UNIVERSITÉ
DE GENÈVE
FACULTÉ DE PSYCHOLOGIE
ET DES SCIENCES DE L'ÉDUCATION

Section de Psychologie

Sous la direction du Professeur David SANDER (Université de Genève)
et la co-direction du Professeur Gilles POURTOIS (Université de Gand)

RELEVANCE DETECTION AS A PSYCHOLOGICAL DETERMINANT OF EMOTIONAL LEARNING

THESE

Présentée à la
Faculté de psychologie et des sciences de l'éducation
de l'Université de Genève
pour obtenir le grade de Docteur en Psychologie

par

Yoann STUSSI

de

Genève, Suisse

Thèse No 734

GENEVE

Mai 2019

Numéro d'étudiant : 09-325-341

ACKNOWLEDGMENTS

Science is a fundamentally collaborative endeavor and process. This thesis is no exception and definitely falls in line with this notion. During all the past years, I had the opportunity to encounter many amazing and brilliant people. Through their help, advice, support, and friendship, they have greatly contributed to the present work and to making this period of my life incommensurably more pleasant and rewarding. I feel extremely grateful and lucky to have been able to conduct this project in such a stimulating and caring environment.

First and foremost, I would like to thank my thesis advisors, Profs. David Sander and Gilles Pourtois. David, you are one of the kindest and most caring people that I had the chance to ever meet. Your continuous encouragement, enthusiasm, positivity, and support were an invaluable source of motivation and helped me perform and complete this thesis. Besides your extraordinary human skills, I have also greatly learned from your scientific rigor and knowledge, as well as your critical thinking, which allowed me to become a better researcher and scientist. Gilles, thank you very much for your guidance, along with your constant optimism, encouragement, and support. Your always faster-than-light feedback and positive comments were consistently enlightening and strongly contributed to improving and strengthening the work reported in this thesis. I could not have wished for a better duo of supervisors; I thank both of you for accepting to embark on this project with me in the first place and, above all, for believing in me.

I would also like to express my gratitude to Prof. Tobias Brosch. I am thankful for letting me the opportunity back in the time to start studying emotional learning and to give electric shocks to people ("for science") when I was still an innocent Master's student. I really appreciate all the valuable feedback and advice that you gave me through the years about my work and scientific career. Finally, thank you for accepting to be a member of my thesis commission and jury.

Thank you very much as well to Prof. Mauricio R. Delgado for kindly accepting to be part of my jury. It is truly a great pleasure and honor to have you as a jury member.

A special thanks goes to Prof. Andreas Olsson for welcoming me in his laboratory at Karolinska Institutet for a six-month internship during my PhD. I really enjoyed visiting your laboratory and collaborating with you on one of the studies included in thesis. Thank you for your enthusiasm, along with your expertise and knowledge on emotional learning, which I greatly benefited from.

I would also like to deeply thank Dr. Sylvain Delplanque for his continued help and support over the years. I owe you a debt of gratitude for teaching me about everything I know on psychophysiology, as well as for always being keen on sharing your scientific knowledge. Many thanks for answering to any (more or less weird) questions that I have come up with and which I did not know who else to ask. I am also sincerely thankful for your insightful feedback and guidance, and for the fact that you have always been present when I needed it.

In addition, I thank all the members, past and present, of the Laboratory for the study of Emotion Elicitation and Expression, with whom I had the pleasure to work with: Catherine Audrin, Tiffany Baer, Tobias Brosch, Patricia Cernadas Curotto, Boris Cheval, Chiara Chilla, Géraldine Coppin, Florian Cova, Sylvain Delplanque, Danny Dukes, Matthieu Ischer, Olga Klimecki, Nathalie Mella, Alison Montagrin, Christian Mumenthaler, Ryan Murray, Aline Pichon, Eva R. Pool, Gwladys Rey, Jeanne Richard, Andrea Samson, David Sander, Vanessa Sennwald, and Alexandra Zaharia. Thank you as well to the students that I had the chance to supervise: Seline Coraj, Chloé Da Silva Coelho, Aude Ferrero, and Alessio Giarrizzo. A particular thank you to Aude for your humor and for genuinely enjoying my dad jokes (as bad as they can be); to Aline for your cheeriness; to Catherine for your constant swearing and subsequent apologizing that inevitably set up a nice and funny atmosphere; to Christian for acting as a mentor when I first joined the lab as a research assistant and for the many opportunities to learn from you; to Géraldine for your advice and your help with the twists and turns of ethics committees; to Matthieu for your lack of inhibition and your social-butterfly skills; to Ryan for your kindness and support; and to Patricia for your cheerfulness, thoughtfulness, and optimism (the lab is in good hands with you).

My appreciation extends as well to my colleagues and friends at the Swiss Center for Affective Sciences and at the Faculty of Psychology and Educational Sciences for their social support and for contributing to establishing a pleasant working environment: Bruno Bonet, Leo Ceravolo, Beatrice Conte, Giada Dirupo, Kim Doell, Didier Grandjean, François Jaquet, Stephanie Mertens, Ben Meuleman, Marcello Mortillaro, Cristina Soriano, Céline Tarditi, Fabrice Teroni, along with all the others, and – of course – Marion Gumy, Sandrine Perruchoud, Daniela Sauge, Sylvain Tailamée, and Carole Varone. In particular, thank you very much to Beatrice and Stephanie for our sharing sessions on the PhD life's ups and downs and for being wonderful mates in the framework of the Swiss Doctoral School in Affective Sciences and beyond – I really enjoyed the time that we spent together; to Céline for always

being honest and for your continued support; to Giada for your kindness and radiant personality; and to Leo for your cheerfulness and the walking breaks around Unimail.

This acknowledgments section would obviously be incomplete without thanking Eva R. Pool and Vanessa Sennwald (a.k.a. Evan) at least a million times. Eva, thank you for our countless scientific and nonscientific conversations that were invariably fun, for sharing your knowledge with such excitement and providing me a lot of opportunities to learn from you, for always giving relevant feedback in the clearest imaginable way, for your never-ending optimism and thoughtfulness, and for being the loveliest nerd ever to exist. Vanessa, thank you for your sharp humor and awesome personality, for all your help and your pertinent comments on my work, as well as for being there on so many occasions to cheer me up and being so caring and understanding. Thank you to both of you for being the best and brightest colleagues and friends that one could dream of, for your endless support, and for making my everyday life infinitely easier and funnier. A lot of heartfelt thanks to Thomas Stefanelli too for being such a great and humorous friend over the years.

Last but not least, I would like to express my deepest gratitude to my family for their unconditional (and unconditioned) love and support.

FOREWORD

The research reported in the present thesis was supported by a Doc.CH grant from the Swiss National Science Foundation (P0GEP1_159057).

ABSTRACT

Emotional learning is a pivotal process that enables organisms to predict and anticipate stimuli with high importance in the environment, thereby helping them flexibly shape appropriate behaviors promoting survival and well-being. Surprisingly, mechanisms underlying preferential emotional learning remain however unclear. Previous research has suggested that only specific stimuli that have threatened survival across evolution are learned preferentially. Here, we seek to challenge this view by testing an alternative theoretical model deriving from appraisal theories of emotion, which holds that stimuli detected as relevant to the organism's concerns benefit from preferential emotional learning independently of their valence and evolutionary status per se. Across a series of empirical studies, we provide evidence showing that (a) similar to threat-relevant stimuli, positive stimuli with affective relevance are likewise readily and persistently associated with a naturally aversive event during Pavlovian aversive conditioning, (b) initially neutral stimuli with no inherent threat value and biological evolutionary significance can induce facilitated Pavlovian aversive conditioning after being temporarily associated with higher goal-relevance relative to goal-irrelevant stimuli, and (c) such preferential learning critically depends on inter-individual differences in affect and motivation. Additionally, we show that (d) the postauricular reflex is a sensitive psychophysiological measure of human Pavlovian appetitive conditioning, and could be used to establish whether our results can generalize to appetitive learning. These findings suggest that learning biases in Pavlovian conditioning are driven by a general mechanism of relevance detection that is not specific to threat, and contribute to fostering new insights into the basic mechanisms underlying emotional learning.

TABLE OF CONTENTS

1. INTRODUCTION & OVERVIEW

Learning is an adaptation that helps organisms navigate, survive, and reproduce in a changing environment. It enables organisms to produce behaviors promoting the avoidance of dangers and the procurement of rewards through the prediction and anticipation of impending threats and rewards in the environment. A central influence on learning is represented by emotions. Emotions are event-focused, two-step, rapid processes involving (a) a relevance-based emotion elicitation mechanism that (b) shapes a multicomponential emotional response, encompassing action tendency, autonomic reaction, expression, and subjective feeling (Sander, 2013). Emotions are generally highly adaptive in that they allow for responding flexibly to environmental contingencies by decoupling stimulus and response (Scherer, 1994), thus facilitating action (e.g., approach or avoidance behaviors) in situations relevant to the organism, as well as modulating cognitive processes, such as attention, learning, memory, and decision-making (see, e.g., Brosch, Scherer, Grandjean, & Sander, 2013; Sander, 2013). In that sense, learning and emotion are closely intertwined phenomena that are critical in enhancing organisms' survival and well-being. Importantly, a pivotal process in which learning and emotion inextricably interact is emotional learning, which refers to the process whereby a stimulus acquires an emotional value (e.g., Phelps, 2006) or whereby a stimulus' emotional value is updated.

Emotional learning is mainly studied by means of Pavlovian conditioning (Pavlov, 1927; Phelps, 2006). It consists of one of the most fundamental forms of learning in the animal kingdom, which is ubiquitous across a large variety of species ranging from simple (e.g., fruit flies and marine snails) to more complex (e.g., rats and humans) organisms (LeDoux, 1994). In Pavlovian conditioning, the organism learns to associate an environmental stimulus (the conditioned stimulus) with a motivationally significant outcome (the unconditioned stimulus). Through single or repeated contingent pairing with the unconditioned stimulus, the conditioned stimulus acquires a predictive and emotional value, and comes to elicit a preparatory response (the conditioned response; Pavlov, 1927; Rescorla, 1988b).

Pavlovian conditioning has substantially contributed to advancing our knowledge of learning, memory, and emotion, along with their complex interactions and neurobiological underpinnings (Büchel, Morris, Dolan, & Friston, 1998; Dunsmoor, Murty, Davachi, & Phelps, 2015; Dunsmoor, Niv, Daw, & Phelps, 2015; LaBar & Cabeza, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; LeDoux, 2000, 2012, 2014; Phelps, 2006; Phelps & LeDoux, 2005). Pavlovian aversive conditioning has notably helped outline the psychological and brain mechanisms responsible for the development, expression, and modification of fear and

defensive responses, as well as assess whether animal models of fear can be applied to humans (Delgado, Olsson, & Phelps, 2006; Phelps & LeDoux, 2005). This research has highlighted the fundamental role of the amygdala in the acquisition, storage, expression, and extinction of conditioned threat responses and memories (e.g., Büchel et al., 1998; LaBar & Cabeza, 2006; LaBar et al., 1998; LeDoux, 2000; Phelps, 2006; Phelps, Delgado, Nearing, & LeDoux, 2004; Phelps & LeDoux, 2005), as well as the involvement of the ventromedial prefrontal cortex in the retention of extinction learning (Phelps et al., 2004). These findings thereby substantiated that the neural substrates underlying aversive learning are highly conserved across species and that animal models of fear learning can largely be translated to humans (Delgado et al., 2006; Phelps & LeDoux, 2005). Additionally, Pavlovian aversive conditioning processes are considered to represent a crucial mechanism in the etiology, maintenance, treatment, and relapse of fear-related clinical disorders, such as anxiety disorders and specific phobias, hence serving as a valid laboratory or experimental model thereof (Lissek et al., 2005; Milad & Quirk, 2012; Mineka & Zinbarg, 2006; Seligman, 1971).

Interestingly, whereas Pavlovian aversive conditioning has drawn a large interest in the study of emotion, the role of Pavlovian appetitive conditioning has, however, been rarely investigated systematically in humans by comparison (e.g., Martin-Soelch, Linthicum, & Ernst, 2007). Animal research on Pavlovian appetitive conditioning has indeed been mainly related to basic learning processes rather than to emotion (e.g., Hull, 1943), animal models of positive emotions being scarcer than animal models of fear for instance (but see, e.g., Berridge & Robinson, 2003). Moreover, Pavlovian appetitive conditioning has been suggested to be more complex to study in humans, thus explaining this asymmetry. This complexity is in particular exemplified by the difficulty in finding appropriate appetitive stimuli that are able to elicit physiological responses that are similarly intense to the ones elicited by the aversive unconditioned stimuli, such as electric stimulations used in Pavlovian aversive conditioning (Hermann, Ziegler, Brimbauer, & Flor, 2000; Martin-Soelch et al., 2007) and/or a possible lack of sensitivity of the psychophysiological measures commonly used to systematically detect appetitive conditioned responses (Stussi, Delplanque, Coraj, Pourtois, & Sander, 2018). Accordingly, developing and validating sensitive psychophysiological indicators of human Pavlovian appetitive conditioning is important to eventually remedy the relative scarcity of knowledge in the study thereof.

In general, research on Pavlovian conditioning has sought to uncover the general principles of learning, delineating in particular the key role of two computational learning

signals in associative learning: prediction error and stimulus' associability (e.g., Mackintosh, 1975; Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Schultz, Dayan, & Montague, 1997). Prediction error corresponds to the discrepancy between the actual and the predicted outcome. It is a critical signal in driving learning: Organisms learn when there is a prediction error (Rescorla & Wagner, 1972). Prediction errors arise when the actual outcome is not predicted or more than predicted by the conditioned stimulus in presence (i.e., positive prediction error), thus triggering excitatory learning that increases the conditioned stimulus' predictive value, or when the actual outcome is omitted or less than predicted by the conditioned stimulus (i.e., negative prediction error), thereby eliciting inhibitory learning diminishing the conditioned stimulus' predictive value. By contrast, no learning takes place when the observed outcome is perfectly predicted by the conditioned stimulus, its predictive value remaining unchanged as a result. Neural correlates of prediction-error signals have been observed in midbrain dopaminergic neurons (especially for reward prediction error, Schultz et al., 1997; but see M. Matsumoto & Hikosaka, 2009), the striatum (e.g., Delgado, Li, Schiller, & Phelps, 2008; Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003), and the amygdala (e.g., Boll, Garner, Gluth, Finsterbusch, & Büchel, 2013; see also Niv & Schoenbaum, 2008). As for stimulus' associability, it refers to the amount of attention paid to the conditioned stimulus as a function of the extent to which it is a reliable predictor of the outcome (e.g., Mackintosh, 1975; Le Pelley, 2004; Li et al., 2011; Pearce & Hall, 1980). Associability modulates learning by influencing the conditioned stimulus' effectiveness to be established as a predictive signal of the outcome, with stimuli that are better attended to (i.e., with a high associability) being more easily associated with the outcome (Le Pelley, 2004; Mackintosh, 1975; Pearce & Hall, 1980). The computations of associability have been reported to principally involve the amygdala (Boll et al., 2013; Li et al., 2011; M. Matsumoto & Hikosaka, 2009).

