsson, Nearing, & Phelps, 2007), we chose to use electrical stimuli. This can be considered an ethical concern since the stimuli are uncomfortable. All participants knew electric shocks would be delivered to them and signed up to the experiments voluntarily, knowing this fact. Additionally, every participant was allowed to terminate the experiment whenever they wanted, with full compensation. Finally, the level of electric shocks was determined by the participants themselves. We used a standardized "work-up" procedure in which we delivered a barely noticeable voltage and increased it until the participant found the stimuli "uncomfortable but not painful". There are no known side-effects of the electric stimulation, except for slight irritations on the skin in very rare cases.

The deception constituted another ethical consideration. Similar procedures have been used several times both in our group and on other research groups without any known adverse implications. Participants were debriefed afterwards and informed about the deception.

4.6.2 Privacy

During neuroimaging, sensitive data such as the participants' personal number, MRI image of their whole head, as well as certain demographic information (e.g., weight, age) are collected. All data are saved with unique and fully anonymized participant numbers on a server where only authorized researchers have access. In publications, individual data that can be traced back to the actual identification are never revealed.

4.6.3 Neuroimaging

During the neuroimaging sessions, participants enter the MRI scanner, and need to be fitted inside a head-coil. This coil fits quite snug, and the MRI scanner is a semi-closed environment. Additionally, the scanner room is locked from the outside while the participant is inside due to security measures. All these aspects of the setting make it uncomfortable with individuals suffering from claustrophobia. During the pre-recruitment screening, participants are asked about their feelings about being in a closed environment, and individuals with claustrophobia are not recruited. Additionally, every participant has the right to terminate the experiment at any point in time with full compensation and are informed of this beforehand.

The strong magnetic field of the scanner is another concern when thinking about using fMRI. This creates a risk for individuals who might have metal in their bodies (e.g., pacemaker) or are in a potential risk group (i.e., pregnant individuals). Thus, during the pre-recruitment procedure we do a thorough screening to make sure there cannot be any risks to the participants' health.

Any neuroimaging session can lead to incidental findings, such as a tumor in the brain. Every scanning session we perform is checked by radiologists and the participant is informed of this during recruitment. In the case of an abnormality, the participant is contacted by the hospital for further evaluation. Furthermore, the participant is well-informed that our experiment is not a medical procedure, and thus cannot be conclusive to test for pathology.

5 Results

5.1 Study I. Neural pattern similarity unveils the integration of social information and aversive learning

Throughout our lives, we have countless social interactions. From shopping groceries to growing up with a childhood-friend, our lives are colored by the interactions we have and the relationships we build. To function around these interactions, we need to infer others' thoughts, emotions and learn about them and their actions. Learning to avoid a dangerous situation, a person or a combination thereof is of essential importance to our mental and physical well-being. Intentionality of an action is one of the social information we consider when we judge individuals' actions, motives and in turn lead to judgements about the individual. Intentionally performed harmful actions are deemed more harmful than unintentional ones, and the individuals performing the are found more blame-worthy (Monroe & Malle, 2019). Furthermore, an intentionally harmful event hurt more (Gray & Wegner, 2008b).

In **Study I, Study II** and **Study III**, we used a socio-interactive aversive learning paradigm where participants interacted with an intentional and an unintentional confederate. The confederates delivered shocks by making choices between two images, one was associated with a shock, and another was not. This design allowed us to investigate the differences between an intentionally delivered shock and an unintentional one. As participants learned the association between the images and the shocks, we were able to observe changes that occurred that are related to learning.

5.1.1 Study I results and conclusions

Post-experimental questionnaires revealed that the participants have learned the association of shocks with CS+ images ($CS+_{intent}$ M = 6.07, SD = 3.83; $CS+_{unintent}$ M = 3.41, SD = 3.41; $CS-_{intent}$ M = 0.71, SD = 1.12; $CS-_{unintent}$ M = 1.10; SD = 2.28) where there were no effects of intentionality but of CS type ($F(1,32) = 73.96$, $P < 0.001$, $\eta^2 = 0.69$). Intentional shocks led to feelings of revenge ($t(31) = 2.47$, $P = 0.032$, $d = 0.33$) and participants reported wanting to deliver more shocks to the intentional confederate if given the chance (Figure 4 C). They were angrier towards the intentional confederate ($t(31) = 4.48$, $P < 0.001$, $d = 0.82$) (Figure 4 D), and reported a greater dislike after the learning phase, compared to before ($F(1,58) = 11.27$, $P = 0.004$, $\eta^2 = 0.13$) (Figure 4 E).

Participants reported receiving greater number of shocks from the intentional confederate ($t(31) = 2.41$, $P = 0.022$, $d = 0.46$) (Figure 4 B), as well as feeling greater discomfort from intentional shocks, compared to unintentional ones ($t(31) = 2.56$, $P = 0.016$, $\eta^2 = 0.53$) (Figure 4 A). Interestingly, the number of shocks reported for intentional confederate was higher than the actual number participants have received (mean = 1.65; SD = 3.88; $t(31) = 2.413$, $P = 0.022$). Participants also expected to receive a shock more

often upon seeing a CS+ image that the intentional confederate has chosen, compared to one that the unintentional confederate chose (Interaction of CS type (2) and intentionality (2) ($F(1,31) = 4.38$, $P = 0.044$, $\eta^2 = 0.12$)) (Figure 4).

We asked participants if they ever doubted the experimental manipulation. 11 out of the 33 participants used in the final sample indicated doubting the experimental manipulation was a set-up, more than 50% of the learning phase. When we investigated the effects of this measure, we saw that the effects of discomfort from shocks, revenge, and likeability disappeared ($P > 0.05$). Participants showed greater arousal (indexed by pupil dilation) to CS+ images, than they did to CS−, regardless of intentionality ($F(1,33) = 21.95$, $P < 0.001$, $\eta^2 = 0.399$) (Figure 4 F).

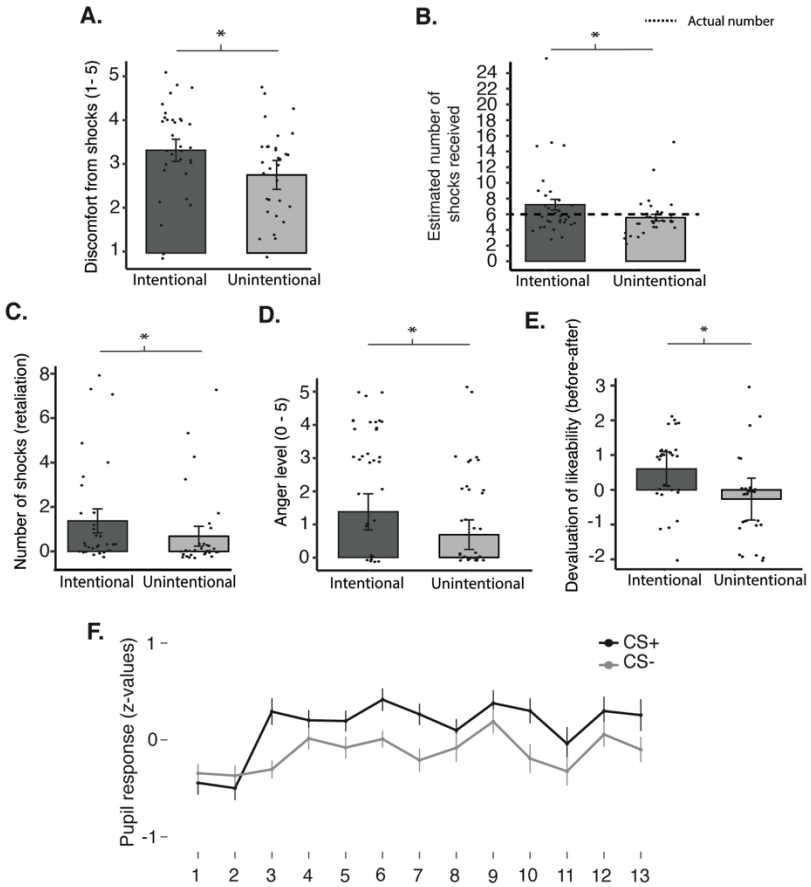

Figure 4. Behavioral responses to the learning task, and pupillometry results from **Study I**. (A) Discomfort of shocks received from each confederate, (B) Number of shocks the participant reported to receive via the presentation of each choice image, (C) How many shocks the participant would like to deliver back if given the chance, (D) How angry the participant felt towards the co-participants. (E) The change in participants' evaluations of how likable each co-participant was, from before the learning task to after. (F) Pupil dilation responses to the stimuli. The Error bars represent standard error of the mean (SEM).