Nonetheless, this line of research has generally omitted to consider the relative importance of the stimuli at stake for the organism. Although early learning theorists initially posited that all stimuli can be associated with equal ease regardless of their nature (e.g., Pavlov, 1927; Watson & Rayner, 1920), certain associations have been revealed to be more easily formed and maintained than others (Garcia & Koelling, 1966; Öhman & Mineka, 2001; Seligman, 1970, 1971), thus reflecting the existence of learning biases. Surprisingly, mechanisms underlying such preferential emotional learning remain yet unclear. Influential theoretical models put forward to account for these preferential associations, such as the

preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, adopt an evolutionary perspective according to which organisms are biologically prepared to preferentially associate stimuli that have threatened survival across evolution with naturally aversive events. Consistent with this view, a series of empirical studies have shown that "evolutionarily prepared" threat stimuli – such as snakes, angry faces, or outgroup faces – are more readily and persistently associated with an aversive outcome than threat-irrelevant stimuli – such as birds, happy faces, or ingroup faces (e.g., Atlas & Phelps, 2018; Ho & Lipp, 2014; Öhman & Dimberg, 1978; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001; Olsson, Ebert, Banaji, & Phelps, 2005). Extending preparedness theory, Öhman and Mineka (2001) further proposed that the preferential processing of, and emotional learning to, evolutionary threat-relevant stimuli would be specifically subserved by an evolved fear module centered on the amygdala in the human brain, allowing the organism to readily detect and react to these stimuli. Thus, the preparedness and fear module theories emphasize the importance of negative stimuli carrying threat-related information from phylogenetic origin in emotional learning, suggesting that preferential emotional learning is underlain by an evolved threat-specific mechanism.

In contrast, we offer here a different view by suggesting that preferential emotional learning is not specific to evolutionary threat-related stimuli but can extend to all stimuli that are relevant to the organism's concerns, such as their needs, goals, motives, values, or well-being (Frijda, 1986, 1988). Deriving from appraisal theories of emotion (e.g., Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005, 2018), this alternative model holds that such preferential learning is driven by a general mechanism of relevance detection as opposed to a threat-specific mechanism (Stussi, Brosch, & Sander, 2015; Stussi, Ferrero, Pourtois, & Sander, in press; Stussi, Pourtois, & Sander, 2018). Relevance detection is a rapid evaluation process, which enables the organism to appraise, detect, and determine whether a stimulus encountered in the environment is relevant to the their concerns (Frijda, 1986, 1988; Pool, Brosch, Delplanque, & Sander, 2016; Sander et al., 2003, 2005). According to this model, threat-relevant stimuli from evolutionary origin are preferentially processed and learned not because they have been associated with threat through evolution, but because they are highly relevant to the organism's survival. More specifically, the relevance detection model predicts that stimuli that are detected as relevant to the organism's concerns benefit from preferential emotional learning, independently of their valence and evolutionary status per se. Accordingly, relevance detection may provide a promising theoretical framework to move beyond the fear-

centered view positing that only threat-related stimuli are preferentially learned, and thereby foster new insights into the understanding of emotional learning in humans.

An overview of the major aims and structure of this thesis

The purpose of this thesis is to investigate the links between the appraisal processes involved in emotion elicitation and the basic mechanisms underlying learning in humans. More precisely, our goal is to empirically assess whether relevance detection is a general determinant of emotional learning in humans, as well as establish and characterize its role therein. To do so, we conducted a series of experiments in healthy adult participants that aimed to systematically test the theoretical prediction deriving from appraisal theories of emotion, which states that stimuli that are detected as highly relevant to the organism's concerns are preferentially learned during Pavlovian conditioning, independently of their intrinsic valence and evolutionary status per se. These experiments had the following objectives: (a) to examine whether, similar to evolutionary threat-relevant stimuli, positive stimuli that are biologically relevant to the organism are likewise preferentially conditioned to threat during Pavlovian aversive conditioning (Studies 1 and 2), (b) to characterize the influence of the stimulus' affective relevance on Pavlovian aversive learning (Studies 1 and 2), (c) to assess whether preferential Pavlovian aversive conditioning extends to stimuli detected as relevant to the organism's concerns beyond biological and evolutionary considerations (Study 3), (d) to investigate the role of inter-individual differences in the organism's concerns in preferential Pavlovian aversive conditioning (Studies 2 and 3), and (e) to test and validate a new psychophysiological measure of Pavlovian appetitive conditioning in humans that could be subsequently used to investigate the generality of a relevance detection mechanism in appetitive learning (Study 4).

In this perspective, the present thesis is structured as follows: In the theoretical part (chapter 2), we first define and delimit the concept of emotional learning as used in the context of this thesis. We subsequently present the basic principles of Pavlovian conditioning, the main paradigms used to study Pavlovian conditioning in humans, and the major formal models of Pavlovian conditioning. We then introduce the notion of preferential emotional learning and the dominant theoretical models thereof, namely the preparedness and fear module theories, and review evidence in the Pavlovian conditioning literature supporting and challenging the core assumption of these models, which posit that only stimuli that have posed threats to survival across evolution are preferentially conditioned to threat. Afterward, we elaborate the relevance detection framework based on appraisal theories, which provides an alternative

model to the biological preparedness perspective. Finally, we delineate the thesis objectives specified above in more detail.

In the experimental part (chapter 3), we report the series of experiments performed to assess the role of relevance detection in human emotional learning. In brief, Study 1 investigated across three experiments whether, similar to biologically threat-relevant stimuli (angry faces and snakes), positive emotional stimuli (baby faces and erotic stimuli) are more readily associated with an aversive event (electric stimulation) during Pavlovian aversive conditioning than neutral stimuli with less relevance (neutral faces and colored squares). Study 2 examined whether, similar to angry faces, preferential Pavlovian aversive conditioning may be observed to happy faces, and the role of inter-individual differences therein. Study 3 assessed whether preferential Pavlovian aversive conditioning can occur to initially neutral stimuli devoid of any inherent biological evolutionary significance, but acquiring goal-relevance through experimental manipulation, and whether such preferential learning is modulated by inter-individual differences in achievement motivation. Study 4 investigated whether the postauricular reflex – a vestigial muscle microreflex that is potentiated by pleasant stimuli relative to neutral and unpleasant stimuli – may provide a valid psychophysiological indicator of Pavlovian appetitive conditioning in humans.

To conclude, in the general discussion (chapter 4) we integrate the findings of our empirical studies within the theoretical framework outlined in the theoretical part. There, we seek to discuss the contribution of this work to the conceptualization of the basic mechanisms underlying emotional learning in humans and the role of relevance detection in the modulation of cognitive functions. We also outline the limitations of this thesis and elaborate on potential new avenues for future research in this area.

2. THEORETICAL PART

2.1. EMOTIONAL LEARNING

Emotional learning refers to a fundamental mental process by which a neutral stimulus or behavior acquires an emotional value, or by which a stimulus' or behavior's emotional value is updated. It is a pivotal adaptive function that enables organisms to predict and detect stimuli with high significance in the environment, and flexibly shape appropriate behaviors in response to these stimuli promoting the organism's survival and well-being.

Two fundamental forms of emotional learning are Pavlovian (or classical) conditioning and instrumental (or operant) conditioning. Pavlovian conditioning is concerned with how organisms learn about stimuli in their environment, and especially with their predictive relationship. Instrumental conditioning is related to how organisms learn about the relations between behaviors and their consequences. Specifically, in Pavlovian conditioning organisms learn that a stimulus predicts the occurrence of a motivationally significant outcome, such as food or pain (Bouton, 2007). It therefore refers to the association between two stimuli (i.e., the conditioned stimulus and the unconditioned stimulus), which comes to elicit a preparatory response (i.e., the conditioned response) allowing the organism to prepare for the upcoming reinforcing event before its actual occurrence (e.g., Domjan, 2005). By comparison, in instrumental conditioning organisms learn to associate a behavior (or a response) with a motivationally significant outcome, the reinforcing properties of which either increase (when an appetitive stimulus is obtained [positive reinforcement] or when an aversive stimulus is omitted or withdrawn [negative reinforcement]) or decrease (when an aversive stimulus is obtained [positive punishment] or when an appetitive stimulus is omitted or withdrawn [negative punishment]) the probability or frequency of the behavior according to the law of effect (Thorndike, 1898). Instrumental conditioning thereby allows organisms to learn to perform behaviors that enhance their fitness (Bouton, 2007), and to control the achievement of their goals and needs. Although Pavlovian and instrumental conditioning share many similarities (e.g., occurrence of extinction processes when the reinforcer is no longer delivered, sensitivity to the reinforcer magnitude and timing), behavioral and neurobiological evidence has shown some degree of independence between these learning systems (e.g., O'Doherty, Cockburn, & Pauli, 2017; Rangel, Camerer, & Montague, 2008; Rescorla & Solomon, 1967). As we exclusively examined Pavlovian conditioning in the experimental part of this thesis (see chapter 3), we hereafter focus on this learning process by describing its basic principles and determining conditions, as well as the main paradigms used in humans to investigate it and the major formal models thereof.

2.2. PAVLOVIAN CONDITIONING

2.2.1. Basic principles

Pavlovian conditioning was originally discovered by Pavlov (e.g., Pavlov, 1927). In his research on the physiology of dog's digestion, he observed that by repeatedly pairing an initially neutral stimulus (e.g., a metronome sound) with the delivery of food, these animals started to salivate in response to the stimulus, rather than merely salivating in the presence of the food. Pavlov interpreted this finding as reflecting the fact that the stimulus progressively acquired the ability to elicit an anticipatory salivary response through its systematic association with the food.

Formally, Pavlovian conditioning corresponds to the learning process and procedure by which an environmental stimulus (the conditioned stimulus, CS) is associated with a motivationally salient aversive or appetitive event (the unconditioned stimulus, US) by virtue of a single or repeated contingent pairing (Fanselow & Wassum, 2016; Pavlov, 1927; Rescorla, 1988b; see Figure 2.1). Usually (but not necessarily), the CS is an initially neutral stimulus (e.g., a tone or a light) that does not evoke a specific response (except orienting in some situations) prior to conditioning; by contrast, the US is biologically potent (e.g., electric stimulation or pleasant food) and automatically triggers an emotional response (the unconditioned response, UR) without prior learning, such as physiological reactions (e.g., increased skin conductance, heart rate, blood pressure, or salivation). After its pairing with the US, the organism learns that the CS predicts the US, and the presentation of the CS alone produces a conditioned response (CR), which often encompasses a constellation of emotional reactions (e.g., Phelps, 2006). These reactions can generally be observed at the behavioral (e.g., freezing, avoidance or approach behaviors), physiological (e.g., stress hormone release, increased startle responses, skin conductance, heart rate, or salivation), neural (e.g., increase in BOLD signal in the amygdala), and subjective (e.g., feelings of fear or pleasure) levels (e.g., LeDoux, 2012; Phelps, 2006). Importantly, the responses evoked by the CS are not identical to those elicited by the US. Albeit generally similar, the CR and the UR may differ, or even be opposite in certain cases. For instance, studies on conditioned analgesia have shown that whereas an electric stimulation (i.e., the US) typically elicited freezing behaviors in rats (i.e., the UR), the presentation of the tone that was established as a predictive cue for the electric stimulation (i.e., the CS) triggered endorphin release (i.e., the CR), thereby preparing the organism to cope with the upcoming electric stimulation by diminishing their pain sensitivity,

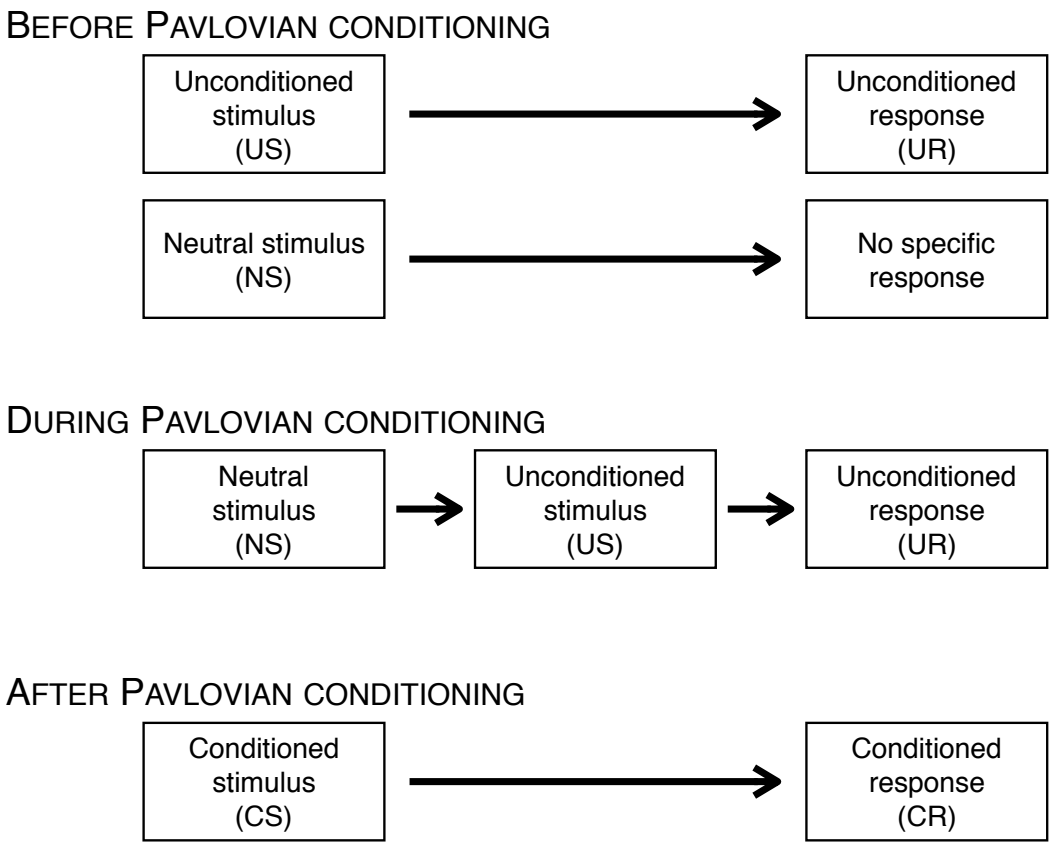

Figure 2.1. Basic principles of Pavlovian conditioning. Through a single or repeated contingent pairing with the unconditioned stimulus (US) that automatically triggers an unconditioned response (UR) without prior learning, a typically (but not necessarily) initially neutral stimulus (NS) becomes conditioned (conditioned stimulus, CS), thereby acquiring an emotional and predictive value eliciting a preparatory response (conditioned response, CR).

which ultimately resulted in a reduction in their amount of freezing (e.g., Fanselow & Bolles, 1979). This result notably indicates that the key mechanism in Pavlovian conditioning pertains to the association between the CS and the US rather than the CS substituting for the US in eliciting the UR in a reflexive manner as initially proposed by Pavlov (1927).