We used a trial-by-trial representational similarity analysis on the neural data during learning to capture the correlation between brain activity patterns that unveiled in time. We saw a differential increase in trial-by-trial similarity patterns to CS+ images, compared to CS-, (main effect of CS type (2)) in the insula ($F(1,32) = 7.89$, $P = 0.008$, $\eta^2 = 0.198$) (Figure 5 A), the ACC ($F(1,32) = 6.98$, $P = 0.013$, $\eta^2 = 0.179$) (Figure 5 B), the lFG ($F(1,32) = 15.60$, $P < 0.001$, $\eta^2 = 0.328$) (Figure 5 C), the dmPFC ($F(1,32) = 5.874$, $P = 0.021$, $\eta^2 = 0.155$), the arSTS ($F(1,32) = 8.892$, $P = 0.005$, $\eta^2 = 0.217$), the prSTS ($F(1,32) = 5.443$, $P = 0.026$, $\eta^2 = 0.145$), the vmPFC ($F(1,32) = 4.459$, $P = 0.043$, $\eta^2 = 0.122$) and lTPJ ($F(1,32) = 8.38$, $P = 0.003$, $\eta^2 = 0.241$).

We found a main effect of intentionality in in the insula ($F(1,32) = 5.42$, $P = 0.02$, $\eta^2 = 0.145$) (Figure 5 A), the ACC ($F(1,32) = 5.97$, $P = 0.02$, $\eta^2 = 0.153$) (Figure 5 B), the lFG ($F(1,32) = 5.96$, $P = 0.02$, $\eta^2 = 0.157$) (Figure 5 C), the dmPFC ($F(1,32) = 4.70$, $P = 0.03$, $\eta^2 = 0.128$), and the arSTS ($F(1,32) = 4.06$, $P = 0.05$, $\eta^2 = 0.113$), suggesting a preference in these regions to the intentional CS's. However, these effects did not survive FDR corrections.

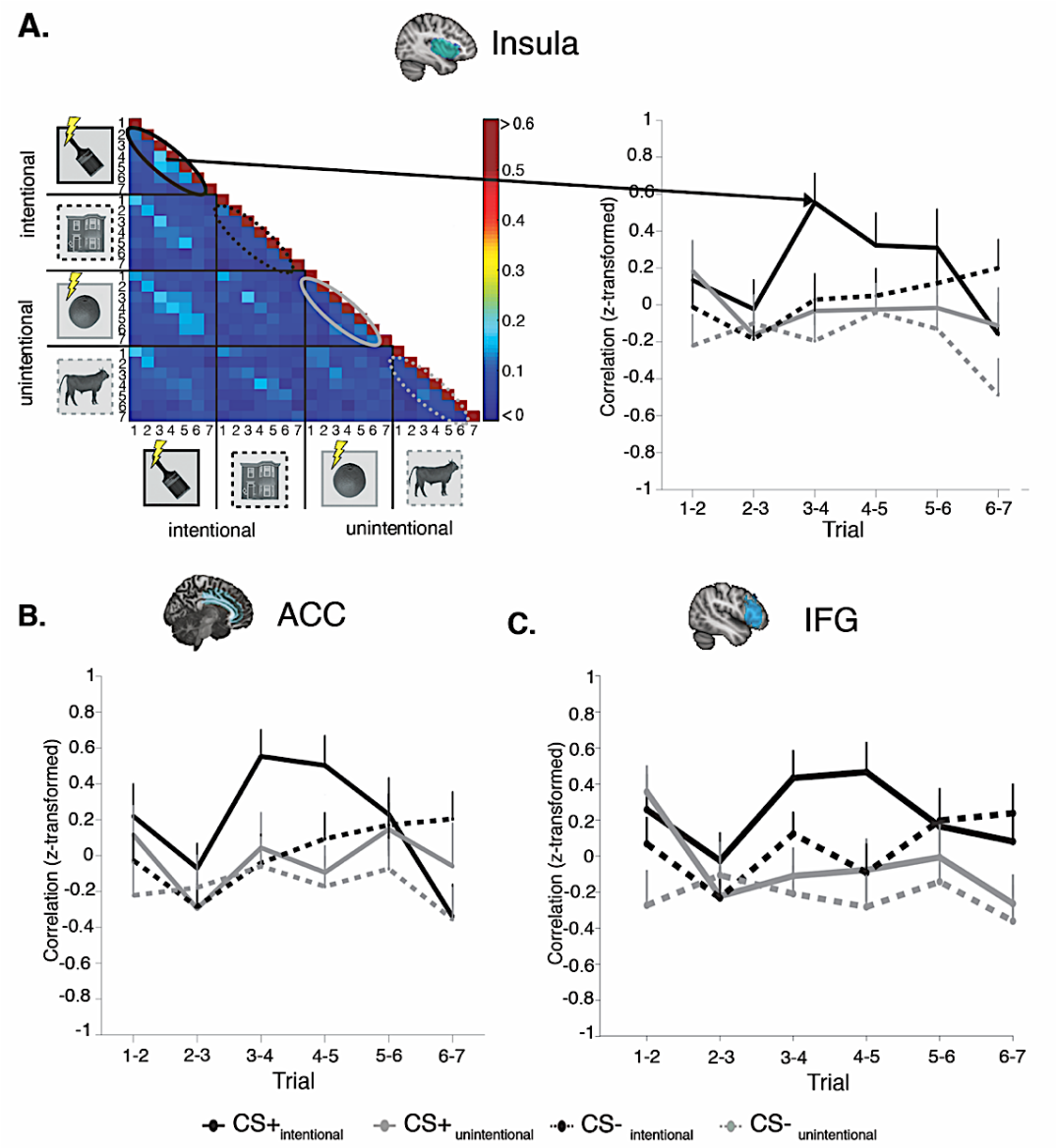

Figure 5. Neural responses to the learning task: trial-by-trial pattern similarity in (A) the insula, (B) the ACC and (C) the IFG. The correlation matrix represents correlations of neural patterns during learning, in each indicated ROI. The removed upper diagonal is a mirror image of the lower. Error bars represent SEM.

In conclusion we introduced a novel experimental paradigm that can capture both aversive and mentalizing properties during a social interaction task. We found that the aversiveness of an action is regulated by the intentionality behind it. Our results tentatively suggest the intentionality and the aversiveness of actions are represented by unique signatures in overlapping brain regions, across the cortex.

5.2 Study II. Model based representational similarity analysis of bold fMRI captures threat learning in social interactions

Study I provided evidence of the importance of mentalizing in learning aversive associations from a social source. Our aim here was to replicate these findings and to apply a different statistical approach that can test our neural hypotheses in a single statistical model, instead of multiple ANOVAs we conducted in previous research. Additionally, we aimed to investigate the update of learned aversive associations with an extinction learning task. Extinction learning following aversive learning can both inform about the learning task and can be used to index the strength of the learned association. For instance, certain social information such as race lead to a resistance to extinction (Navarrete et al., 2009; Olsson et al., 2005).

5.2.1 Study II results and conclusions

Participants in the study sample reported receiving a greater number of shocks from CS+ images, compared to CS- ones (t23 = 3.17, p = 0.004, d = 0.65). This confirmed that participants knew which images were associated with shocks and which were not. However, we did not find effects of learning indexed by greater pupil dilation to the CS+ images, compared to CS- in this study.

We replicated the results of the **Study I** and saw effects of intentionality on shock expectancy (CS [2] × Intentionality [2], F1,24=8.74, p=0.007, η^2=0.26), and numbers (CS [2] × Intentionality [2], F1,24 = 4.72, p = 0.047, η^2 = 0.16), where participants reported expecting more shocks from the intentional CS+ and receiving more shocks from intentional decisions. Participants rated intentional shocks more uncomfortable than unintentional ones (t_{23} = 3.02, p = 0.006, d = 0.62), were angrier (t_{23} = 3.21, p = 0.004, d = 0.65), and more revengeful to the intentional confederate (t_{23} = 3.33, p = 0.003, d = 0.68). Here, different from the previous study, we found no differences between the decrease in likeability of the intentional and unintentional confederate after learning. Five participants out of the same reported doubting the experimental manipulation during more than 50% of the learning task. The effects of the estimated number of shocks received and discomfort disappeared when we included this information (Figure 6).

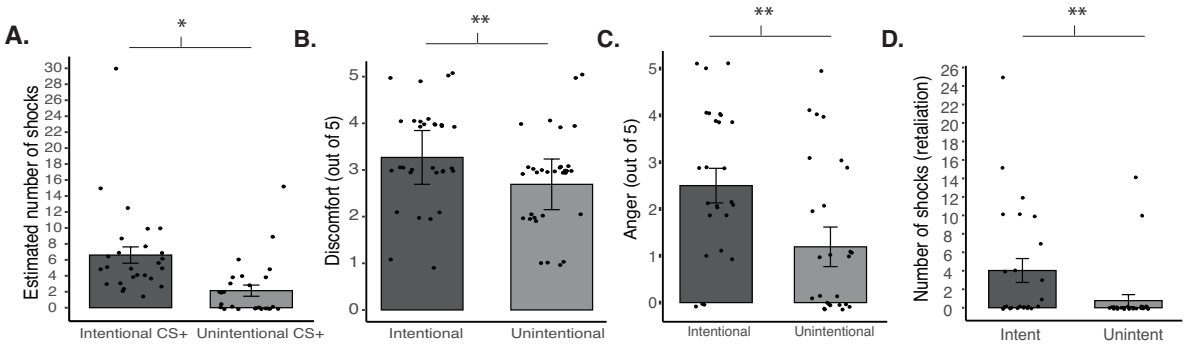

Figure 6. Behavioral responses to the learning task, and pupillometry results from **Study II**. Panels (A) number of shocks the participant reported to receive via the presentation of each choice image, (B) discomfort of shocks received from each confederate, (C) how angry the participant felt towards the co-participants. (D) how many shocks the participant would like to deliver back if given the chance, Error bars represent SEM.

We found that CS+ images were represented by a trial-by-trial increase in pattern correlations during aversive learning in the insula ($t_{25} = 4.02$, $p < 0.001$, $d = 0.78$) (Figure 7) and the IFG ($t_{25} = 4.31$, $p < 0.001$, $d = 0.84$). We found no effects of intentionality.

Using a univariate approach, we were able to see the effects of intentional harm in the insula, the dmPFC, and the rTPJ. During the early anticipation phase of the extinction learning task, we found effects of a gradual increase in response to faces of the unintentional confederate in the hippocampus. We found a differential decrease for the $CS+_{intent}$ in the amygdala.

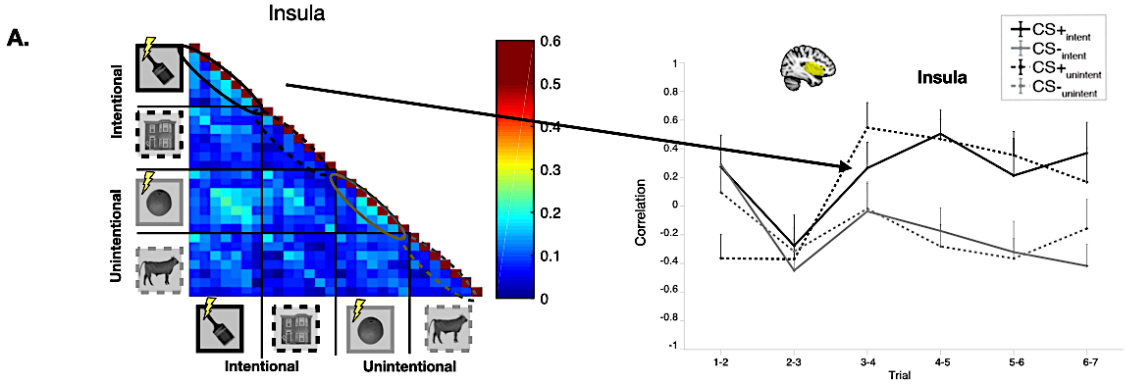

Figure 7. (A) Neural responses to the learning task: trial-by-trial pattern similarity in the insula. The correlation matrix represents correlations of neural patterns during learning. Error bars represent SEM.

In conclusion, we replicated the effects of intentionality on aversive learning on behavioral measures. We failed to replicate our findings regarding pupillometry. While participants reported knowing the association between the CS+ images and shock delivery, their arousal levels failed to show a differential increase to the CS+. Nevertheless, we were able to capture trial-by-trial representational similarity increase for the CS+ images. As for intentionality, we only saw effects of a gradual decrease during the extinction phase.

5.3 Study III. Memories of intentional harm: Changes in perception and neural connectivity following an aversive social interaction

Learning leads to changes in brain connectivity (Tambini et al., 2017). **Study III**, builds up on **Study I**, as we collected rs-fMRI before and after the learning task used in **Study I**. In this study, we correlate rs-fMRI findings with behavioral tests related to face perception and CS image memory generalization 24 hours following learning.

To this end, we used a face perception paradigm where participants were asked if they remembered individuals presented to them from the day before. These images were the faces of confederates from the day before, morphed with novel faces in varying degrees. This was followed by a CS memory test, where participants were presented CS images that were used during the learning task, as well as similar images from the same category. Participants were asked if they saw these images the day before.

5.3.1 Study III results and conclusions

We found that participants reported remembering the confederates with higher confidence ($\beta = 0.92$, SE = 0.03, $p < 0.001$), as the percentage of their presence increased in the morphed images. Surprisingly, participants reported receiving more shocks from the images containing the unintentional confederates face than the intentional one ($\beta = 0.36$, SE = 0.09, $p < 0.001$). In **Study I**, we showed that participants over-estimated the number of shocks received from the intentional confederate immediately after learning (Figure 4 B). Participants reported correctly to which CS image was presented to them the day before, with higher confidence ($\beta = 0.53$, SE = 0.01, $p < 0.001$).

We found a higher Δ connectivity (calculated as connectivity after – connectivity before) between the insula-FFA (Unintentionality X FFA – insula: β = -26.13, SE = 12.43, p = 0.036) and dmPFC-FFA(Unintentionality X FFA – dmPFC: β = -37.51, SE = 15.75, p = 0.017) correlated with earlier detection of faces from the learning task, for the intentional confederate's face (Figure 8).

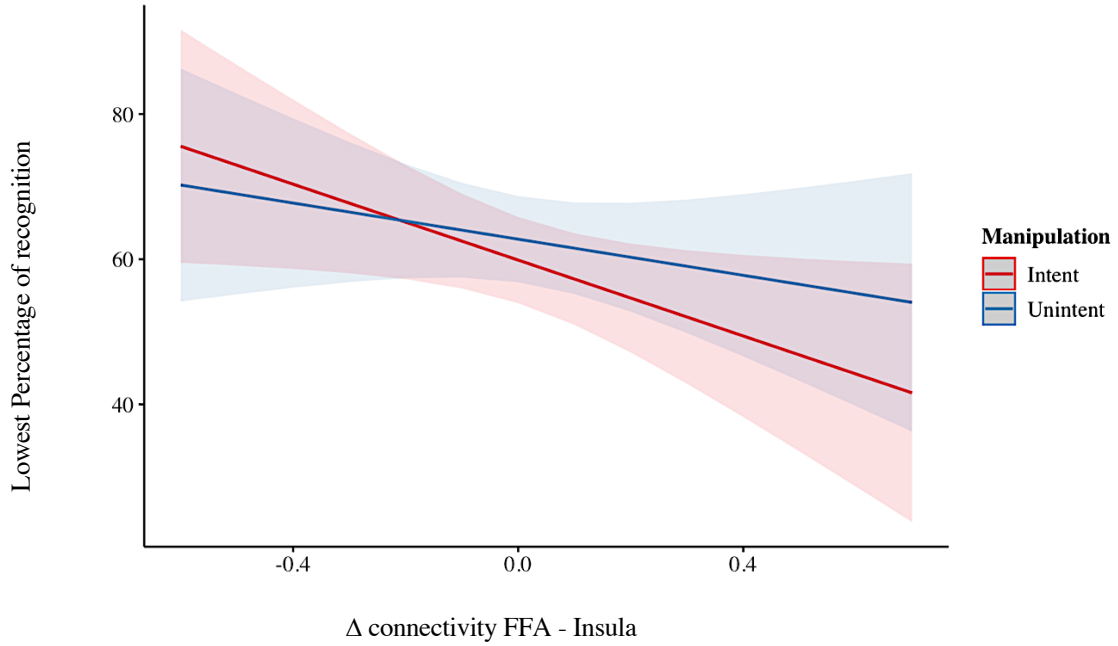

Figure 8: The lowest recognition percentage for each confederate, and its correlation between the fixed effect estimate values from the regression model for the change in connectivity between the hippocampus and the insula.

When analyzing the CS memory task to investigate memory generalization, we found that a higher Δ connectivity between the hippocampus - the amygdala correlated with generalization to $CS+_{intent}$ (Unintentionality X CS+ X Hipp – dmPFC: β = -0.36, SE = 0.13, p = 0.009) (Figure 9).