CS-US contingency

Correspondingly, the relation between the CS and the US is a central aspect of Pavlovian conditioning. Whereas the mere co-occurrence or contiguity between the CS and the US was initially considered as the determining condition for Pavlovian learning (e.g., Pavlov, 1927), it has been demonstrated that such contiguity is neither sufficient nor necessary. Instead, Pavlovian conditioning only occurs if the CS has a predictive relationship with the US (i.e., the presentation of the US is contingent upon the occurrence of the CS; Rescorla, 1967, 1988b).

For instance, the importance of the CS-US contingency is illustrated by Rescorla's (1968) experiment, in which he trained three groups of rats to associate a tone CS with an electric stimulation as the US. In the first group, the probability of the US was greater when the CS was presented than when it was absent (i.e., positive contingency). In the second group, the probability of the US was the same whether or not the CS was presented (i.e., zero contingency), the number of CS-US pairings being equivalent as in the first group and additional US presentations being delivered during the intertrial interval. In the third group, the probability of the US was lower when the CS was presented than when it was absent (i.e., negative contingency). Results showed that excitatory conditioning to the CS occurred in the first group, while inhibitory conditioning happened in the third group, the CS becoming a safety signal. By contrast, no conditioning was observed in the second group, even though an identical number of CS-US pairings was received as in the first group, which indicates that Pavlovian conditioning only occurred when there was a contingency relationship between the CS and the US.

Another example of the CS-US contingency relevance is the blocking effect (Kamin, 1968, 1969). *Blocking* refers to the finding that the conditioning of the association between a conditioned stimulus, CS_A, and the US is impaired if CS_A is presented in compound with another conditioned stimulus, CS_B, that has been previously associated with the US during conditioning trials, CS_B "blocking" conditioning to CS_A (Bouton, 2007; Fanselow & Wassum, 2016; Kamin, 1969). This effect suggests that simply pairing a CS with a US is not sufficient for producing Pavlovian conditioning, and that conditioning is achieved only when a CS has informational value (i.e., it provides new information about the US beyond what is already predicted by other CSs; Bouton, 2007). It is worth noting that explicit awareness of the contingency between the CS and the US is, however, likely not necessary for Pavlovian conditioning to occur in animals (see, e.g., Papini & Bitterman, 1990) and in humans (e.g., Bechara et al., 1995), although the specific role of contingency awareness in humans remains debated (e.g., Lovibond & Shanks, 2002; Öhman & Mineka, 2001).

Extinction

In addition to acquisition (i.e., learning resulting from the association between the CS and the US), an essential phenomenon observed in Pavlovian conditioning is extinction. It occurs when the US is no longer delivered after the CS and/or when the CS is no longer predictive of the US, which results in a gradual weakening or decrease in the probability of the CR occurrence over time. Extinction refers both to the procedure of presenting the CS in

absence of the US and to the phenomenon resulting from that procedure. It constitutes a crucial process in behavior change (e.g., Bouton, 2007; Dunsmoor, Niv, et al., 2015). Extinction indeed allows the organism to adapt to a changing environment by stopping producing responses and behaviors that are no longer reinforced. Extinction processes also have a high clinical significance for the treatment of a variety of psychiatric conditions (e.g., Dunsmoor, Niv, et al., 2015; Milad & Quirk, 2012), as well as serve as the basis for exposure therapy, which is one of the most effective treatment for anxiety disorders, phobias, stress-related disorders, and addictions (see, e.g., Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014; Dunsmoor, Niv, et al., 2015).

Importantly, extinction has been theorized to involve unlearning or erasure of the original CS-US association (e.g., Rescorla & Wagner, 1972), or, alternatively, to induce new inhibitory learning between the CS and the US, the CS thereby acquiring inhibitory properties that reduce or suppress the CR (e.g., Bouton, 2002; Bouton, Westbrook, Corcoran, & Maren, 2006; Konorski, 1967; Pearce & Hall, 1980). In line with the latter view, research on extinction has unveiled the existence of at least four phenomena suggesting that original learning (i.e., acquisition) is not merely erased: (1) spontaneous recovery, which consists of the recovery of the CR when the CS is tested after time has passed following the end of extinction training, (2) reinstatement, which refers to the CR recovery occurring after the organism is exposed to the US alone after extinction, (3) renewal, which corresponds to the CR recovery that can happen when the context is changed after extinction, and (4) rapid reacquisition, which pertains to the rapid CR recovery when the CS is paired with the US again after extinction (Bouton, 2002). Nevertheless, recent theoretical advances in the study of computational mechanisms underlying extinction processes have suggested that extinction may induce both the formation of new inhibitory learning that interferes with the original excitatory CS-US association and attenuation or updating of this association with new information (Dunsmoor, Niv, et al., 2015; Gershman, Blei, & Niv, 2010; Gershman & Hartley, 2015; Gershman, Monfils, Norman, & Niv, 2017; Gershman & Niv, 2012).

Further, extinction learning is more fragile than the original CS-US association acquired through Pavlovian conditioning, as well as more sensitive to contextual information (Bouton, 2002; Bouton et al., 2006; Dunsmoor, Niv, et al., 2015). The asymmetry between extinction fragility and conditioning strength and persistence is however likely adaptive: Although signals for danger or reward may not always, or even only rarely, contingently co-occur with an actual threatening or rewarding event in the environment, producing a rapid

defensive or approach response, respectively, remains critical in promoting survival when the occurrence of threat or reward is still possible (Dunsmoor, Niv, et al., 2015). Maintaining a memory trace of the original CS-US association may therefore help prepare the organism against the remote possibility of future threat or reward (Dunsmoor, Niv, et al, 2015).

2.2.2. Factors influencing Pavlovian conditioning

Whereas a central aim of this thesis is to elucidate the mechanisms underlying why some stimuli are preferentially learned above others during Pavlovian conditioning, investigation of Pavlovian conditioning in animals and in humans has been mainly guided by the aim of uncovering the general laws of learning that apply across different kinds of stimuli, rather than highlighting differences between classes of stimuli. This line of research has notably identified a number of key variables that exert an influence on Pavlovian learning. Before addressing in more depth the theoretical models advanced to explain the occurrence of preferential Pavlovian conditioning in response to certain stimuli (see chapter 2.3), we briefly detail these key factors below.

Stimulus novelty and intensity

The CS and the US have been demonstrated to be more rapidly and effectively learned when they are novel at the beginning of conditioning (Bouton, 2007; Nasser & Delamater, 2016). Preexposure to the CS before Pavlovian conditioning can indeed interfere with learning by reducing the rapidity with which organisms learn about the CS-US association when they are subsequently paired, an effect called *latent inhibition* (e.g., Lubow, 1973; Vaitl & Lipp, 1997). Similarly, preexposure to the US alone before conditioning can reduce its effectiveness as a US and delay subsequent Pavlovian learning (Bouton, 2007), this effect being referred to as *US preexposure effect* (Randich & LeLordo, 1979).

The US intensity also critically influences Pavlovian learning, with more intense USs producing faster and stronger Pavlovian conditioning (Bouton, 2007; Nasser & Delamater, 2016; Rescorla & Wagner, 1972). Likewise, the intensity or "salience"[1] of the CS plays an

[1] It is important to note that the construct of salience here refers to the stimulus' physical properties (e.g., brightness, contrast, loudness, color intensity; e.g., Kamin, 1965; Öhman & Mineka, 2001), as traditionally used in the Pavlovian conditioning literature, and is closely related to the notion of stimulus intensity (e.g., Pearce & Hall, 1980). Although it has been acknoweldged that stimulus salience is not only confined to mere stimulus' charachteristics but also relates to the stimulus' importance to the organism's motivational contingencies (see, e.g., Öhman & Mineka, 2001; Rescorla, 1988a), the role of motivational salience, as opposed to physical or

important role by affecting the rate of learning; with more intense or salient CSs being generally learned more rapidly (Kamin & Schaub, 1963; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla, 1988a; Rescorla & Wagner, 1972). As salient CSs are particularly attention-grabbing, this effect may reflect the influence of attentional processes on Pavlovian conditioning, whereby CSs that are better attended to become more readily conditioned (Mackintosh, 1975; Pearce & Hall, 1980). Another phenomenon highlighting the importance of the CS salience relates to *overshadowing* (e.g., Kalat & Rozin, 1970; Mackintosh, 1976; Pavlov, 1927), which corresponds to the observation that weaker Pavlovian conditioning occurs to a CS when it is combined with a more salient CS during conditioning trials.

Number of CS-US pairings

The number of CS-US pairings is one of the most basic variables affecting Pavlovian learning. Whereas Pavlovian conditioning can occur through a single CS-US pairing in specific cases (e.g., Garcia & Koelling, 1966; Öhman, Eriksson, et al., 1975; Seligman, 1970, 1971), conditioned responding typically increases with the number of CS-US pairings, even when controlling for the overall amount of time of the conditioning procedure (Nasser & Delamater, 2016).

Temporal and spatial contiguity

The time that elapses between the CS and the US is also a crucial factor in Pavlovian conditioning. Usually, Pavlovian conditioning is the most effective when the CS precedes the US, thus acting as a signal thereof, the optimal time interval between the CS-onset and the US-onset varying across species and conditioning paradigms. Accordingly, the most efficient form of Pavlovian conditioning is *forward delay conditioning* (see Figure 2.2A), in which the CS is first presented and then coterminates with the US. Because of its high efficiency in producing Pavlovian conditioning, we notably implemented this procedure in all the empirical studies reported in the experimental part of this thesis (see chapter 3). Another form of forward conditioning is *trace conditioning* (see Figure 2.2B), where the CS begins and ends before the US is presented, the interval between the CS and the US being called the *trace interval*. Trace conditioning is in general less effective than delay conditioning and becomes less efficient the longer the trace interval. A third procedure is to present the CS and the US simultaneously in time, which is referred to as *simultaneous conditioning* (see Figure 2.2C). Simultaneous

perceptual salience, remains largely underinvestigated in Pavlovian conditioning. This distinction will be an important point for discussion later on.

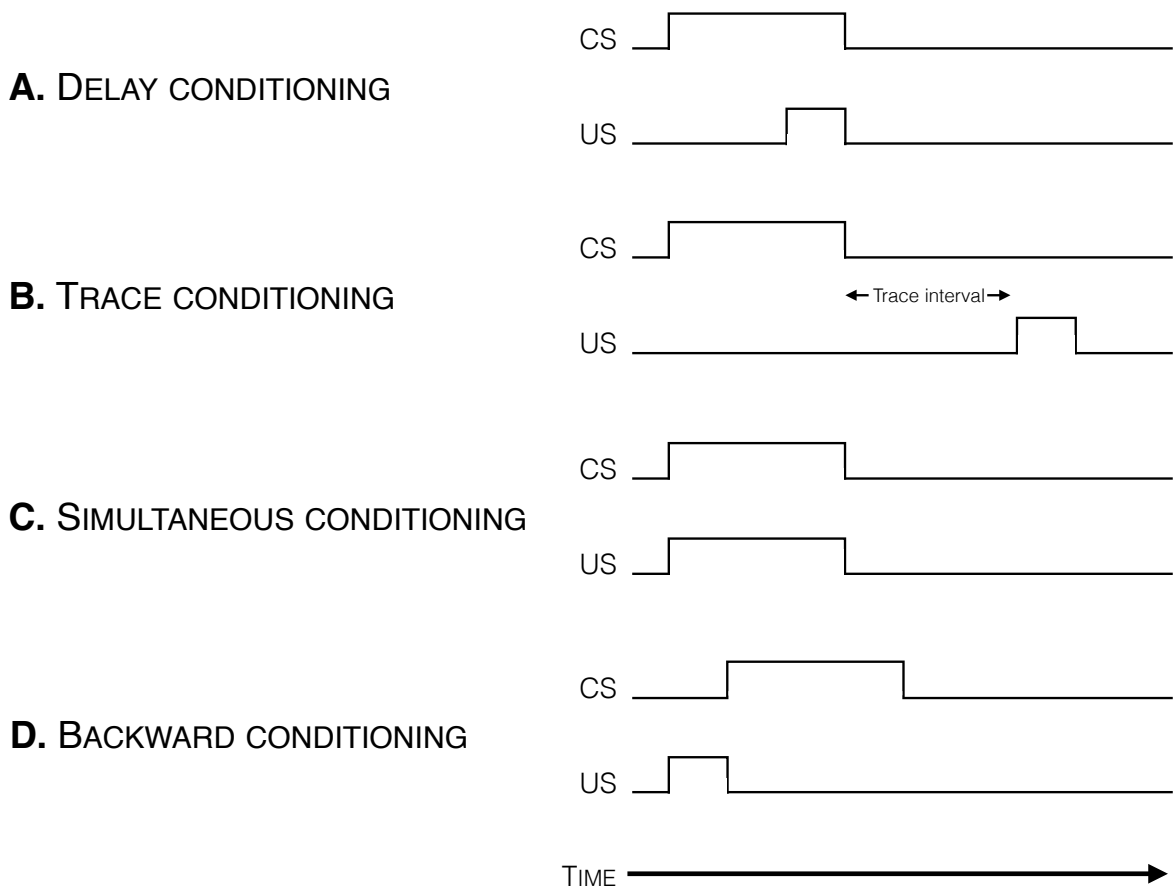

Figure 2.2. Forms of Pavlovian conditioning implementing different time intervals between the conditioned stimulus (CS) and the unconditioned stimulus (US). Adapted from Bouton (2007).

conditioning generally causes weaker Pavlovian conditioning than forward conditioning procedures. A final procedure is *backward conditioning* (see Figure 2.2D), in which the CS is presented after the US and which is usually the least effective in producing Pavlovian conditioning, though it might sometimes result in conditioned inhibition, the CS becoming a signal for the absence of the US (Bouton, 2007).

Albeit much less explored than temporal contiguity, spatial contiguity has also been shown to influence Pavlovian learning (see Nasser & Delamater, 2016). More precisely, learning can be promoted when the CS and the US are contiguous in space (Nasser & Delamater, 2016).

CS-US similarity

Although research on this specific factor is scant, CS-US similarity has been hypothesized to play an important role in the formation of an association between the CS and the US (e.g., Rescorla & Holland, 1976). In agreement with this view, Pavlovian conditioning

is generally enhanced when the CS and the US are more than less similar (Nasser & Delamater, 2016).

US surprisingness or prediction error

The extent to which the US is surprising (i.e., because of a prediction error) is a fundamental determinant of Pavlovian learning. This idea originated from Kamin's (1969) seminal work on blocking, based on which he suggested that Pavlovian conditioning only happens when the US is surprising. According to this view, blocking occurs because the CS previously associated with the US already predicts its occurrence, and the US is therefore no longer surprising (i.e., there is no prediction error) when a new CS is presented along with the CS, no learning occurring to the new CS as a result. The influence of US surprisingness has been furthermore supported by evidence showing that surprising USs are more effective reinforcers than expected USs (Rescorla & Wagner, 1972), thereby reflecting the key role of prediction error in learning (Niv & Schoenbaum, 2008; Rescorla & Wagner, 1972).