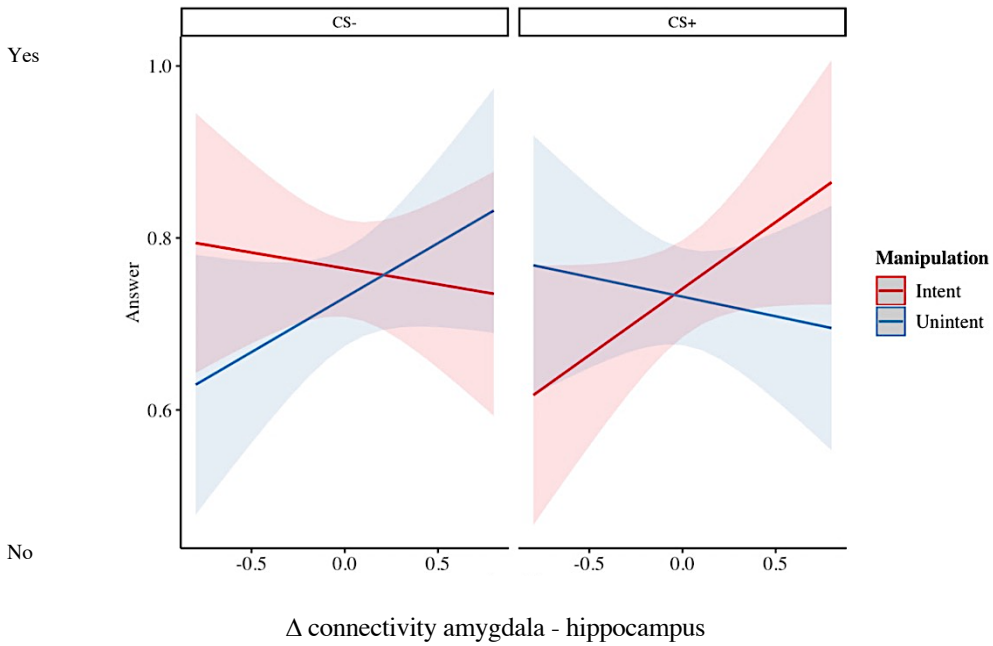

Figure 9: The memory test answers 24-hours following learning to exemplars from CS categories, and its correlation between the fixed effect estimate values from the regression model for the change in connectivity between the hippocampus and the amygdala.

To conclude, we found that the insula, a region that was involved in representing intentionality during learning, facilitated the detection of a harmful intentional agent 24-hours later, via an increased connectivity to the face processing region (FFA). In turn, an increase in connectivity between the amygdala and the hippocampus played a role in generalizing of the actions that led to intentional harm, to its category exemplars.

6 Discussion

The general aim of this thesis was to understand how social interactions can lead to aversive learning. In three studies, we investigated the contributions of brain regions commonly reported in social cognition and aversive learning tasks during an aversive social interaction. We examined the neural mechanisms at play during learning, updating and expression of fear memories related to others' actions and mind states.

In the upcoming section, I will discuss our findings regarding intentionality of an aversive action. First, I will start by presenting the behavioral results from Study I-III to set the stage for a discussion of the neural correlates we observed along with this behavior. In the following section, I will discuss our findings about neural processing during learning from Study I and II. Then, I will discuss Study III, examining the neural changes that occur after learning, and how they relate to behavior 24-hours following learning. Finally, I will present a discussion about how these studies collectively contribute to our understanding of how we learn to fear others and their actions.

6.1 Intentional vs. unintentional harm

To date, multiple studies have shown that the intentionality of another individual's actions alter judgements about them and their actions (Cushman, Sheketoff, Wharton, & Carey, 2013; Liljeholm et al., 2014; Martin & Cushman, 2016). Most of these studies are conducted from a third-person point of view, where participants read about a possible scenario about an agent, that intentionality hurts another. Albeit their strength in understanding certain aspects of our lives, like legal judgements, these fictitious scenarios do not capture the effects on intentionality when the harmful action is experienced first-hand. In this thesis, we aimed to find out the effects of an intentionally harmful action first-hand and collected information from participants about their experience. We replicated previous findings that report increased discomfort from intentionally delivered shocks (Gray & Wegner, 2008b), increased anger to an intentional and aversive action, decreased likeability for the intentional agent after the aversive encounter, and increased their desire to revenge (Liljeholm et al., 2014). In Study I and II we provide evidence of how participants also believed the frequency of the intentionally harmful act was higher than that of an unintentional one. Additionally, we show that this inflation was present immediately after learning but disappeared in the 24-hour later task in Study III.

In both Study I and Study II, participants received an identical number of shocks from the intentional and the unintentional confederate. Additionally, these two confederates made the choice between CS+ and CS- an equal number of times. Nevertheless, participants reported receiving more shocks from the intentional confederate, as well as greater discomfort from them. Past research has shown that fear-related images (such as snakes) are more readily associated with shocks than fear-irrelevant ones (such as flowers)

(Öhman & Mineka, 2001). In Study I and Study II we showed that the intentionality behind an action can lead to the same effect, causing so-called "illusory correlations". Studies on illusory correlations report increased arousal to fear-relevant stimuli, however, we did not observe such effects of intentionality in either of the studies. We found no difference in arousal towards intentionally delivered harm, compared to unintentional ones. Importantly, the preparedness theory that lays the ground to illusory correlations are currently challenged. A recent study examining 23 different aversive learning studies has found out that 69% of these studies failed to show support for this hypothesis (Åhs et al., 2018). Although we did not test for preparedness per se, we provide evidence for social information behind an aversive action, in this case the intentionality, can modulate how an aversive experience is reported (Study I and Study II) and remembered (Study III).

Surprisingly, in Study II we found no learning effects in pupillometry that reflect aversive learning (i.e., a differential increase in arousal to the CS+ compared to the CS- in time). This finding was surprising, as we saw a clear learning effect in pupillometry in Study I. One difference between the two studies was sample size. Study I pupillometry results include data from 35 individuals, whereas Study II from 23. A power analysis we conducted based on previous literature using a similar design (Visser et al., 2013) revealed that a sample size of 26 was necessary to capture the effects of learning, suggesting that Study II might be underpowered to detect the expected learning effects. Participants in the sample from Study II reported expecting shocks more from the CS+ and reported receiving more shocks from them as well. Additionally, as it will be discussed in the next section, we saw effects of CS+ and CS- differentiation in our neural measures. These findings add on to the aversive learning literature, where recently measures of fear expression have been under investigation.

6.2 Intentionality is integrated with aversive learning throughout the cortex

In both Study I and II, we replicated results from previous research that shows associating a neutral stimulus with an aversive one (i.e., US – CS relationship) leads to a refined activity pattern in regions such as the IFG and the insula (Visser, Haan, et al., 2016; Visser et al., 2011). In Study III, we added on to these findings and showed that the insula is involved in regulating the perception of intentional faces 24-hours later through its altered connectivity to the FFA.

Our findings in Study I replicate previous findings and show an increase in neural correlation patterns in response to an aversive stimulus in the insula and the ACC (Dunsmoor, Kragel, et al., 2014; Visser et al., 2013). We found additional brain regions that were not reported previously, such as the IFG, the dmPFC, the arSTS, and the prSTS. These regions have been associated with their role in mentalizing in past research (C. D. Frith & Frith, 2006; Koster-Hale & Saxe, 2013). We found tentative effects of intentionality in the

insula, the ACC and the lFG (Figure 5), however, these effects are small and did not survive FDR correction and thus must be interpreted cautiously. We did find univariate activation in the lFG in response to $CS+_{intent}$ compared to the $CS+_{unintent}$. In addition, we showed that the vmPFC is involved in representing a neural signature for the non-harmful CS- stimuli. These findings highlight the involvement of brain regions commonly reported for mentalizing in a socio-interactive aversive learning task. Our research supports previous findings that report activity in the insula in response to cognitive (Atlas & Wager, 2012) or emotional (Orenius et al., 2017) states that alter the perception of a sensory stimulus. Past research has shown that social information can alter pain perception, mediated by activity in a large network that includes the ACC and the insula (Koban, Jepma, López-Solà, & Wager, 2019).

Study II failed to replicate our previous findings about intentionality. Here, we found no evidence for the representation of intentional harmful acts during learning. Instead, we replicated findings in aversive learning one more time, showing representation of aversive stimuli in the lFG and the insula. As discussed in relationship to the behavioral findings, one striking difference between the two studies was the pupil responses during learning. We found no differential increase in pupil sizes in response to the CS+ compared to CS- during the experiment, and thus did not have the same learning index as the previous experiment.

A procedural difference between Study I and Study II was the timing of the shocks. In Study I, participants received shocks during the choice period that ended immediately before the choice the confederate made was revealed. In contrast, in Study II, participants received shocks that were delivered after the choice was revealed. In addition, for both studies we report only trials that were not reinforced with a shock. This means that participants first saw two trials of each stimulus that were not reinforced with a shock (i.e., a habituation period), where they did not know the association between shocks and CS+ images. Following that, participants went through "filler" blocks where they received an electrical stimulation, if the confederates chose a CS+ image. Following this block, participants were again presented with choices that represented each CS image, but now the CS+ was not paired with a shock (i.e., the "target" blocks, for more information on filler and target trials please refer to the manuscripts). This means that the responses we report for are trials where participants had already been presented with a CS+ that was paired with a shock but did not deliver a shock. Here, the difference between shock timing becomes important. In Study I, participants realize the lack of shock simultaneously as they see a CS+ was chosen. Thus, when we model the choice period, we are capturing the participants' responses to the lack of shock, although a shock decision was made. In Study II, we are capturing the expectation of shocks. In a trial that includes a shock the participant would receive a shock in the end, but now did not. Thus, the time leading to the shock is identical. Taken together, our neural findings tentatively point out to the role

of the ACC, the insula and the IFG in representing an intentional threatening event in the realization of the lack of an aversive outcome (Study I), but not expectation (Study II). A recent study comparing social and physical pain supports this possibility, where physical pain relief but not a social one engages the IFG (Meyer, Williams, & Eisenberger, 2015).