CS-US relevance or belongingness

The notion of *CS-US relevance* or *belongingness* refers to the fact that some combinations of CS and US are more readily conditioned ("belong" together) than others depending on their functional relation. Garcia and Koelling's (1966) experiment on taste aversion learning provided the most compelling support for this notion. In this experiment, thirsty rats drank water that was paired with an audiovisual stimulus and a gustatory stimulus (bright-noisy-flavored water). Consumption of the water was paired with either nausea-inducing drug or footshock as the US in different groups of rats. Rats were subsequently tested several days later either with the flavored water in the absence of the audiovisual features, or with the bright-noisy water in the absence of the flavor. Results demonstrated that rats trained with the illness US avoided drinking the flavored water but consumed the bright-noisy water, whereas rats trained with the footshock US avoided consuming the bright-noisy water but drank the flavored water. This result hence indicated that specific CS-US associations can be learned preferentially. These selective associations provided evidence for the existence of biological constraints on learning (Garcia & Koelling, 1966; Seligman, 1970), and challenged the idea that all stimuli are equally associable and follow the same general laws of learning. Of importance, selective associations have also been demonstrated in human Pavlovian aversive conditioning in response to threat-relevant stimuli (e.g., Öhman et al., 1976; Öhman & Mineka, 2001; Hamm, Vaitl, & Lang, 1989; Hugdahl & Johnsen, 1989; see chapter 2.3).

2.2.3. Pavlovian conditioning paradigms in humans

In this section, we provide an overview of the main methods used to investigate Pavlovian conditioning in humans. To this end, we briefly review the stimuli, procedures, conditioning phases, and measures that are typically employed in human Pavlovian conditioning research (see Lonsdorf et al., 2017, for a more detailed review on methodological considerations in human Pavlovian aversive conditioning studies), and specify the methodology used in the experiments reported in the experimental part of this thesis (see chapter 3).

Conditioned stimuli

In humans, Pavlovian conditioning studies predominantly use discrete exteroceptive cues as CSs (Lonsdorf et al., 2017; but see De Peuter, Van Diest, Vansteenwegen, Van den Bergh, & Vlaeyen, 2012, for a review on interoceptive aversive conditioning). Typical CSs consist of visual stimuli, such as colored squares (e.g., LaBar et al., 1998; Phelps et al., 2004; Schiller et al., 2010), geometric shapes (e.g., Gottfried, O'Doherty, & Dolan, 2003), human faces (e.g., Öhman & Dimberg, 1978; Öhman & Mineka, 2001; Olsson et al., 2005; Olsson & Phelps, 2004) or images of animals (e.g., Ho & Lipp, 2014; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005), though auditory, olfactory, gustatory, and tactile CSs have also been employed (Lonsdorf et al., 2017). In our studies (see chapter 3), we used visual CSs corresponding to images of threat-relevant stimuli (angry faces and snakes in Study 1, angry faces in Study 2), positive relevant stimuli (baby faces and erotic stimuli in Study 1, happy faces in Study 2), and neutral stimuli (neutral faces and colored squares in Study 1, neutral faces in Study 2, geometric figures in Study 3 and 4).

Unconditioned stimuli

In human Pavlovian aversive conditioning paradigms, the most commonly used and among the most efficient USs are electro-tactile stimulations, which consist of the delivery of mild electric currents to the skin, and auditory stimuli, such as loud noise or human screams (Lonsdorf et al., 2017). Interoceptive discomfort (e.g., Pappens, Smets, Vansteenwegen, Van den Bergh, & Van Diest, 2012) and unpleasant odors (e.g., Gottfried, O'Doherty, & Dolan, 2002; Hermann et al., 2000), as well as secondary reinforcers, such as loss of money (e.g., Delgado, Jou, & Phelps, 2011; Delgado, Labouliere, & Phelps, 2006), have also been employed as aversive USs. As electric stimulation constitutes a highly, if not the most, efficient aversive

US, it was used as US in all our experiments employing a Pavlovian aversive conditioning paradigm (see chapter 3).

Pavlovian appetitive conditioning studies in humans have both utilized primary reinforcers, such as food (e.g., Andreatta & Pauli, 2015), water (e.g., Kumar et al., 2008), tastes (e.g., Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013), pleasant odors (e.g., Gottfried et al., 2002, 2003), or erotic pictures (e.g., Klucken et al., 2009), and secondary reinforcers, such as money (e.g., Austin & Duka, 2010), as appetitive USs. Of note, these appetitive USs however typically (a) elicit physiological reactions that are less intense (e.g., Martin-Soelch et al., 2007), (b) are more sensitive to the organism's psychological and physiological state (e.g., the organism generally needs to be in a hunger state for food to be rewarding; Clark, Hollon, & Phillips, 2012), and (c) are more difficult to administer (see, e.g., Andreatta & Pauli, 2015) than aversive USs, such as electric stimulations, thereby making Pavlovian appetitive conditioning processes more complex than Pavlovian aversive conditioning ones to investigate in humans. In Study 4 (see chapter 3.4), we used a pleasant odor as appetitive US that was individually selected among various pleasant odors according to liking and intensity ratings in order to ensure it was rewarding for each participant and constituted an appropriate appetitive US.

Procedures

Single-cue Pavlovian conditioning procedures and differential Pavlovian conditioning procedures are typically used to study Pavlovian conditioning in animals and humans (e.g., Lonsdorf et al., 2017). In single-cue procedures, a single CS (or a compound of CSs) is associated with the US. The CRs elicited by the CS are then compared either with the responses of a control group (e.g., receiving the same number of US administrations that are not contingent upon the CS presentation or explicitly unpaired with the CS) or with the responses in the absence of the CS (e.g., during the intertrial interval), thus involving exclusively between-subject comparisons (e.g., Lonsdorf et al., 2017; Rescorla, 1967). While these procedures are the most used in rodents, they are not very common in humans (Lonsdorf et al., 2017). Human research on Pavlovian conditioning indeed mostly employs differential procedures (Lonsdorf et al., 2017). In a differential Pavlovian conditioning procedure (see Figure 2.3), two types of CSs are used: One stimulus (the reinforced stimulus, CS+) is contingently paired with the US, whereas another stimulus (the unreinforced stimulus, CS-) is never associated with the US. The CRs are operationalized as the differential response to the CS+ versus the CS- (e.g., Lonsdorf et al., 2017; Olsson et al., 2005). Differential Pavlovian

conditioning procedures thereby allows for controlling for (a) between-subject differences in responding, (b) nonassociative processes (e.g., orienting responses, habituation, sensitization) that are thought to affect the responses to the CS+ and the CS- in a similar manner, and (c) the possible confounding role of preexisting differences in the CS categories' emotional salience when several categories of CS are used and compared with each other in the same experiment (e.g., Lonsdorf et al., 2017; Olsson et al., 2005). In a similar vein, differential Pavlovian conditioning procedures provide more statistical power than single-cue procedures (Lonsdorf et al., 2017). It is however important to note that the CS- may not constitute a completely neutral control stimulus as it signals the absence of the US, hence possibly inducing inhibitory learning (Lonsdorf et al., 2017). Because of its many advantages over single-cue procedures, a differential Pavlovian conditioning paradigm was used in each study of the experimental part of this thesis (see chapter 3).

Conditioning phases

Differential Pavlovian conditioning paradigms in humans generally include three successive and distinctive phases: habituation, acquisition, and extinction (see Figure 2.3). During the habituation phase, all the CSs are presented repeatedly without being associated with the US. This phase notably serves to (a) establish a baseline response rate, allowing to determine and correct for potential pre-existing differences in responding to the to-be-CS+ and the to-be-CS-, and (b) to reduce the initial reactivity of certain physiological measures that show a decline in responding over the first number of trials (e.g., orienting responses; Lonsdorf et al., 2017). The habituation phase may be preceded by a US calibration or selection procedure. US calibration procedures are commonly implemented in Pavlovian aversive conditioning paradigms, with the aim of individually adjusting the US intensity (mostly for electric stimulations) to a pre-defined subjective level of unpleasantness or aversiveness across participants (Lonsdorf et al., 2017). US selection procedures are also sometimes used in Pavlovian appetitive conditioning studies (e.g., Stussi, Delplanque, et al., 2018) to individually select the most liked stimulus as appetitive US, thus optimizing its chances to have rewarding properties for each and every participant. In the acquisition phase, the CS+ is contingently paired with the US, while the CS- is never associated with it. In humans, delay conditioning and, to a lesser extent, trace conditioning procedures are predominantly used given their greater efficiency to produce Pavlovian conditioning. Finally, during the extinction phase, all the CSs are presented in the absence of the US, the delivery of which is discontinued. These three phases (habituation, acquisition, and extinction) were included in all the experiments that we

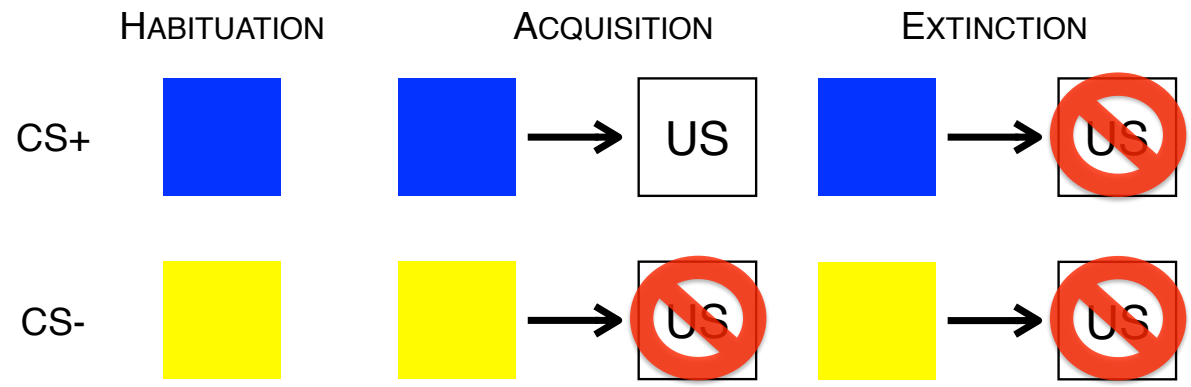

Figure 2.3. Illustration of a differential Pavlovian conditioning procedure. During the habituation phase, the conditioned stimuli (CSs; e.g., colored squares) are presented without being reinforced. In the acquisition phase, the reinforced CS (CS+) is associated with the unconditioned stimulus (US), whereas the unreinforced CS (CS-) is never paired with the US. In the extinction phase, the US is no longer delivered.

performed (see chapter 3). Studies investigating the phenomena of the CR recovery after extinction may contain additional phases after extinction, such as reextinction, reinstatement, or renewal for instance (e.g., Schiller et al., 2010; see Haaker, Golkar, Hermans, & Lonsdorf, 2014, for a review on human reinstatement studies).

Conditioned response measures

Table 2.1 summarizes the main dependent variables used in the existing literature to measure the CR in humans (see also Lonsdorf et al., 2017, for a review of the outcome measures used in human Pavlovian aversive conditioning research). Psychophysiological indicators are commonly applied to index Pavlovian conditioning (see, e.g., Leuchs, Schneider, & Spoormaker, 2019). Among those, the most widely employed index of the CR is electrodermal activity (EDA), and more specifically skin conductance response (SCR). EDA is primarily under sympathetic control and originates from the eccrine sweat glands, which are most dense on the palms and soles of the feet (Dawson, Schell, & Fillion, 2016). Whereas eccrine sweat glands primary function is thermoregulation, those located on the palms have been suggested to be more responsive to emotional than thermal stimuli (Dawson et al., 2016), thereby reflecting EDA's sensitivity to emotional processes. SCR is considered a valid indicator of autonomic arousal and corresponds to a phasic response to a stimulus that triggers an increase in skin conductance (e.g., Lykken & Venables, 1971). Accordingly, the CS+ typically elicits SCRs of larger amplitude than the CS- during Pavlovian conditioning

(Lonsdorf et al., 2017). Whereas SCR has been extensively used as a reliable and sensitive psychophysiological indicator of Pavlovian aversive conditioning processes in humans, Pavlovian appetitive conditioning studies using SCR as an indicator of appetitive CRs have yielded rather mixed results, some studies reporting enhanced SCRs to the CS+ associated with the appetitive US compared with the CS- (e.g., Andreatta & Pauli, 2015; Klucken et al., 2009), while others observing no difference in SCR to the CS+ versus the CS- (e.g., Hermann et al., 2000; Stussi, Delplanque, et al., 2018). It has been argued that this asymmetry might arise from the fact that SCR may be particularly sensitive to the US intensity, hence possibly failing to detect subtle changes caused by Pavlovian appetitive conditioning due to a weaker US intensity than in Pavlovian aversive conditioning (Stussi, Delplanque, et al., 2018). As SCR is one of the most well-established psychophysiological measures of Pavlovian conditioning in humans, it was used as the main dependent variable of the CR in all the experiments that we conducted (see chapter 3).

Another widely applied psychophysiological measure of the CR in human Pavlovian conditioning research is the startle reflex, which is an automatic defensive response to a sudden, intense, and unexpected sensory event. The startle reflex is elicited with a startle probe typically consisting of a brief white noise burst (acoustic startle probe). In humans, the startle reflex is frequently indexed with the eyeblink reflex. The eyeblink reflex has been shown to be potentiated in response to a CS+ paired with an aversive US (e.g., Andreatta & Pauli, 2015; Grillon, 2002; Grillon & Davis, 1997; Hamm, Greenwald, Bradley, & Lang, 1993), and attenuated in response to a CS+ paired with an appetitive US (Andreatta & Pauli, 2015; but see Bradley, Zlatar, & Lang, 2018; Stussi, Delplanque, et al., 2018), relative to a CS-. We used startle eyeblink reflex as a measure of appetitive CRs in Study 4 (see chapter 3.4) in order to assess its sensitivity in measuring Pavlovian appetitive learning in comparison with the postauricular reflex. Other common but less frequently used psychophysiological measures of Pavlovian conditioning include heart rate (e.g., Hamm et al., 1993), finger-pulse volume responses (e.g., Hamm et al., 1989), and pupillary response (e.g., Korn, Staib, Tzovara, Categnetti, & Bach, 2017; Leuchs et al., 2019; O'Doherty et al., 2003; Reinhard & Lachnit, 2002).

At the neural level, electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) have been used to measure Pavlovian CRs. Research using EEG (e.g., Bacigalupo & Luck, 2018; Stolarova, Keil, & Moratti, 2006; Wieser, Miskovic, Rausch, & Keil, 2014) and MEG (e.g., Yuan, Giménez-Fernández, Méndez-

Bértolo, & Moratti, 2018) has indicated that Pavlovian learning induces amplification of cortical sensory and attentional processing of the CS+ versus the CS- mainly in aversive (for a review, see Miskovic & Keil, 2012) but also in appetitive paradigms (e.g., Franken, Huijding, Nijs, & van Strien, 2011). Studies using fMRI have shown that the acquisition, expression, and extinction of CRs during Pavlovian aversive conditioning involve a distributed neural network, including in particular the amygdala (e.g., Büchel et al., 1998; LaBar et al., 1998; Phelps et al., 2004; Sehlmeyer et al., 2009), the anterior cingulate cortex, the insula, the hippocampus, the striatum, the ventromedial prefrontal cortex, the thalamus, and motor and sensory cortices (see Fullana et al., 2016; Sehlmeyer et al., 2009, for a recent meta-analysis and a systematic review, respectively). In comparison, effects of Pavlovian appetitive conditioning have been observed across a distributed brain network, including in particular the amygdala, the striatum, the orbitofrontal cortex, and the anterior cingulate cortex (e.g., Gottfried et al., 2002, 2003; Martin-Soelch et al., 2007; O'Doherty et al., 2003).