In Study I, we found tentative results that indicate intentionality of aversive actions led to the development of a neural signature throughout learning. In Study II, we quantified these signatures' development with templates that model the expected change in pattern correlations but failed to replicate previous findings on intentionality. These approaches were implemented as we integrate research from two different domains: aversive learning and mentalizing. Since our experimental design relies on learning about the aversive properties of an action (i.e., the choice of CS+) in the presence (or absence) of the intentionality behind it, we could observe how the aversive properties are integrated with that of intentionality. In both studies that report findings on learning, the first two trials of CS+ choices were not reinforced. This means that the participants have learned that these images are coupled to electrical shocks only after the first two presentations. This allowed us to observe these stimuli in the absence and presence of the association with shock. In Study I, we observed the integration of intentionality with that of the CS+, prominently for the trials that are following the first shock delivery in the experiment (Figure 5). In Study II, we deliberately modeled a linear increase from trial-to-trial, to capture any linear effects. Importantly, in Study I, we used data from consecutive trials. In Study II, we were able to model an increase between trials that are farther apart. Here, we modeled an increase between consecutive trials (i.e., 3-4, 4-5) and the rest of the trials (i.e., 3-5, 4-6). This means that the effects we captured in Study II also capture the stability of a neural pattern that has formed in consecutive trials. In Study I, we were lacking this aspect.

6.3 Neural changes that regulate subsequent memory

Learning leads to changes in brain connectivity and these changes can in turn affect subsequent memory (Hermans et al., 2016; Tambini & Davachi, 2013; Tambini et al., 2017). In Study III, we investigated the effects of neural connectivity changes before and after aversive learning, and its relation to memory 24-hours later. Here, we collected rs-fMRI before and after the learning task presented in Study I (Figure 2). We were specifically interested in face perception, as literature points to a generalization of the features of the perpetrator, and re-living of the traumatic event when faced with another individual with similar features in individuals with PTSD from interpersonal trauma (Ehlers & Clark, 2000). As we presented participants with both faces and other neutral stimuli that served as CS's during learning, we also investigated the generalization effect on the CS's. Past literature suggests generalization for CS+ related stimuli (Dymond et al., 2015), and here we investigate how and if the generalization can be altered by social information (i.e., intentionality).

Our findings suggest that an increased connectivity between the insula and the FFA, and the dmPFC and the FFA, after a social interaction with aversive learning, predicts recognizing the intentional confederate's face at a lower percentage, compared to the unintentional one (Figure 8). FFA has been studied thoroughly in the past years, and has been reported in brain activity in response to faces (N Kanwisher et al., 1997), and more specifically to face similarity, social traits, and gender (Tsantani et al., 2020). In Study I, we found preliminary evidence that during learning a neural signature forms in response to the intentional actions in the insula. In Study III, we showed that the same region has increased connectivity with the FFA that leads to differences in perception of intentional faces 24-hours later. Our findings are in line with the previous literature that suggests a lower perceptual threshold for stimuli related to the traumatic event in patients in PTSD (Ehlers & Clark, 2000). Patients with PTSD also exhibited biased activity in the insula, correlated with fear generalization to faces with the same emotional expression as the conditioned stimuli that was used during aversive learning (Morey et al., 2015).

Past research suggests that social information encoding (such as the name of an individual) leads to an increase in right hippocampus-MPFC connectivity, compared to non-social information (Meyer et al., 2018). Our results did not suggest a role for hippocampus-MPFC connectivity and its role in expression of memories of a social interaction. One interpretation could be that the MPFC is involved in social information encoding, such as remembering a person's age and name, and the insula is involved in the mentalizing and emotional aspects. Our study did not include biographical information about the confederates and thus we cannot make a direct comparison.

Many regions reported as a part of the DMN overlap greatly with that of the mentalizing network: the TPJ, PC, dmPFC, and vmPFC are a part of both networks (Buckner et al., 2008; Spreng & Andrews-Hanna, 2015). Recent work has unveiled a distinguishable neural counterpart of the mentalizing network and have shown that the mentalizing network can be differentiated on the level of fiber tracts (Wang et al., 2021). This means that activity captured during rest after a mentalizing related task does not necessarily reflect DMN activation (which would always be captured, no matter the task) but specific aspects related to mentalizing. In this study, the authors defined the mentalizing network as consisting of the TPJ, the precuneus, the anterior temporal lobes, the dmPFC and the vmPFC. Our set of ROIs included only the TPJ, the vmPFC and the dmPFC among these regions. Our findings regarding an increased dmPFC-FFA RSFC after learning adds on to the current literature and provides evidence for the role of dmPFC in the perceptual aspects of processing social memory.

Finally, we found evidence of the involvement of the amygdala-hippocampus RSFC in the generalizing of CS+ stimuli, specifically with intentional harm. Generalization is quantified by the transfer of behavioral responses that are specific to one exemplar, to others. This transfer often occurs across perceptually similar items and is well studied in humans as

well as non-primates (for a review see (Dymond et al., 2015)). Our results are in line with previous literature in generalization of aversive responses to the CS+ category, that indicate hippocampus-amygdala connectivity increase during learning in response to the CS+ category, regulated by the typicality of the stimulus (Dunsmoor, Kragel, et al., 2014). Recent studies in fear generalization suggest that generalization of responses to the CS+ images are correlated with connectivity changes between the hippocampus and they amygdala (Webler et al., 2021). In Study III, we provide evidence that when an aversive stimulus is experienced from an intentional confederate, the changes in hippocampus-amygdala RSFC after learning predicts specifically the generalized memory for the aversive intentional CS category, and not the others. Of note is that studies that report generalization effects that were mentioned above include physiological measures that show generalization of arousal responses to category exemplars, but we report findings that show participants reported more category exemplars that they never saw before as images they remembered.

7 Conclusions

This thesis aimed to understand the effects of social information, such as the intentionality of an individual, on learning and memory. The results we present suggests that the intentionality of an action leads both the action to be perceived worse and the individual performing the action to be judged in a more negative light. We showed preliminary evidence that suggests intentionality of an aversive action is represented by a unique neural pattern across the cortex, however, that this pattern formation during learning relied on certain factors. One of the factors was the timing of the aversive stimulus, and the other was the size of the study sample. We saw that resting state connectivity changes before and after learning can predict the expression of memory of both the actions and the individuals involved in the social emotional learning task we implemented.

Studies included in this thesis highlight the importance of social information and mentalizing during a traumatic incident, and the expression of it in the aftermath. These studies only partially cover the highly complex mechanisms that underlie emotional learning during a social interaction, and its neural correlates. However, I hope that the introduction of a novel experimental paradigm to study social aversions, as well as our findings will pave the way for future researchers examining the interaction of various social and non-social sources of information during learning and memory.

8 Points of perspective

In Study I we found tentative results for a unique neural representation of intentional harm that develops during learning, as well as CS+ representations. In Study II, we only observed increased neural correlations in response to CS+'s but not intentionality. Study I and II raise the question if physiological arousal and sample size are essential to capture the effects of intentionality. Our three studies that are reported in this thesis do not contain the necessary measures to explore that, but future research can investigate the importance of physiological arousal and sample size in learning.

In Study II, we implemented an extinction learning paradigm. Our goal was to observe how neural signatures from the learning task update to safety. To this end, we used templates that model our hypothesis of how this safety update would look like. An important addition here would be to see if neural signatures during learning are indeed the same as extinction. Future research can use a similar approach as we did and use representational similarity analysis to investigate if the same patterns are indeed activated during extinction.

Similarly, neural patterns that were captured during the learning task can be investigated during rs-fMRI. Here, we show tentative evidence that the insula-hippocampus connectivity is increased after learning and can predict subsequent memory and the insula represents intentional harm with a unique neural pattern. If these patterns are reactivated during post-learning rest can be explored in the future.

Finally, findings in this thesis are relevant to psychopathology such as PTSD, but also autism and social anxiety. Future work can investigate different patient populations to understand how and if these effects are reproduced or modified.

9 Acknowledgements

I would like to start by thanking my supervisor **Andreas Olsson**, who have provided me with the opportunity to pursue my PhD degree. Thank you for allowing me an incredible amount of ownership and creativity in my work, and trusting me to take on big projects, without which this work would not be possible.

I also thank my co-supervisors **Fredrik Åhs** and **Armita Törngren Golkar**, for the many discussions we had that helped shape the work in this thesis.

Perhaps even of equal importance to the creating of this thesis, and of me as a scientist I am today are the past and current members of the Emotion Lab: **Lisa, Philip, Ida, Tanaz, Amy, Tobias, Jessica, Troy, Tove, Björn, Jan,** and **Jonathan,** thank you for inspiring me to be a better scientist, supporting me, and creating a fun research group to work in. A special spot is reserved for **Joana,** my mentor, collaborator, and very dear friend. Thank you for the countless hours of listening to me, getting angry with me, helping me and being there for me no matter what. I couldn't have asked for a better person to "teach me something".

Additional thanks to my department buddies: **Arnaud, Danja, Georgia, Frida, Moa, Robin** for being there both at work and outside of work: the lunches, birthdays, celebrations, and the cakes are all what made this journey so special.

A special thanks to my Swedish family. Dear **Annica, Lennart,** thank you for your support and love. Your support was one of the strongest driving forces that pushed me through the finish line. I'm grateful to have you!

Patrik, it's hard to put into words the amount of strength you gave me and continue to give. Thank you for allowing me the space to be me and making every hurdle I met along the way so easy to overcome. It's a joy to explore, to be curious, to learn, to take leaps of faith, to fall and to get up again together with you.

Ve en önemlisi, canım aileme sonsuz teşekkürlerimi iletirim. Sevgili çekirdek ailem **annem**, **babam**, **teyzem** ve **Ergin**, bu zorlu süreçte bana destek verdiğiniz için ve bugüne kadarki emeğiniz için teşekkür ederim. Canım **süper babaannem**, iyi ki varsın! Her zaman olduğu gibi doktora çalışmalarım sırasında da bana ilham olmaya ve güç vermeye devam ettin.

Diğer ailem, **MBDZ**'m. İyi ki varsınız, uzaklara dağılmış da olsak her zaman bıraktığımız yerden devam edip her konuda beni desteklediğiniz için, yanımda olduğunuz için çok çok teşekkür ederim. Neredeyse hepimiz doktor oluyoruz, nereden nereye diyor ve devamını okuyanların hayal gücüne bırakıyorum.

Last but not least, **Luiza**, my soul sister. I thank the universe, the luck, the destiny, the birds, the bees, whatever that made us meet. Thank you, for being you.

10 References

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., & Lundström, J. N. (2018). Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: A systematic review. Neuroscience and Biobehavioral Reviews, 95(October), 430–437. https://doi.org/10.1016/j.neubiorev.2018.10.017

Albert, N. B., Robertson, E. M., & Miall, R. C. (2009). The Resting Human Brain and Motor Learning. Current Biology, 19(12). https://doi.org/10.1016/j.cub.2009.04.028

American Psychiatric Association, A. and others. (2013). Diagnostic and Statistical Manual of Mental Disorders. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596.dsm05

Ames, D. L., & Fiske, S. T. (2015). Perceived intent motivates people to magnify observed harms. Proceedings of the National Academy of Sciences, 112(12), 201501592. https://doi.org/10.1073/pnas.1501592112

Amodio, D. M. (2018). Social Cognition 2.0: An Interactive Memory Systems Account. Trends in Cognitive Sciences, xx, 1–13. https://doi.org/10.1016/j.tics.2018.10.002

Atlas, L. Y., & Wager, T. D. (2012). How expectations shape pain. Neuroscience Letters, 520(2), 140–148. https://doi.org/10.1016/j.neulet.2012.03.039

Bolles, R. C., & Fanselow, M. S. (1980). A perceptual-defensive-recuperative model of fear and pain. Behavioral and Brain Sciences, 3(2), 291–301. https://doi.org/10.1017/S0140525X0000491X

Boucsein, W. (2012). Electrodermal activity. Techniques in psychophysiology (Vol. 3). https://doi.org/10.1007/978-1-4614-1126-0

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. Human Brain Mapping, 40(11), 3362–3384. https://doi.org/10.1002/hbm.24603

Bradley, M. B., Miccoli, L. M., Escrig, M. a, & Lang, P. J. (2008). The pupil as a measure of emotional arousal and automatic activation. Psychophysiology, 45(4), 602. https://doi.org/10.1111/j.1469-8986.2008.00654.x.The

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(2), 379–390. https://doi.org/10.1037/0278-7393.18.2.379

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. Annals of the New York Academy of Sciences. https://doi.org/10.1196/annals.1440.011

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (2007). Handbook of psychophysiology. Cambridge University Press.

Chadwick, M. J., Bonnici, H. M., & Maguire, E. A. (2012). Decoding information in the human hippocampus: A user's guide. Neuropsychologia, 50(13), 3107–3121. https://doi.org/10.1016/j.neuropsychologia.2012.07.007

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition, 108(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. Cognition, 127(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

Davis, M. (1992). The role of the amygdala in fear and anxiety. Annual Review of Neuroscience. https://doi.org/10.1146/annurev.neuro.15.1.353

de Voogd, L. D., Fernández, G., & Hermans, E. J. (2016). Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. NeuroImage, 134, 563–572. https://doi.org/10.1016/j.neuroimage.2016.04.026

Delgado, M. R., Olsson, A., & Phelps, E. a. (2006). Extending animal models of fear conditioning to humans. Biological Psychology, 73(1), 39–48. https://doi.org/10.1016/j.biopsycho.2006.01.006

Dunbar, R. I. M., & Shultz, S. (2007). Evolution in the social brain. Science. https://doi.org/10.1126/science.1145463

Dunsmoor, J. E., Ahs, F., Zielinski, D. J., & LaBar, K. S. (2014). Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. Neurobiology of Learning and Memory, 113, 157–164. https://doi.org/10.1016/j.nlm.2014.02.010

Dunsmoor, J. E., Kragel, P. a, Martin, A., & LaBar, K. S. (2014). Aversive Learning Modulates Cortical Representations of Object Categories. Cerebral Cortex, 24(11), 2859–2872. https://doi.org/10.1093/cercor/bht138

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: how humans generalize fear. Trends in Cognitive Sciences, 1–5. https://doi.org/10.1016/j.tics.2014.12.003

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

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. Behaviour Research and Therapy, 38(4). https://doi.org/10.1016/S0005-7967(99)00123-0

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences of the United States of America, 113(28). https://doi.org/10.1073/pnas.1602413113

Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., … Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. Nature Protocols, 15(7), 2186–2202. https://doi.org/10.1038/s41596-020-0327-3

Feng, P., Zheng, Y., & Feng, T. (2015). Spontaneous brain activity following fear reminder of fear conditioning by using resting-state functional MRI. Scientific Reports, 5(1), 16701. https://doi.org/10.1038/srep16701

Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. Neuron, 50(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Frith, C. D., & Frith, U. (2005). Theory of mind. Current Biology : CB, 15(17), 644–645. https://doi.org/10.1016/j.cub.2005.08.041

Fullana, M. A., Dunsmoor, J. E., Schruers, K. R. J., Savage, H. S., Bach, D. R., & Harrison, B. J. (2020). Human fear conditioning: From neuroscience to the clinic. Behaviour Research and Therapy, 124(September 2019), 103528. https://doi.org/10.1016/j.brat.2019.103528

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

Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., … Harrison, B. J. (2018). Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. Neuroscience and Biobehavioral Reviews, 88(December 2017), 16–25. https://doi.org/10.1016/j.neubiorev.2018.03.002

Giourou, E., Skokou, M., Andrew, S. P., Alexopoulou, K., Gourzis, P., & Jelastopulu, E. (2018). Complex posttraumatic stress disorder: The need to consolidate a distinct clinical syndrome or to reevaluate features of psychiatric disorders following interpersonal trauma? World Journal of Psychiatry, 8(1). https://doi.org/10.5498/wjp.v8.i1.12

Gray, K., & Wegner, D. M. (2008a). The sting of intentional pain. Psychological Science, 19(12), 1260–1262. https://doi.org/10.1111/j.1467-9280.2008.02208.x

Gray, K., & Wegner, D. M. (2008b). The Sting of Intentional Pain. Psychological Science, 19(12), 1260–1262. https://doi.org/10.1111/j.1467-9280.2008.02208.x

Hamann, S. B. (2001). Cognitive and neural mechanisms of emotional memory. Trends in Cognitive Sciences, 5(9), 394–400. https://doi.org/10.1016/S1364-6613(00)01707-1

Hamann, S. B., Cahill, L., McGaugh, J. L., & Squire, L. R. (1997). Intact enhancement of declarative memory for emotional material in amnesia. Learning and Memory, 4(3). https://doi.org/10.1101/lm.4.3.301

Hamlin, J. K. (2012). A Developmental Perspective on the Moral Dyad. Psychological Inquiry, 23(2), 166–171. https://doi.org/10.1080/1047840X.2012.670101

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nature Neuroscience, 8(5). https://doi.org/10.1038/nn1445