Behavioral measures are also sometimes used as indices of Pavlovian CRs. They essentially consist of reaction time measures for detecting the CSs (e.g., Gottfried et al., 2003; Pool, Brosch, Delplanque, & Sander, 2015), generally resulting in faster reaction times in response to the CS+ than to the CS-. Reaction times have also been used to index avoidance and approach tendencies (e.g., Krypotos, Effting, Arnaudova, Kindt, & Beckers, 2014; Van Gucht, Vansteenwegen, Van den Bergh, & Beckers, 2008), with faster reaction times to avoid the CS+ compared with the CS- in Pavlovian aversive conditioning and faster reaction times to approach the CS+ relative to the CS- in Pavlovian appetitive conditioning. Reaction times can be measured trial-by-trial during the Pavlovian conditioning procedure (e.g., Pool et al., 2015), intermittently after each conditioning phase or group of trials, or even retrospectively after the entire conditioning procedure (e.g., Van Gucht et al., 2008).

Subjective ratings or verbal reports provide another complementing measure of Pavlovian conditioning at the behavioral level. Subjective ratings include for instance CS-US contingency or US expectancy ratings (e.g., Boddez et al., 2013), CS pleasantness (or valence or liking) ratings, CS arousal ratings, and subjective feeling ratings, such as fear ratings in Pavlovian aversive conditioning (e.g., Lonsdorf et al., 2017). CS-US contingency ratings assess the extent to which the CS+ and the CS- are deemed predictive of the US, and provide an indicator of whether participants are aware of the contingencies between the CSs and the US (but see Boddez et al., 2013; Lovibond & Shanks, 2002, for discussions on methodological considerations for these ratings). Ratings of the CS pleasantness evaluate to what extent the

Table 2.1

Overview of the main measures of the conditioned response commonly used in human Pavlovian conditioning paradigms.

Level	Measure	Paradigm	Measurement time	Typical effect of Pavlovian conditioning
Psychophysiological	Skin conductance response	Aversive and appetitive	Continuous	CS+ > CS-
	Startle eyeblink reflex	Aversive and appetitive	Continuous	*Aversive*: CS+ > CS-; *Appetitive*: CS+ < CS-
	Heart rate	Aversive	Continuous	CS+ < CS- (orienting response) or CS+ > CS- (defensive response)
	Finger-pulse volume	Aversive	Continuous	CS+ < CS-
	Pupillary response	Aversive and appetitive	Continuous	CS+ > CS-
Neural	EEG/MEG	Aversive and appetitive	Continuous	ERPs/ERFs/ssVEPs/ssVEFs: CS+ > CS- in sensory brain regions
	fMRI	Aversive and appetitive	Continuous	*Aversive*: BOLD signal CS+ > CS- in a brain network including the amygdala, the ACC, the insula, the hippocampus *Appetitive*: BOLD signal CS+ > CS- in a brain network including the amygdala, the striatum, the ACC, the OFC
Behavioral	Reaction times	Aversive and appetitive	Continuous, intermittent, or retrospective	CS+ < CS-
Subjective	CS-US contingency/ US expectancy	Aversive and appetitive	Continuous, intermittent, or retrospective	CS+ > CS-
	CS pleasantness	Aversive and appetitive	Continuous, intermittent, or retrospective	*Aversive*: CS+ < CS-; *Appetitive*: CS+ > CS-
	CS arousal	Aversive and appetitive	Continuous, intermittent, or retrospective	CS+ > CS-
	Fear ratings	Aversive	Continuous, intermittent, or retrospective	CS+ > CS-

Note. ERPs = event-related potentials, ERFs = event-related field time averaged responses, ssVEPs = steady-state visual evoked potentials, ssVEFs = steady-state visual evoked fields, BOLD = blood-oxygen-level dependent, ACC = anterior cingulate cortex, OFC = orbitofrontal cortex.

CSs are pleasant or unpleasant, thereby reflecting the evaluative effects of Pavlovian conditioning (see, e.g., De Houwer, Thomas, & Baeyens, 2001; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). CS arousal ratings assess the degree to which the CSs are deemed subjectively arousing. Fear ratings are employed to evaluate how strong participants' subjective feelings of fear are in response to the CSs (Lonsdorf et al., 2017). All of these subjective rating measures can be assessed online trial-by-trial, in an intermittent fashion after each conditioning phase or group of trials, or in retrospect after the end of the conditioning procedure. We measured CS-US contingency ratings and CS liking or pleasantness ratings after the conditioning procedure in all the empirical studies reported in the experimental part (see chapter 3), whereas we additionally measured CS arousal ratings and CS subjective relevance ratings in Studies 2, 3, and 4, and in Studies 2 and 3, respectively. These subjective ratings primarily served as manipulation checks with the aim of assessing participants' awareness of the reinforcement contingencies, as well as the evaluative effects of the Pavlovian conditioning procedure.

2.2.4. Formal models of Pavlovian conditioning

After a review of the main paradigms employed to study Pavlovian conditioning in humans, we introduce the major theories of Pavlovian conditioning (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972) that have been proposed to account for the wealth of empirical findings reported in the literature, as well as some of their recent development (Li et al., 2011). These theories are particularly relevant for explaining and characterizing the fundamental computational mechanisms underlying Pavlovian conditioning. To gain insights into the computations underlying the influence of affectively relevant stimuli on Pavlovian conditioning relative to neutral stimuli with less relevance, we accordingly used and adapted some of these models in Study 1 and Study 2 (see chapter 3.1 and 3.2, respectively).

Rescorla-Wagner model

The most influential model of Pavlovian conditioning is the Rescorla-Wagner model (Rescorla & Wagner, 1972). Inspired by the earlier Bush and Mosteller's (1951) model, the Rescorla-Wagner model is built on the idea that learning depends on the effectiveness or surprisingness of the US (see Kamin, 1968, 1969, for a similar view), thus being often referred to as a US-processing model (e.g., Le Pelley, 2004). It postulates that organisms only learn when events violate their expectations. This model thereby formalizes the role of prediction

error in learning by providing a simple computational and algorithmic account of Pavlovian conditioning according to which learning is directly driven by the discrepancy between the CS current associative strength or predictive value (i.e., the extent to which the CS predicts the US on that trial) and the maximum amount (or asymptote) of learning determined by the magnitude of the US. More formally, the Rescorla-Wagner model states that the predictive value V at trial $t + 1$ of a given CS j is updated based on the sum of the current predictive value V_j at trial t, and the prediction error between the summed predictive value ΣV of all CSs presented at trial t and the maximum associative strength λ that the US can support, weighted by two constants α and β:

$$V_j(t+1) = V_j(t) + \alpha_j \cdot \beta \cdot (\lambda - \sum V(t))$$

where α_j is a learning-rate parameter within the range [0, 1] that is determined by the CS intensity or salience, and β is a parameter within the range [0, 1] that relates to the US intensity. Whereas the summation term ΣV is critical in the Rescorla-Wagner model as it allows for explaining conditioning phenomena involving compound stimuli, such as blocking (see Rescorla, 1988b; see chapter 2.2.1, section "CS-US contingency" here above), this model can be simplified in many cases when no compound stimuli are used by considering only the predictive value V of the CS presented at trial t. Similarly, the constant β parameter is often omitted in the event only a single type of US is used, thus limiting the number of parameters included in the model. The parameter λ is also commonly replaced by the value R corresponding to the reinforcement at trial t (e.g., $R(t) = 1$ if the US is delivered at trial t, whereas $R(t) = 0$ if the US is not presented). This model further assumes that the CS predictive value is monotonically related to the observed CR.

Altogether, the Rescorla-Wagner model successfully accounts for a wide range of behavioral phenomena observed during Pavlovian conditioning (see Miller, Barnet, & Grahame, 1995, for a review), such as the effects of the US magnitude or intensity (through different values of λ and β, respectively; see Figure 2.4A) and the CS salience (through different values of α; see Figure 2.4B), compound conditioning phenomena (e.g., blocking, overshadowing, and conditioned inhibition), the role the CS-US contingency (i.e., the contingency between the CS and the US must be stronger than the contingency between no-CS or the context and the US for Pavlovian conditioning to occur; see Rescorla, 1988b; Rescorla & Wagner, 1972), and the occurrence of extinction when the US is no longer presented (i.e., $\lambda = 0$; see Figure 2.4C). In spite of these strengths, this model however cannot

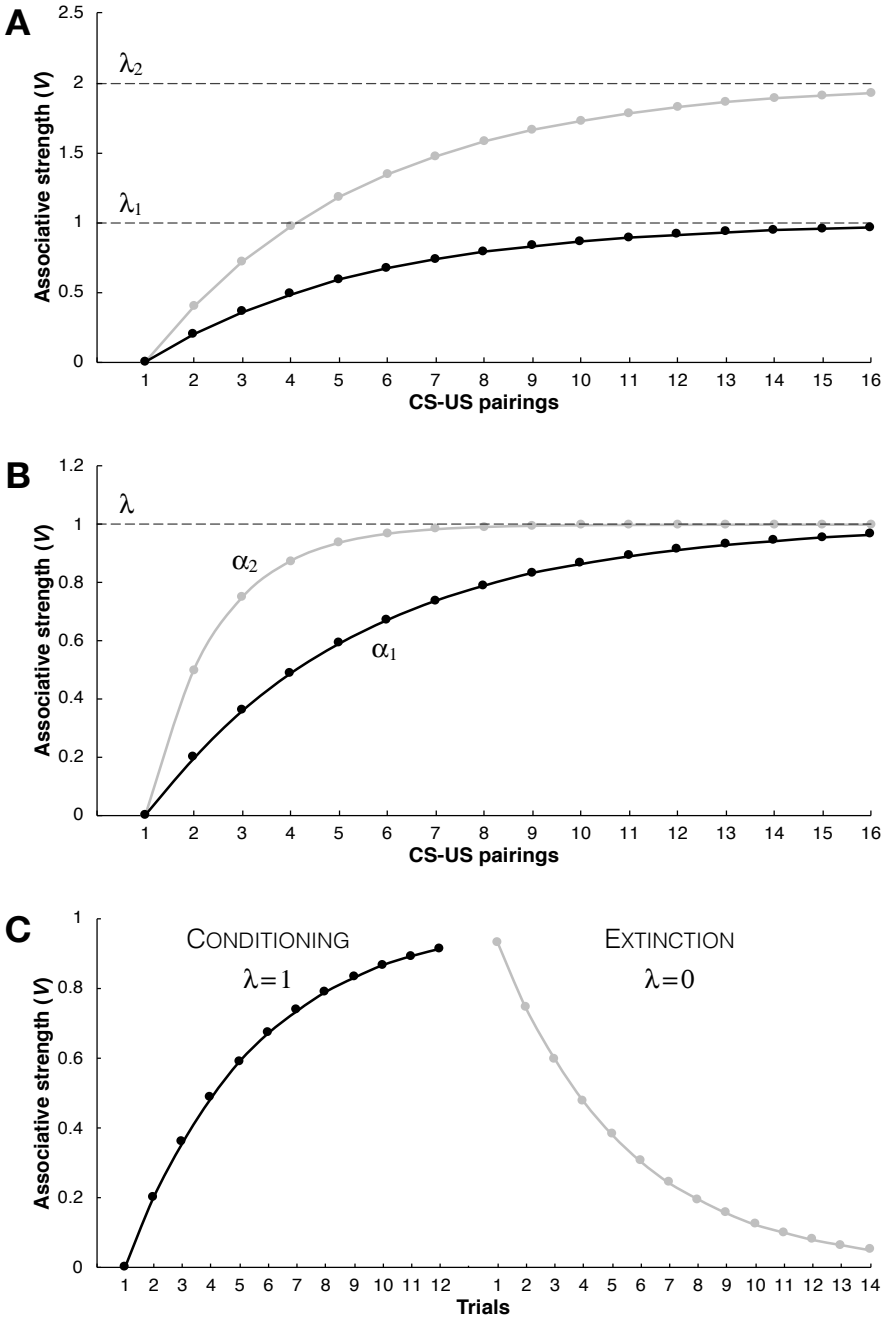

Figure 2.4. Illustrations of learning curves according to the Rescorla-Wagner model. (A) Effect of the unconditioned stimulus magnitude (λ) on learning. (B) Effect of conditioned stimulus salience (α) on learning. (C) Conditioning (left) and extinction (right) curves. In extinction, the Rescorla-Wagner model predicts that the associative strength approaches a new asymptote as the conditioned stimulus predicts no unconditioned stimulus (λ = 0). Adapted from Bouton (2007).

easily accommodate some important findings in the Pavlovian conditioning literature, such as latent inhibition, the extinction of inhibition, and the role of the CS-US relevance (see Miller et al., 1995, for a review). As the Rescorla-Wagner model conceives extinction as unlearning of the CS-US association, it additionally cannot satisfactorily explain the phenomena of the

CR recovery from extinction, such as spontaneous recovery and rapid reacquisition. Another limitation pertains to the fact that the Rescorla-Wagner model discretizes predictions in terms of trials, whereas predictions are likely to be continuous in time. Such shortcoming is notably addressed by temporal difference learning models (e.g., Sutton & Barto, 1998), which represent an extension of the Rescorla-Wagner model to the temporal domain. These limitations notwithstanding, the Rescorla-Wagner model provides an elegant and simple account of Pavlovian conditioning that still holds a high heuristic value nowadays.

Mackintosh model

At variance with the Rescorla-Wagner model, Mackintosh (1975) proposed a CS-processing model of Pavlovian conditioning, according to which learning is not only driven by the effectiveness of the US, but also by the associability of the CS. The basic idea of the Mackintosh model is that the CS associability is determined by how much attention is devoted to the CS, with CSs that are better attended to having in turn higher associability. More specifically, Mackintosh suggested that the amount of attention paid to the CS (a) depends on the extent to which the CS reliably predicts the US and (b) changes over time as a function of experience. If the CS is experienced as a reliable predictor of the US, then more attention is allocated to it and it is hence more easily conditioned; by contrast, attention to the CS declines if it does not predict the US more reliably than other CSs. The Mackintosh model incorporates this notion of variable associability by enabling this variable characterizing the CS to change as a result of its predictiveness of the US. In this model, the associative strength V at trial $t + 1$ of a given CS j is changed based on the sum of the current CS associative strength V_j at trial t, and the prediction error between the current CS associative strength V_j at trial t and the maximum associative strength λ supported by the US, weighted by a constant θ and the CS associability α at trial *t*:

$$V_j(t+1) = V_j(t) + \theta \cdot \alpha_j(t) \cdot (\lambda - V_j(t))$$

where θ is a learning-rate parameter within the range [0, 1] determined by the US intensity. As for CS associability, it is updated as follows (see Le Pelley, 2004):

$$\Delta\alpha_j(t) > 0 \; if \left|\lambda(t) - V_j(t-1)\right| < \left|\lambda(t) - V_X(t-1)\right|$$
$$\Delta\alpha_j(t) < 0 \; if \left|\lambda(t) - V_j(t-1)\right| \geq \left|\lambda(t) - V_X(t-1)\right|$$

where $\Delta\alpha_j(t)$ is the change of associability of the CS_j at trial t, $\lambda(t)$ is the US magnitude at trial t, $V_j(t-1)$ is the associative strength of the CS_j at trial $t-1$ (i.e., before it was updated at trial t), and $V_X(t-1)$ is the associative strength of all the other CSs present at trial $t-1$ before associative strengths were updated at trial t. Mackintosh suggested that the change in α_j should be proportional to the magnitude of these two inequalities, without however providing an algorithm for computing this change (Le Pelley, 2004).