Hermans, E. J., Kanen, J. W., Tambini, A., Fernández, G., Davachi, L., & Phelps, E. A. (2016). Persistence of Amygdala–Hippocampal Connectivity and Multi-Voxel Correlation Structures During Awake Rest After Fear Learning Predicts Long-Term Expression of Fear. Cerebral Cortex, bhw145. https://doi.org/10.1093/cercor/bhw145

Hitti, F. L., & Siegelbaum, S. A. (2014). The hippocampal CA2 region is essential for social memory. Nature, 508(7494), 88–92. https://doi.org/10.1038/nature13028

Joshi, S. A., Duval, E. R., Kubat, B., & Liberzon, I. (2020). A review of hippocampal activation in post-traumatic stress disorder. Psychophysiology, 57(1), 1–11. https://doi.org/10.1111/psyp.13357

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. Trends in Cognitive Sciences, 24(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005

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), 4302–4311. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9151747

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America, 107(25), 11163–11170. https://doi.org/10.1073/pnas.1005062107

Karatzias, T., Hyland, P., Bradley, A., Cloitre, M., Roberts, N. P., Bisson, J. I., & Shevlin, M. (2019). Risk factors and comorbidity of ICD-11 PTSD and complex PTSD: Findings from a trauma-exposed population based sample of adults in the United Kingdom. Depression and Anxiety. https://doi.org/10.1002/da.22934

Kindt, M., & Soeter, M. (2013). Reconsolidation in a human fear conditioning study: a test of extinction as updating mechanism. Biological Psychology, 92(1), 43–50. https://doi.org/10.1016/j.biopsycho.2011.09.016

Kitamura, T. (2017). Driving and regulating temporal association learning coordinated by entorhinal-hippocampal network. Neuroscience Research, 121, 1–6. https://doi.org/10.1016/j.neures.2017.04.005

Kleim, B., Ehlers, A., & Glucksman, E. (2007). Early predictors of chronic post-traumatic stress disorder in assault survivors. Psychological Medicine, 37(10), 1457–1467. https://doi.org/10.1017/S0033291707001006

Koban, L., Jepma, M., López-Solà, M., & Wager, T. D. (2019). Different brain networks mediate the effects of social and conditioned expectations on pain. Nature Communications, 10(1), 4096. https://doi.org/10.1038/s41467-019-11934-y

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

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. Neuron, 79(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. Neuron, 60(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kurdi, B., Krosch, A. R., & Ferguson, M. J. (2020). Implicit evaluations of moral agents reflect intent and outcome. Journal of Experimental Social Psychology, 90, 103990. https://doi.org/10.1016/j.jesp.2020.103990

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology, 30(3). https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

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 and Emotion, 24(8), 1377–1388.

Ledoux, J. E. (2000). Emotion circuits in the brain. Annual Review of Neuroscience, 23(1), 155–184. https://doi.org/2000. 23:155–184

LeDoux, J. E. (1993). Emotional memory systems in the brain. Behavioural Brain Research, 58(1–2), 69–79. https://doi.org/10.1016/0166-4328(93)90091-4

Leuchs, L., Schneider, M., Czisch, M., & Spoormaker, V. I. (2017). Neural correlates of pupil dilation during human fear learning. NeuroImage, 147, 186–197. https://doi.org/10.1016/j.neuroimage.2016.11.072

Levine, S. M., Kumpf, M., Rupprecht, R., & Schwarzbach, J. V. (2020). Supracategorical fear information revealed by aversively conditioning multiple categories. Cognitive Neuroscience, 12(1), 28–39. https://doi.org/10.1080/17588928.2020.1839039

Levine, S., Mikhail, J., & Leslie, A. M. (2018). Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. Journal of Experimental Psychology: General, 147(11), 1728–1747. https://doi.org/10.1037/xge0000459

Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2014). Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 34(34), 11339–11348. https://doi.org/10.1523/JNEUROSCI.1126-14.2014

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., … Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neuroscience and Biobehavioral Reviews. https://doi.org/10.1016/j.neubiorev.2017.02.026

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain." Frontiers in Human Neuroscience, 6(JUNE 2012). https://doi.org/10.3389/fnhum.2012.00189

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. Cognition, 147, 133–143. https://doi.org/10.1016/j.cognition.2015.11.008

Masserman, J. H., Wechkin, S., & Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. American Journal of Psychiatry, 121(6), 584–585. https://doi.org/10.1176/ajp.121.6.584

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. Social Cognitive and Affective Neuroscience, 8(6), 623–631. https://doi.org/10.1093/scan/nss040

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2018). Evidence That Default Network Connectivity During Rest Consolidates Social Information. Cerebral Cortex, (May), 1–11. https://doi.org/10.1093/cercor/bhy071

Meyer, M. L., Williams, K. D., & Eisenberger, N. I. (2015). Why Social Pain Can Live on: Different Neural Mechanisms Are Associated with Reliving Social and Physical Pain. PloS One, 10(6), e0128294. https://doi.org/10.1371/journal.pone.0128294

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. https://doi.org/10.1016/j.neuroimage.2015.07.015

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. Journal of Personality and Social Psychology, 116(2), 215–236. https://doi.org/10.1037/pspa0000137

Morey, R. A., Dunsmoor, J. E., Haswell, C. C., Brown, V. M., Vora, A., Weiner, J., … Szabo, S. T. (2015). Fear learning circuitry is biased toward generalization of fear associations in posttraumatic stress disorder. Translational Psychiatry, 5(12). https://doi.org/10.1038/tp.2015.196

Navarrete, C. D., Olsson, A., Ho, A. K., Mendes, W. B., Thomsen, L., & Sidanius, J. (2009). Fear extinction to an out-group face: The role of target gender. Psychological Science. https://doi.org/10.1111/j.1467-9280.2009.02273.x

Norman, D. A. (1970). Comments on the information structure of memory. Acta Psychologica, 33, 293–303. https://doi.org/10.1016/0001-6918(70)90141-1

Öhman, A. (1986). Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. Psychophysiology. https://doi.org/10.1111/j.1469-8986.1986.tb00608.x

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

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). Psychology: The role of social groups in the persistence of learned fear. Science. https://doi.org/10.1126/science.1113551

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. Social Cognitive and Affective Neuroscience, 2(1), 3–11. https://doi.org/10.1093/scan/nsm005

Olsson, A., & Undeger, I. (2017). Evolved Physiological Reactions. In Encyclopedia of Evolutionary Psychological Science (pp. 1–7). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-16999-6_2993-1

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

Orenius, T. I., Raij, T. T., Nuortimo, A., Näätänen, P., Lipsanen, J., & Karlsson, H. (2017). The interaction of emotion and pain in the insula and secondary somatosensory cortex. Neuroscience, 349, 185–194. https://doi.org/10.1016/j.neuroscience.2017.02.047

Pape, H. C., & Pare, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiological Reviews. https://doi.org/10.1152/physrev.00037.2009

Paré, D. (2003). Role of the basolateral amygdala in memory consolidation. Progress in Neurobiology. https://doi.org/10.1016/S0301-0082(03)00104-7

Parkinson, M., & Byrne, R. M. J. (2018). Judgments of moral responsibility and wrongness for intentional and accidental harm and purity violations. Quarterly Journal of Experimental Psychology, 71(3), 779–789. https://doi.org/10.1080/17470218.2016.1276942

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. International Journal of Human Computer Studies, 59(1–2), 185–198. https://doi.org/10.1016/S1071-5819(03)00017-X

Pavlov, I. P. (1927). Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford England.: Oxford University Press.

Phelps, E. A. (2006). Emotion and Cognition: Insights from Studies of the Human Amygdala. Annual Review of Psychology. https://doi.org/10.1146/annurev.psych.56.091103.070234

Phelps, E. A., Delgado, M. R., Nearing, K. I., & Ledoux, J. E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. Neuron, 43(6), 897–905. https://doi.org/10.1016/j.neuron.2004.08.042

Poldrack, R. A. (2007). Region of interest analysis for fMRI. Social Cognitive and Affective Neuroscience, 2(1), 67–70. https://doi.org/10.1093/scan/nsm006

Popa, D., Duvarci, S., Popescu, A. T., Léna, C., & Paré, D. (2010). Coherent amygdalocortical theta promotes fear memory consolidation during paradoxical sleep. Proceedings of the National Academy of Sciences of the United States of America, 107(14). https://doi.org/10.1073/pnas.0913016107

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. Nature Reviews Neuroscience, 20(August), 495–505. https://doi.org/10.1038/s41583-019-0179-4

Reinhard, G., & Lachnit, H. (2002). Differential conditioning of anticipatory pupillary dilation responses in humans. Biological Psychology, 60(1), 51–68. https://doi.org/10.1016/S0301-0511(02)00011-X

Reisberg, D., & Heuer, F. (1992). Remembering the details of emotional events. In Affect and Accuracy in Recall (pp. 162–190). Cambridge University Press. https://doi.org/10.1017/CBO9780511664069.009

Roesler, R., Roozendaal, B., & McGaugh, J. L. (2002). Basolateral amygdala lesions block the memory-enhancing effect of 8-Br-cAMP infused into the entorhinal cortex of rats after training. European Journal of Neuroscience, 15(5), 905–910. https://doi.org/10.1046/j.1460-9568.2002.01924.x

Schacter, D. L. (1990). Introduction to "Implicit memory: Multiple perspectives." Bulletin of the Psychonomic Society, 28(4), 338–340. https://doi.org/10.3758/BF03334038

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neuroscience and Biobehavioral Reviews, 42, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Seidenbecher, T., Laxmi, T. R., Stork, O., & Pape, H. C. (2003). Amygdalar and hippocampal theta rhythm synchronization during fear memory retrieval. Science, 301(5634). https://doi.org/10.1126/science.1085818

Seligman, M. E. P. (1971). Phobias and preparedness. Behavior Therapy. https://doi.org/10.1016/S0005-7894(71)80064-3

Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., & Thase, M. E. (2003). Do the seconds turn into hours? Relationships between sustained pupil dilation in response to emotional information and self-reported rumination. Cognitive Therapy and Research, 27(3), 365–382. https://doi.org/10.1023/A:1023974602357

Spreng, R. N., & Andrews-Hanna, J. R. (2015). The Default Network and Social Cognition. Brain Mapping: An Encyclopedic Reference, 3(January 2018), 165–169. https://doi.org/10.1016/B978-0-12-397025-1.00173-1

Tambini, A., & D'Esposito, M. (2020). Causal Contribution of Awake Post-encoding Processes to Episodic Memory Consolidation. Current Biology, 30(18), 3533-3543.e7. https://doi.org/10.1016/j.cub.2020.06.063

Tambini, A., & Davachi, L. (2013). Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. Proceedings of the National Academy of Sciences, 110(48), 19591–19596. https://doi.org/10.1073/pnas.1308499110

Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced Brain Correlations during Rest Are Related to Memory for Recent Experiences. Neuron, 65(2), 280–290. https://doi.org/10.1016/j.neuron.2010.01.001

Tambini, A., Rimmele, U., Phelps, E. A., & Davachi, L. (2016). Emotional brain states carry over and enhance future memory formation. Nature Neuroscience, 20(2). https://doi.org/10.1038/nn.4468

Tambini, A., Rimmele, U., Phelps, E. A., & Davachi, L. (2017). Emotional brain states carry over and enhance future memory formation. Nature Neuroscience, 20(2), 271–278. https://doi.org/10.1038/nn.4468

Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (2020). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. Molecular Psychiatry, 25(10), 2342–2354. https://doi.org/10.1038/s41380-019-0520-3

Thompson, R. (1986). The neurobiology of learning and memory. Science, 233(4767), 941–947. https://doi.org/10.1126/science.3738519

Tomarken, A. J., Mineka, S., & Cook, M. (1989). Fear-relevant selective associations and covariation bias. Journal of Abnormal Psychology, 98(4), 381–394. https://doi.org/10.1037/0021-843X.98.4.381

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., & Feldman Barrett, L. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. NeuroImage, 60(4), 1947–1958. https://doi.org/10.1016/j.neuroimage.2012.02.012

Tranel, D., & Damasio, H. (1994). Neuroanatomical correlates of electrodermal skin conductance responses. Psychophysiology, 31(5), 427–438. https://doi.org/10.1111/j.1469-8986.1994.tb01046.x

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2020). FFA and OFA encode distinct types of face identity information, (December 2020). https://doi.org/10.1101/2020.05.12.090878

Tulving, E. (1972). 12: Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of memory/Eds E. (pp. 381–403). New York: Academic Press.

Van de Vondervoort, J. W., & Hamlin, J. K. (2016). Evidence for Intuitive Morality: Preverbal Infants Make Sociomoral Evaluations. Child Development Perspectives, 10(3), 143–148. https://doi.org/10.1111/cdep.12175

Visser, R. M., de Haan, M. I. C., Beemsterboer, T., Haver, P., Kindt, M., & Scholte, H. S. (2016). Quantifying learning-dependent changes in the brain: Single-trial multivoxel pattern analysis requires slow event-related fMRI. Psychophysiology, 53(8), 1117–1127. https://doi.org/10.1111/psyp.12665

Visser, R. M., Haan, M. I. C. De, Scholte, H. S., & Kindt, M. (2016). Trial-by-trial analysis of BOLD-MRI patterns uncovers the formation of associative fear memory : a protocol Trial-by-trial analysis of BOLD MRI patterns uncovers the formation of associative fear memory : a protocol, (August). https://doi.org/10.13140/RG.2.1.2508.8888

Visser, R. M., Kunze, A. E., Westhoff, B., Scholte, H. S., & Kindt, M. (2015). Representational similarity analysis offers a preview of the noradrenergic modulation of long-term fear memory at the time of encoding. Psychoneuroendocrinology, 55, 8–20. https://doi.org/10.1016/j.psyneuen.2015.01.021

Visser, R. M., Scholte, H. S., Beemsterboer, T., & Kindt, M. (2013). Neural pattern similarity predicts long-term fear memory. Nature Neuroscience, 16(4), 388–390. https://doi.org/10.1038/nn.3345

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative Learning Increases Trial-by-Trial Similarity of BOLD-MRI Patterns. Journal of Neuroscience, 31(33), 12021–12028. https://doi.org/10.1523/JNEUROSCI.2178-11.2011

Wang, Y., Metoki, A., Xia, Y., Zang, Y., He, Y., & Olson, I. R. (2021). A large-scale structural and functional connectome of social mentalizing. NeuroImage, 236(March), 118115. https://doi.org/10.1016/j.neuroimage.2021.118115

Webler, R. D., Berg, H., Fhong, K., Tuominen, L., Holt, D. J., Morey, R. A., … Lissek, S. (2021). The neurobiology of human fear generalization: meta-analysis and working neural model. Neuroscience and Biobehavioral Reviews. https://doi.org/10.1016/j.neubiorev.2021.06.035

Weibert, K., & Andrews, T. J. (2015). Activity in the right fusiform face area predicts the behavioural advantage for the perception of familiar faces. Neuropsychologia, 75. https://doi.org/10.1016/j.neuropsychologia.2015.07.015

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. Behavior Research Methods, 42(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020). Mentalizing during social InterAction: A four component model. Cortex, 126, 242–252. https://doi.org/10.1016/j.cortex.2019.12.031

Wundt, W. M. (1912). An introduction to psychology. G. Allen, Limited.

Yang, Q., Shao, R., Zhang, Q., Li, C., Li, Y., Li, H., & Lee, T. (2019). When morality opposes the law: An fMRI investigation into punishment judgments for crimes with good intentions. Neuropsychologia, 127(February), 195–203. https://doi.org/10.1016/j.neuropsychologia.2019.01.020

Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., & Wager, T. (2011). NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. In Frontiers in Neuroinformatics Conference Abstract: 4th INCF Congress of Neuroinformatics (p. doi: 10.3389/conf.fninf.2011.08.00058). https://doi.org/10.3389/conf.fninf.2011.08.00058

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: An emotional binding account. Trends in Cognitive Sciences, 19(5), 259–267. https://doi.org/10.1016/j.tics.2015.02.009

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. NeuroImage, 40(4), 1912–1920. https://doi.org/10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity☆. Neuropsychologia, 47(10), 2065–2072. https://doi.org/10.1016/j.neuropsychologia.2009.03.020

Young, M. P. (1993). The organization of neural systems in the primate cerebral cortex. Proceedings of the Royal Society of London. Series B: Biological Sciences, 252(1333), 13–18. https://doi.org/10.1098/rspb.1993.0040

Yu, H., Li, J., & Zhou, X. (2015). Neural Substrates of Intention-Consequence Integration and Its Impact on Reactive Punishment in Interpersonal Transgression. Journal of Neuroscience, 35(12), 4917–4925. https://doi.org/10.1523/JNEUROSCI.3536-14.2015

Zhang, H., & Mo, L. (2016). Mentalizing and information propagation through social network: Evidence from a resting-state-fMRI study. Frontiers in Psychology, 7(NOV). https://doi.org/10.3389/fpsyg.2016.01716

Zhang, Y., Yu, H., Yin, Y., & Zhou, X. (2016). Intention Modulates the Effect of Punishment Threat in Norm Enforcement via the Lateral Orbitofrontal Cortex. Journal of Neuroscience, 36(35), 9217–9226. https://doi.org/10.1523/JNEUROSCI.0595-16.2016

Zimring, F. E. (2000). Penal Proportionality for the Young Offender: Notes on Immaturity, Capacity, and Diminished Responsibility. In Youth on Trial: A Developmental Perspective on Juvenile Justice.