According to the Mackintosh model, both prediction error and associability govern learning. By formalizing the role of attention in Pavlovian conditioning, this model addresses some of the Rescorla-Wagner model's shortcomings. It notably allows for explaining the phenomenon of latent inhibition by suggesting that when the CS is preexposed without being associated with the US, the CS associability decreases, thereby having only weak increments in associative strength at the outset of conditioning. Furthermore, the Mackintosh model provides an alternative explanation to the blocking effect: Because the new CS is a redundant predictor of the US, organisms learn to pay less attention to the new CS and ultimately ignore it, which results in no learning occurring to that CS after the first blocking trial (but see Balaz, Kasprow, & Miller, 1982, for a study reporting complete blocking on the first trial in line with the prediction of the Rescorla-Wagner model). Mackintosh's (1975) attentional model of Pavlovian conditioning has received support from research on the interactions between associative learning and attention in humans, which has shown that attention tends to be biased toward stimuli that predict their consequences in a reliable manner (see Le Pelley, Mitchell, Beesley, George, & Willis, 2016, for a review). Nevertheless, this model also presents some limitations, being notably unable to capture the role of the CS-US relevance. In addition, it has been demonstrated that in some situations, the associability of a CS actually decreases rather than increases when the CS becomes a reliable predictor of the US (see, e.g., Le Pelley, 2004; Pearce & Hall, 1980).

Pearce-Hall model

Pearce and Hall (1980; Pearce, Kaye, & Hall, 1982) proposed an alternative CS-processing model of Pavlovian conditioning, in which the associability of the CS constitutes the key determinant of learning. Unlike Mackintosh's (1975) rule for how attention to the CS changes during Pavlovian conditioning, the Pearce-Hall model posits that the CS associability declines when the CS accurately predicts the US, whereas it increases or remains high when the CS is an unreliable predictor of the US. This proposal reflects the notion that organisms

need to attend to stimuli the consequences of which are uncertain, rather than spending their attentional resources on stimuli that already predict their consequences in a reliable fashion (Bouton, 2007; Pearce & Hall, 1980). The Pearce-Hall model accordingly states that the associability α at trial $t+1$ of a given CS j is determined by the sum of the current associability α_j and the absolute value of the prediction error between the actual reinforcer λ and the difference between the current sums of the associative strengths of all the CSs for excitatory learning ΣV and for inhibitory learning $\Sigma \bar{V}$ at trial t, as follows (see Le Pelley, 2004):

$$\alpha_j(t+1) = \eta \cdot \left| \lambda(t) - (\sum V(t) - \sum \bar{V}(t)) \right| + (1-\eta) \cdot \alpha_j(t)$$

where η is a weighting parameter within the range [0, 1] that determines the extent to which changes in α_j are determined by the events of the current and preceding trials. For instance, if $\eta = 1$, changes in α_j are determined exclusively by the events of the current trial, with no influence of preceding trials; by contrast, if $\eta = 0$, changes in α_j are determined by earlier trials, with the current trial having no effect (Le Pelley, 2004).

As can be seen above, the Pearce-Hall model additionally assumes the existence of separated associative strengths for excitatory and inhibitory learning. Indeed, this model conceptualizes extinction of the excitatory CR as being caused by an inhibitory relationship between the CS-US association and the CS-no US association (Dunsmoor, Niv, et al., 2015; Konorski, 1967; Pearce & Hall, 1980). This assumption contrasts with the Rescorla-Wagner and Mackintosh models, which both consider inhibition as a symmetrical opposite of excitation, excitation and inhibition both affecting a single CS associative strength. According to the Pearce-Hall model, whether excitatory or inhibitory learning occurs depends on the prediction error δ between the actual reinforcer at trial t and the difference between the sums of the associative strengths of all the CSs for excitatory learning ΣV and for inhibitory learning $\Sigma \bar{V}$ at trial $t-1$, which is computed as follows (see Le Pelley, 2004):

$$\delta(t) = \lambda(t) - (\sum V(t-1) - \sum \bar{V}(t-1))$$

If δ is positive (i.e., positive prediction error), the excitatory associative strength V_j at trial $t+1$ is updated based on the sum of the current excitatory associative strength V_j at trial t and the product of the constant S_j, the CS associability α_j at trial t, and the excitatory reinforcer λ at trial t, as follows (see Le Pelley, 2004):

$$V_j(t+1) = V_j(t) + S_j \cdot \alpha_j(t) \cdot \lambda(t)$$

where S_j is a learning-rate parameter within the range [0, 1] that is determined by the CS intensity or intrinsic salience.

Inversely, if δ is negative (i.e., negative prediction error), the inhibitory associative strength $\bar{V}_j$ at trial $t+1$ is updated based on the sum of the current inhibitory associative strength $\bar{V}_j$ at trial t and the product of the constant learning rate S_j, the CS associability α_j at trial t, and the absolute value of the prediction error δ at trial t (see Le Pelley, 2004):

$$\bar{V}_j(t+1) = \bar{V}_j(t) + S_j \cdot \alpha_j(t) \cdot |\delta(t)|$$

In sum, the Pearce-Hall model assigns an even more important role to associability processes in learning than the Mackintosh model, in that modulation of excitatory learning is directly driven by the processing of the CS, prediction error only having an indirect influence thereon through the alteration of the CS associability. Accordingly, this model can account for a number of attentional phenomena seen during Pavlovian conditioning, while accommodating many of the conditioning phenomena addressed by the Rescorla-Wagner model (Pearce & Hall, 1980). It furthermore aligns with the observation that Pavlovian conditioning to an uncertain or inaccurate predictor of the US is faster than to an accurate predictor of the US under certain conditions (Le Pelley, 2004; Pearce & Hall, 1980). As for the Rescorla-Wagner and the Mackintosh models, the Pearce-Hall model however also has several limitations, which likewise include the difficulty in explaining the impact of the CS-US relevance. In addition, the idea that CSs associated with high uncertainty are allocated more attention has received to date only little empirical support in human attentional learning research (Le Pelley et al., 2016).

Hybrid model

With the aim of capitalizing on the different model's strengths and providing an even more satisfactory account of Pavlovian conditioning, attempts have been made to propose hybrid models that combine several features of the aforementioned models (e.g., Le Pelley, 2004). For example, various hybrid models have sought to reconcile the role of predictiveness and uncertainty in the modulation of stimulus associability by incorporating both the Mackintosh and the Pearce-Hall associability mechanisms within the same model (e.g., Esber & Haselgrove, 2011; Le Pelley, 2004). Another hybrid model (which we hereafter refer to as the hybrid model) introduced by Li and colleagues (Li et al., 2011) combines both the simplified version of Rescorla-Wagner model and the Pearce-Hall model, where the Rescorla-

Wagner algorithm is implemented for updating the CS predictive value based on prediction error, and the Pearce-Hall associability mechanism is substituted for the constant learning rate α, thereby acting as a dynamic learning rate. Following the Pearce-Hall rule, the CS associability decreases when the CS correctly and reliably predicts the US, whereas it increases when the CS does not reliably predict the US. In this hybrid model, the predictive value V and the associability α of a given CS j are updated as follows:

$$V_j(t+1) = V_j(t) + \kappa \cdot \alpha_j(t) \cdot (R(t) - V_j(t))$$

$$\alpha_j(t+1) = \eta \cdot \left| R(t) - V_j(t) \right| + (1-\eta) \cdot \alpha_j(t)$$

where the initial associability α_0, the constant learning rate κ, and the weighting factor η are free parameters within the range [0, 1].

Li et al.'s (2011) hybrid model has been put forward as a more accurate explanation of both SCR (e.g., Li et al., 2011) and US expectancy rating (e.g., Boll et al., 2013) modulations during aversive reversal-learning tasks (i.e., in which the contingencies between the CSs and the US are changed during the task) in humans than the Rescorla-Wagner model. At the neural level, it has furthermore allowed highlighting that the neural correlates of the Rescorla-Wagner error-driven algorithm and the Pearce-Hall associability mechanism coexist within the human brain (Roesch, Esber, Li, Daw, & Schoenbaum, 2012), with prediction error signals being notably encoded in the striatum (Li et al., 2011) and the central nucleus of the amygdala (Boll et al., 2013), and associability signals encoded in the amygdala (Li et al., 2011), and more specifically in the basolateral nuclei (Boll et al., 2013). Even though the Rescorla-Wagner and the Pearce-Hall models are generally viewed as competing models, the hybrid model conversely suggests that prediction error and associability mechanisms may be largely complementary and interacting with each other to drive Pavlovian learning (e.g., Le Pelley, 2004; Li et al., 2011; Roesch et al., 2012).

2.3. PREFERENTIAL EMOTIONAL LEARNING

As illustrated so far, research on Pavlovian conditioning has mostly focused on uncovering the general principles of learning that apply across different types of stimuli, without considering the putative importance or emotional value of the stimuli at stake for the organism. Departing from this trend, studies on taste aversion learning (e.g., Garcia & Koelling, 1966) have however revealed that certain associations between stimuli are more easily formed and maintained than others. These studies have demonstrated that not all stimuli can be associated with equal ease in Pavlovian conditioning, thereby challenging the general process view of learning held by early learning theorists (e.g., Estes, 1950; Pavlov, 1927; Watson & Rayner, 1920). These results contributed to the emergence of the idea that learning mechanisms are biologically constrained and might have evolved to help organisms handle specific problems that they encounter in their environment (Bolles, 1970; Bouton, 2007; Seligman, 1970). In this vein, research on preferential aversive learning in humans (e.g., E. W. Cook, Hodes, & Lang, 1986; Ho & Lipp, 2014; Öhman & Dimberg, 1978; Öhman, Eriksson, & Olofsson, 1975; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005) and non-human primates (e.g., M. Cook & Mineka, 1989, 1990; Mineka, Davidson, Cook, & Keir, 1984) has demonstrated that stimuli from certain evolutionarily threat-relevant categories, such as snakes and threatening conspecifics, are more readily and persistently associated with aversive outcomes than stimuli from threat-irrelevant categories, such as flowers and nonthreatening conspecifics. These findings emphasize that specific classes of stimuli can produce enhanced Pavlovian conditioning, thus reflecting the existence of learning biases or preferential emotional learning. In general, such preferential emotional learning is characterized by a faster acquisition of a CR, the acquisition of a larger CR, and/or enhanced resistance to extinction of that CR, all of these different indicators being considered as inherently valid (Öhman & Mineka, 2001; Rescorla, 1980).

Nonetheless, psychological mechanisms underlying preferential emotional learning remain surprisingly poorly understood currently. Critically, formal models of Pavlovian conditioning, such as the Rescorla-Wagner, the Mackintosh, and the Pearce-Hall models, or even the hybrid model (see chapter 2.2.4 for their presentation and underlying computational principles), do not accommodate the combined findings of faster and more persistent Pavlovian aversive learning to threat-relevant relative to threat-irrelevant stimuli, as well as the role of the CS-US relevance. Whereas these models account for the effects of accelerated Pavlovian aversive conditioning to threat-relevant stimuli because of their higher salience than that of

threat-irrelevant stimuli, they also predict that, all else being equal, the CR to more salient stimuli should extinguish faster than the CR to less salient stimuli (see Siddle & Bond, 1988; see also Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002, for studies in rats providing either direct or indirect support for this prediction), hence contrasting with the greater CR persistence typically observed to threat-relevant stimuli.

In this section, we provide an overview of the main theoretical models put forward to account for these preferential associations. We first present the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, which both adopt an evolutionary perspective according to which only stimuli that have threatened survival across evolution are preferentially conditioned to threat, and review empirical evidence both supporting and contradicting these theories' core predictions. Finally, we propose an alternative framework deriving from appraisal theories of emotion (e.g., Sander et al., 2003, 2005, 2018), which holds that preferential emotional learning in humans is driven by a more general mechanism of relevance detection as opposed to a threat-specific mechanism.

2.3.1. Preparedness theory

Basic assumptions

The preparedness theory was originally introduced by Seligman (1970, 1971). Challenging the equipotentiality premise that was widely held by general process learning theory at the time (e.g., Estes, 1950; Pavlov, 1927; Watson & Rayner, 1920), this theory offered a new perspective on Pavlovian conditioning by stating that all stimuli are not equivalently associable, and contributed to highlighting the pivotal role of evolution in associative learning. Seligman argued that evolutionary contingencies have prepared organisms to preferentially learn specific associations between events that have facilitated the survival of the species. Specifically, Seligman suggested that associative learning is organized along a continuum of preparedness holding that organisms are either prepared, unprepared, or contraprepared to learn to associate certain events. The dimension of preparedness is defined in terms of ease of learning or amount of input (e.g., number of trials or pairings) that is required for an association to be learned. Prepared learning would require only a single or a few pairings to occur even if the input is highly degraded, whereas unprepared learning would necessitate many pairings to emerge. By contrast, if learning occurs only after an extensive amount of pairings or does not occur at all, then the organism would be deemed contraprepared for learning the association in

question. Taste aversion learning would for instance represent a case of prepared and contraprepared learning: Rats are prepared to associate tastes or flavors with illness, whereas they are contraprepared to associate tastes with footshock (Seligman, 1970). Associations between arbitrary events, such as the association between a colored square or a tone and an electric stimulation, would inversely constitute instances of unprepared learning.

Seligman (1971) applied the concept of preparedness to explain the acquisition of fears and phobias in humans, suggesting that phobias constitute a case of highly prepared learning. While adopting the view that phobias or intense fears may result from Pavlovian conditioning (e.g., Watson & Rayner, 1920), Seligman noted that phobias differ from fears conditioned to arbitrary stimuli through Pavlovian aversive conditioning in the laboratory (i.e., unprepared learning) in four major ways. First, phobias are selective. In fact, phobias are not evenly distributed across all possible stimuli. Instead, humans are more likely to develop fears or phobias of specific stimuli that have provided threats to the species' survival through evolution (e.g., dangerous animals, heights, storms; Agras, Sylvester, & Oliveau, 1969; Fredrikson, Annas, Fischer, & Wik, 1996; Marks, 1969). Second, phobias, unlike fears conditioned to arbitrary stimuli, are highly resistant to extinction. Repeated exposure to the phobic stimulus in the absence of the traumatic event is in general insufficient to induce typical extinction as observed in laboratory Pavlovian conditioning (Mallan et al., 2013). Third, phobias are easily acquired and can result from a single traumatic event (i.e., one-trial learning). Fourth, phobias seem to be irrational and impervious to cognitive influences, such as verbal instructions, in that they are resistant to change by information. Informing a phobic person that the object that corresponds to the source of their phobia is not harmful or dangerous is rarely effective in producing extinction and changing behavior. Based on these observations, a core hypothesis of the preparedness theory thereby is that humans are biologically predisposed to learn to associate evolutionarily prepared threat-relevant stimuli with a naturally aversive event in comparison with threat-irrelevant stimuli or evolutionarily novel threat-relevant stimuli. Following the criteria for prepared learning, these prepared associations are accordingly thought to be (a) resistant to extinction, (b) easily and readily acquired, (c) selective, and (d) resistant to verbal instructions.

Empirical evidence for preparedness theory

Research testing the preparedness theory's basic assumptions has been pioneered by Öhman and his colleagues in the 1970s (e.g., Fredrikson, Hugdahl, & Öhman, 1976; Öhman, 1979; Öhman & Dimberg, 1978; Öhman, Eriksson, et al., 1975; Öhman, Erixon, & Löfberg,

1975; Öhman et al., 1976; Öhman, Fredrikson, & Hugdahl, 1978). These experiments have typically involved comparisons between threat-relevant and threat-irrelevant stimulus categories using a differential Pavlovian aversive conditioning procedure. In this procedure, one reinforced (CS+) and one unreinforced (CS-) stimulus from either threat-relevant or threat-irrelevant categories are presented to two separate groups of participants. Three phases are generally included therein: (a) habituation, in which each stimulus is presented several times without being reinforced, (b) acquisition, in which each CS+ is associated with an electric stimulation (US), whereas each CS- is never paired with the US, and (c) extinction, which involves several unreinforced presentations of each stimulus. Acquisition is defined as larger responding to the CS+ than to the CS- during the acquisition phase, and resistance to extinction is defined as the persistence of the CR (i.e., differential responding between the CS+ and the CS-) during the extinction phase where the US is no longer delivered (McNally, 1987). The CR has been most frequently assessed through SCR magnitude, although the startle eyeblink reflex, heart rate variations, and finger-pulse volume have been also used sometimes (see, e.g., Mallan et al., 2013).

Resistance to extinction. A critical empirical finding supporting the preparedness theory is the enhanced resistance to extinction of the CR to threat-relevant stimuli from phylogenetic origin, whereas the CR to threat-irrelevant stimuli rapidly extinguishes (for reviews, see Mallan et al., 2013; McNally, 1987; Öhman & Mineka, 2001; but see Åhs et al., 2018, for a recent systematic review questioning the strength of this evidence). Such resistance to extinction has been demonstrated across various classes of evolutionarily threat-relevant stimuli. For instance, animal threat-relevant stimuli, such as snakes or spiders, have been shown to produce more persistent differential responding during extinction than threat-irrelevant stimuli, such as birds, flowers, or mushrooms (e.g., Öhman, 1979; Öhman, Eriksson, et al., 1975; Öhman, Erixon, et al., 1975; Öhman et al., 1976, 1978; Öhman & Mineka, 2001; Olsson et al., 2005), or even threat-relevant stimuli from ontogenetic origin, such as weapons or electrical outlets (E. W. Cook et al., 1986; Hugdahl & Kärker, 1981). Similarly, enhanced resistance to extinction has been reported in response to social threat-relevant stimuli with an evolutionary basis, such as angry faces, relative to social nonthreatening stimuli, such as happy or neutral faces (e.g., Öhman & Dimberg, 1978; Rowles, Mallan, & Lipp, 2012; see Dimberg & Öhman, 1996, for a review). More recently, other classes of social threat-relevant stimuli, such as outgroup faces (e.g., Mallan, Sax, & Lipp, 2009; Navarrete et al., 2009; Olsson et al., 2005) or dominant faces (Haaker, Molapour, & Olsson, 2016), have likewise been shown to

induce a greater persistence of the CR in comparison with ingroup faces or subordinate faces, respectively.

Ease of acquisition. Whereas enhanced resistance to extinction has been frequently demonstrated to threat-relevant stimuli (Öhman & Mineka, 2001), evidence for the hypothesis that evolutionarily threat-relevant stimuli are more easily and readily associated with an aversive event remains by comparison scarce (see McNally, 1987). An explanation for this lack of experimental support relates to the possible presence of ceiling effects in the CR acquisition readiness, thus obscuring the emergence of differences between stimulus categories (Ho & Lipp, 2014; Lissek, Pine, & Grillon, 2006; Öhman & Mineka, 2001). Nevertheless, some studies have revealed faster (Ho & Lipp, 2014; see also Atlas & Phelps, 2018) or larger (e.g., Fredrikson et al., 1976; Öhman et al., 1978) acquisition of a CR to animal threat-relevant stimuli (snakes and spiders) than to threat-irrelevant stimuli (e.g., birds and flowers). Likewise, outgroup faces have been reported to produce faster (Navarrete et al., 2012) and larger (Olsson et al., 2005) acquisition of a CR than ingroup faces.

A more stringent form of the ease of acquisition hypothesis contends that Pavlovian aversive conditioning to threat-relevant, but not to threat-irrelevant, stimuli should be observed after a single learning episode (Seligman, 1970, 1971). Although this hypothesis has not been investigated systematically so far (see Mallan et al., 2013), Öhman and colleagues (Öhman, Eriksson, et al., 1975) reported resistance to extinction of the CR to snake images relative to pictures of houses after a single pairing with an aversive US, these effects not differing from the effects obtained after five pairings. These findings therefore suggested that enhanced Pavlovian conditioning to animal threat-relevant stimuli can occur as a result of one-trial learning (but see Lipp, Cronin, Alhadad, & Luck, 2015, for contradictory findings). However, it is important to note that these results should be interpreted with caution as this study implemented a single-cue conditioning procedure rather than a differential conditioning procedure, thus rendering it difficult to disentangle elevated responding in SCR to animal threat-relevant stimuli from enhanced responding in SCR due to conditioning (Mallan et al., 2013).

Selectivity. According to preparedness theory, threat-relevant stimuli from evolutionary origin should be selectively associable with naturally aversive events rather than be generally more effective CSs than neutral stimuli (McNally, 1987; Öhman & Mineka, 2001; Seligman, 1970, 1971). This selectivity of prepared associations between evolutionarily threat-relevant stimuli and an aversive event has been highlighted by the studies of Öhman and

collaborators (Öhman et al., 1976, 1978). Results demonstrated that the CR to animal threat-relevant, but not to threat-irrelevant, stimuli was resistant to extinction after these stimuli were contingently paired with an aversive US (i.e., electric stimulation); by contrast, no resistance to extinction was observed to animal threat-relevant stimuli after they were paired with a nonaversive US (i.e., a tone serving as an imperative stimulus in a reaction time task). Hamm et al. (1989) furthermore reported that high-belongingness pairs of CS and US (e.g., angry faces and loud scream, respectively) elicited superior acquisition and resistance to extinction than low-belongingness pairs (e.g., landscape and loud scream) as measured with finger-pulse volume responses. Taken together, these findings suggested that preparedness effects result from selective or prepared associations between phylogenetically threat-relevant stimuli and aversive outcomes (but see Åhs et al., 2018; McNally, 1987, for a discussion about the equivocal nature of the empirical evidence in favor of the preparedness selectivity hypothesis).

Irrationality. In order to assess the hypothesis that prepared associations involving evolutionarily threat-relevant stimuli and aversive events are a form of noncognitive learning (Seligman, 1971), experimental studies have operationalized irrationality of phobias as the persistence of the CR after instructed extinction, that is the explicit instructions that the US will no longer be administered and/or the removal of the electric stimulation electrode during the extinction phase (e.g., Mallan et al., 2013). Consistent with preparedness theory, it has been shown that the CR to animal threat-relevant stimuli from phylogenetic origin was impervious to instructed extinction, whereas the CR to threat-irrelevant stimuli immediately extinguished after such instructions (Hugdahl, 1978; Hugdahl & Öhman, 1977; Lipp & Edwards, 2002; Öhman, Erixon, et al., 1975; Soares & Öhman, 1993). Nonetheless, these findings have turned out to be relatively difficult to replicate (see, e.g., McNally, 1987; Mertens, Raes, & De Houwer, 2016).

Summary

The reviewed series of experiments on preparedness theory indicates that Pavlovian aversive conditioning to evolutionarily threat-relevant stimuli – especially animal threat-relevant stimuli – largely meets the criteria for prepared learning postulated by Seligman (1970, 1971), in the sense that it is (a) resistant to extinction (e.g., Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001), (b) readily acquired (e.g., Ho & Lipp, 2014; Öhman, Eriksson, et al., 1975), (c) selective to contingent pairings with an aversive US (e.g., Öhman et al., 1976), and (d) impervious to verbal instructions (e.g., Hugdahl & Öhman, 1977). Of note, the most robust finding providing evidence for preparedness theory consists of enhanced

resistance to extinction, whereas the hypotheses regarding ease of acquisition, selectivity, and irrationality have by comparison received less consistent support.

2.3.2. The fear module: An evolved module for fear and fear learning

Fear module theory

Öhman and Mineka (2001) provided an important theoretical extension to preparedness theory. Ascribing a privileged evolutionary status to fear due to its central importance for survival, these authors assumed that natural selection has pressured animals and humans to shape adaptive mechanisms promoting preferential activation of defensive behaviors in response to cues that signal the occurrence of survival threats (e.g., Bolles, 1970; Seligman, 1970, 1971). In this context, they argued that humans have been biologically predisposed to (learn to) fear events and situations that have posed threats to the survival of their ancestors across evolution, thus aligning with the core hypothesis of preparedness theory (Seligman, 1971). Integrating evidence for this assumption with a wide range of findings from affective, cognitive, and behavioral neuroscience in humans and animals, Öhman and Mineka therefore proposed the existence of an evolved fear module implemented in the human brain dedicated to preferentially processing threat-relevant stimuli from phylogenetic origin, thereby subserving preferential Pavlovian aversive conditioning to these stimuli. The fear module is conceptualized as an apparatus that enables the organism to readily activate defensive behaviors and elicit associated psychophysiological responses to threat-related stimuli in aversive contexts (Öhman & Mineka, 2001; see Figure 2.5). It is conceived to have been specifically tailored by evolutionary pressures to help the organism adapt to survival-threatening situations frequently encountered by the species during its evolution (Mineka & Öhman, 2002; Öhman & Mineka, 2001). The fear module presents four principal characteristics, each resulting from evolutionary contingencies: (a) input selectivity, (b) automaticity, (c) encapsulation, and (d) specific neural circuitry.

Characteristics of the fear module

Input selectivity. The fear module is proposed to be selective with regard to the input to which it responds. More specifically, it is assumed to be preferentially activated in aversive contexts by stimuli that have been correlated with fear through evolution. Pavlovian conditioning processes additionally allow for further expanding the range of stimuli that are able to activate the fear module, with a particular emphasis on stimuli that have been associated

with situations providing recurrent survival threats in an evolutionary perspective (Mineka & Öhman, 2002; Öhman & Mineka, 2001). Although arbitrary cues can potentially activate the fear module through Pavlovian aversive conditioning under certain circumstances, associations between these stimuli and fear are thought to be more difficult to learn and less resistant to extinction as opposed to associations between evolutionary threat stimuli and fear (Öhman & Mineka, 2001).

Automaticity. The fear module is proposed to be automatically activated by evolutionarily threat-relevant stimuli without requiring voluntary attention or conscious awareness. Such automaticity would be the result of evolutionary pressures that have encouraged the development of mechanisms enabling rapid processing and prioritization of stimuli related to survival threats with minimal neural computations (Öhman & Mineka, 2001); thereby procuring an adaptive advantage that contributes to promoting the organism's survival.

Encapsulation. Encapsulation refers to the fear module's relative independence from cognition (Mineka & Öhman, 2002). More precisely, the fear module is supposed to be impervious and resistant to conscious cognitive control or higher cognitive influences, such as expectancies or language (Mineka & Öhman, 2002; Öhman & Mineka, 2001). This characteristic is consistent with the preparedness theory assumption of irrationality of phobias. Even though the fear module is impenetrable to cognition, it may conversely exert an influence on cognitive processes (Öhman & Mineka, 2001; see Figure 2.5). Of note, encapsulation differs from automaticity in that automaticity relates to the activity initiation of the fear module even in the absence of conscious processing, whereas encapsulation pertains to the maintaining of the fear response independently of cognitive control once initiated (Mineka & Öhman, 2002; Öhman & Mineka, 2001).

Specific neural circuitry. Öhman and Mineka (2001) proposed that the fear module has a dedicated neural implementation in a subcortical neural network centered on the amygdala. Following the traditional conception of the amygdala as the brain substrate for fear, this proposal primarily relied on animal research, which has shown that the amygdala – a subcortical ensemble of neural nuclei with partially distinct functional features located in the anterior medial temporal lobe (e.g., LeDoux, 2000) – is crucially involved in fear-related processes and behaviors (e.g., Weiskrantz, 1956) and, more particularly, in Pavlovian aversive conditioning (e.g., Davis & Whalen, 2001; Fendt & Fanselow, 1999; LeDoux, 1996, 2000; Maren, 2001; Phelps & LeDoux, 2005). Strongly inspired by LeDoux's (1996) dual-route model, Öhman (2005) further suggested that fear reactions triggered by threat stimuli are

mediated by an automatic and subcortical pathway passing through the superior colliculi and the pulvinar nucleus of the thalamus before accessing the amygdala. This neural circuitry would notably have an ancient evolutionary origin, as suggested by its subcortical location and high conservation across species (Öhman & Mineka, 2001). The fear module's ancient origin and brain location would hence be responsible for its automaticity and relative impenetrability to cognition (LeDoux, 1996; Öhman & Mineka, 2001).

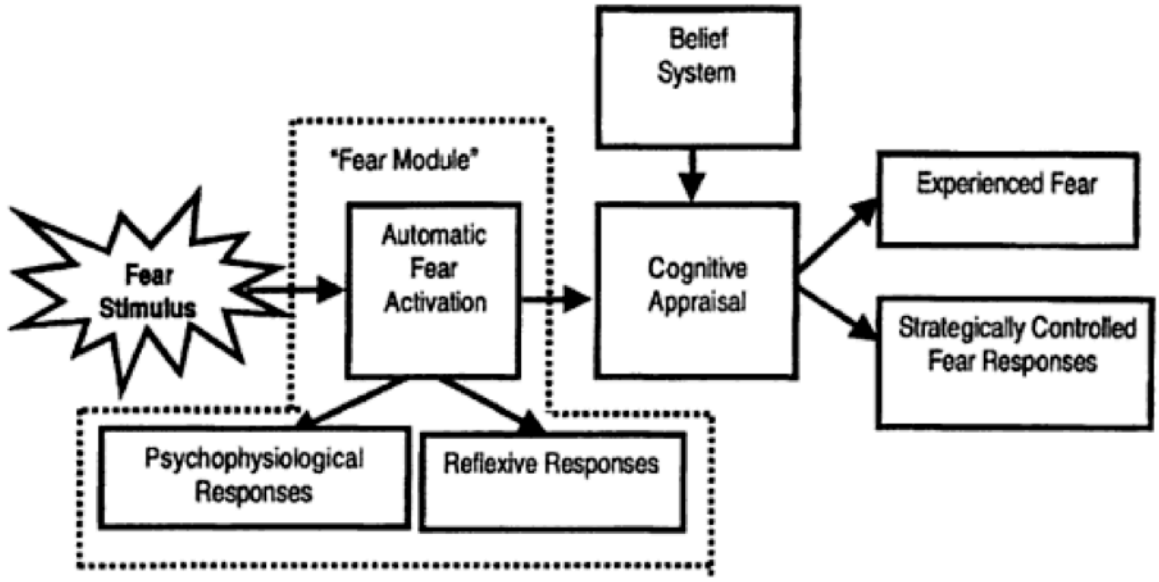

Figure 2.5. Schematic representation of the fear module and its hypothetical relations with other psychophysiological, behavioral, and cognitive systems. Adapted from Öhman and Wiens (2004).

Empirical basis of the fear module

Input selectivity. The hypothesized differential sensitivity of the fear module to evolutionarily threat-relevant versus threat-irrelevant stimuli has been supported by findings from human Pavlovian conditioning experiments on preparedness. As mentioned previously (see chapter 2.3.1 here above), these experiments have demonstrated selective associations between threat-relevant stimuli from phylogenetic origin and aversive USs, as reflected by faster and more persistent Pavlovian aversive conditioning to these stimuli (e.g., Ho & Lipp, 2014; Öhman & Dimberg, 1978; Öhman, Eriksson, et al., 1975; Öhman et al., 1976; Öhman & Mineka, 2001).

In addition, such selective associations have also been reported in non-human primates (e.g., M. Cook & Mineka, 1989, 1990; Mineka et al., 1984). Using a vicarious conditioning paradigm, these studies have shown that lab-reared observer monkeys selectively acquired fear

of snakes after watching videotapes of wild-reared model monkeys behaving fearfully with a (real or toy) snake, but did not acquire fear of flowers when the snake was replaced with a flower in the videotapes (M. Cook & Mineka, 1989, 1990). Similar results have been likewise observed when (real or toy) crocodiles and rabbits were used as animal threat-relevant and threat-irrelevant stimuli, respectively (M. Cook & Mineka, 1989). In contrast, no preferential learning occurred when snake stimuli were used as discriminative stimuli for appetitive food rewards (M. Cook & Mineka, 1990). Importantly, the observer monkeys were all completely naive to the stimuli used in these experiments as they had never been exposed to them previously (Öhman & Mineka, 2001), hence suggesting the involvement of both a biological preparedness mechanism and selective associations (Öhman & Mineka, 2001; Seligman, 1970, 1971).

Empirical support in line with the fear module selectivity have also been provided by human experiments using a covariation bias paradigm (e.g., Tomarken, Mineka, & Cook, 1989). This paradigm assesses whether selective associations between evolutionarily threat-relevant stimuli and aversive outcomes can be evidenced in association or covariation judgments. Specifically, Tomarken et al. (1989) implemented an illusory correlation paradigm, in which participants were exposed to threat-relevant stimuli (snakes or spiders) and threat-irrelevant stimuli (flowers and mushrooms), with each stimulus being randomly followed either with an electric stimulation, a tone, or no outcome. Participants were then asked to evaluate the degree of covariation between the stimulus categories and the outcomes. Results indicated that participants with a high level of fear of the threat-relevant stimuli largely overestimated the contingency between the threat-relevant stimuli and the electric stimulations, which was not the case for the other stimulus categories and outcomes; participants with a low level of fear exhibited a tendency in the same direction. Moreover, this covariation bias was specific to the electric stimulation's aversiveness rather than its salience, as the effect did not occur in a second experiment to a nonaversive outcome that was rated as similarly salient as the electric stimulation. Subsequent studies have replicated and extended these initial findings to the comparison between phylogenetically and ontogenetically threat-relevant stimuli, generally reporting a covariation bias to the former but not to the latter (e.g., Amin & Lovibond, 1997; Kennedy, Rapee, & Mazurski, 1997; Tomarken, Sutton, & Mineka, 1995).

Another related line of research has further shown that evolutionarily threat-relevant stimuli are more effective in capturing attention than threat-irrelevant stimuli (Öhman, Flykt, & Esteves, 2001; Öhman, Lundqvist, & Esteves, 2001). More precisely, animal (snakes or

spiders; Öhman, Flykt, et al., 2001) and social (threatening faces; Öhman, Lundqvist, et al., 2001) threat-relevant stimuli, as well as simple geometric shapes conveying a threat meaning similar to angry faces (i.e., downward-pointing "V"; Larson, Aronoff, & Stearns, 2007), have been reported to be detected more quickly in a visual search paradigm than threat-irrelevant stimuli, such as flowers, friendly faces, or geometric shapes devoid of any threat meaning, respectively. This preferential attentional capture by threat-relevant stimuli from phylogenetic origin has been categorized as preattentive and concurs with the notion that the fear module is selectively activated by evolutionary threat stimuli in an automatic fashion (Öhman & Mineka, 2001).

Automaticity. Evidence for the view that the fear module is automatically activated by evolutionarily threat-relevant stimuli essentially stems from human Pavlovian aversive conditioning studies using backward masking. In this procedure, a target stimulus is presented very briefly (e.g., 30 ms) and is immediately followed by a masking stimulus with the aim of preventing the target stimulus from being consciously recognized (Öhman & Mineka, 2001). Öhman and colleagues (e.g., Öhman & Soares, 1993; Soares & Öhman, 1993; see also Öhman, 1986) notably demonstrated that the CR to animal threat-relevant stimuli persisted even when they were masked during extinction, whereas masking led to extinction of the CR to threat-irrelevant stimuli. Resistance to extinction of the CR to angry faces was reported as well when they were masked with neutral faces during the extinction phase, which was not the case for happy and neutral faces (Esteves, Dimberg, & Öhman, 1994). In a similar vein, resistance to extinction effects were shown in response to unmasked animal threat-relevant (Öhman & Soares, 1998) and angry faces (Esteves, Parra, Dimberg, & Öhman, 1994), but not to unmasked threat-irrelevant stimuli, when these stimuli were masked during the acquisition rather than the extinction phase. Altogether, Öhman and Mineka (2001) interpreted these results as showing that CRs to biologically threat-relevant stimuli are automatic and can be elicited independently from visual awareness, thereby supporting the fear module's automaticity.

Encapsulation. Experiments by Hugdahl and Öhman (Hugdahl, 1978; Hugdahl & Öhman, 1977; Öhman, Erixon, et al., 1975; see also Lipp & Edwards, 2002) provided support for the encapsulation of the fear module in showing that the CR to animal threat-relevant stimuli from evolutionary origin was resistant to extinction even after explicit verbal instructions emphasizing that the US would no longer be administered, whereas the CR to threat-irrelevant stimuli immediately extinguished after verbal instructions. Soares and Öhman (1993) furthermore extended these results by combining a backward masking procedure with

verbal instructions stressing that the US would no longer be presented during extinction. They observed that the CR to animal threat-relevant was persistent during extinction and remained unaffected neither by the masking procedure, nor by verbal instructions; by contrast, both manipulations eliminated the CR to threat-irrelevant stimuli. Thus, Öhman and Mineka (2001) interpreted these findings as corroborating the purported impenetrability of the fear module to influences from higher-level cognitive factors.

Specific neural circuitry. A crucial argument for the conception of the amygdala as the fear module's center revolves around its fundamental role in Pavlovian aversive conditioning across animals and humans (e.g., LeDoux, 1996; Phelps & LeDoux, 2005). Animal lesion studies (see Fendt & Fanselow, 1999, for a review) and human neuropsychological studies in patients with amygdala damage (e.g., Bechara et al., 1995; LaBar, LeDoux, Spencer, & Phelps, 1995) have underlined that amygdalar lesions impair or abolish Pavlovian aversive conditioning. Relatedly, functional neuroimaging studies in humans have delineated the amygdala as a central neural correlate of Pavlovian aversive conditioning (e.g., Büchel et al., 1998; LaBar et al., 1998; Morris, Dolan, & Friston, 1998, 1999; Olsson & Phelps, 2007; Phelps, 2006; Phelps et al., 2004). In particular, Morris et al. (1998) demonstrated that the amygdala was more activated in response to angry face CSs+ than angry face CSs- during extinction, both for masked and unmasked presentations. Morris et al. (1999) further reported that the heightened amygdala activations by masked stimuli, but not by unmasked stimuli, could be predicted by the superior colliculus and the pulvinar activations, which suggests a potential involvement of a subcortical route from the thalamus to the amygdala in processing threat-related stimuli without conscious awareness (see LeDoux, 1996). Accordingly, Öhman (2005) proposed that the amygdala corresponds to an automatic threat detector, thereby accentuating its conceptualization as a fear module.

Summary

The fear module theory (Öhman & Mineka, 2001) proposes that the preferential processing of, and enhanced Pavlovian aversive conditioning to, threat-relevant stimuli from evolutionary origin is subserved by an evolved fear module in the human brain (with a pivotal role of the amygdala hypothesized). These stimuli are assumed to preferentially and automatically activate the fear module in aversive contexts, and therefore readily enter into selective association with aversive events even when they are presented in a highly degraded manner, as substantiated by a large number of Pavlovian aversive conditioning studies in humans and monkeys (e.g., M. Cook & Mineka, 1989, 1990; Esteves, Parra et al., 1994; Öhman

& Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001; Öhman & Soares, 1998). The fear module is considered encapsulated and impenetrable to cognitive control, as supported by experiments indicating that associations between animal threat-relevant stimuli and aversive outcomes are impervious to verbal instructions (e.g., Hugdahl & Öhman, 1978; Soares & Öhman, 1993). Finally, the fear module is thought to reflect the operation of a specific neural network for fear elicitation and fear learning that is centered on the amygdala (Öhman & Mineka, 2001), as highlighted by the central role of this phylogenetically conserved brain structure in Pavlovian aversive conditioning.

2.3.3. Criticisms of the biological preparedness models and alternative explanations

Although the preparedness and fear module theories have received considerable empirical support, criticisms of the biological preparedness perspective have been formulated (see, e.g., Davey, 1995; Mallan et al., 2013; McNally, 1987; see also Åhs et al., 2018). In particular, the biological preparedness accounts have been argued to overly rely on evolutionary explanations, without considering the importance of ontogenetic and cultural factors in modulating preferential emotional learning (e.g., Mallan et al., 2013). In agreement with this view, several findings have cast doubt on the putative superiority of threat-relevant stimuli from phylogenetic origin relative to threat-relevant stimuli from ontogenetic origin. For instance, Hugdahl and Johnsen (1989) reported that enhanced resistance to extinction to pictures of guns pointed toward participants (i.e., a cultural threat) associated with a loud noise as the US was not statistically different from resistance to extinction to snake pictures directed toward participants (i.e., a biological threat) associated with an electric stimulation as the US. Whereas Öhman and Mineka (2001) acknowledged that the fear module can also be activated by ontogenetic threat-related stimuli under certain circumstances (e.g., extensive training) in addition to its preferential activation by phylogenetic threat-relevant stimuli, they however clearly argued that such activations should not occur to threat-relevant stimuli from ontogenetic origin when they are presented under suboptimal, potentially unaware conditions. In contrast to this prediction, Flykt, Esteves, and Öhman (2007) showed an enhanced resistance to extinction to both unmasked and masked presentations of snake and gun pictures during extinction when they were directed toward participants, but not when the masked snake and gun pictures were directed away from them. Jointly, these studies seem to indicate a similar conditioning effect across phylogenetic and ontogenetic threat-relevant stimuli, which is at odds with the tenets of the preparedness and fear module theories. Similar results were also

documented with respect to attention processes. Several experiments (e.g., Blanchette, 2006; Brosch & Sharma, 2005; Fox, Griggs, & Mouchlianitis, 2007) have found that both phylogenetic and ontogenetic threats attract attention more quickly than neutral stimuli in a comparable manner. A recent study further raised the possibility that modern threats might even be detected faster than ancient threats (Subra, Muller, Fourgassie, Chauvin, & Alexopoulos, 2017). Taken together, these empirical data suggest that the key factor modulating Pavlovian aversive conditioning and attentional biases to threat-relevant stimuli may be threat-relevance rather than evolutionary history per se.

In a similar vein, another line of evidence inconsistent with the preparedness and fear module theories concerns Pavlovian aversive conditioning to social threat-relevant stimuli, which appears to be more malleable than Pavlovian aversive conditioning to animal threat-relevant stimuli (Mallan et al., 2013). Whereas learned threat to animal threat-relevant stimuli has been shown to be impervious to instructed extinction (Hugdahl, 1978; Hugdahl & Öhman, 1977; Lipp & Edwards, 2002; Öhman, Erixon, et al., 1975; Soares & Öhman, 1993), learned threat to social threat-stimuli has been reported to swiftly extinguish following instructed extinction and electrode removal both for angry faces (Rowles et al., 2012) and outgroup faces (Mallan et al., 2009). These results suggests that Pavlovian aversive conditioning to social threat stimuli is susceptible to alterations by cognition, unlike Pavlovian aversive conditioning to animal threat stimuli (Mallan et al., 2013). It has therefore been proposed that preferential Pavlovian aversive conditioning to social threat-relevant stimuli may hinge on sociocultural factors, such as negative stereotypes or social norms, rather than purely biological factors, such as biological preparedness (Mallan et al., 2013; see also Olsson et al., 2005). Accordingly, these observations underline the notion that both phylogenetic and ontogenetic factors may play an important role in preferential emotional learning.

Further, a growing body of research has likewise questioned the hypothesized preferential processing of threat-relevant stimuli (Lipp, Kempich, Jee, & Arnold, 2014). For instance, Lipp et al. (2014) conducted an experiment implementing a differential Pavlovian aversive conditioning paradigm in which images of snakes and wallabies were presented supraliminally and suboptimally using a binocular masking procedure transiently blocking awareness thereof. Results showed that supraliminal and suboptimal presentations of both image classes induced reliable differential Pavlovian aversive conditioning, thereby reflecting no preferential emotional learning to evolutionarily threat-relevant stimuli compared with cute, nonthreatening stimuli (i.e., wallabies), even under conditions of highly degraded input. In that

regard, these findings are strongly inconsistent with the predictions made by the preparedness and fear module theories.

Based on these considerations, alternative theories to the biological preparedness models have been elaborated for explaining evolutionary threat-relevance phenomena (see McNally, 1987; Öhman & Mineka, 2001, for reviews). We address some of these alternative theoretical models in the following sections.

Selective sensitization

Several authors (e.g., J. A. Gray, 1987; Lovibond, Siddle, & Bond, 1993) have argued that the enhanced responding to evolutionarily threat-relevant stimuli induced by their association with aversive outcomes is more parsimoniously explained by selective sensitization, a nonassociative process, rather than by selective associations. According to this account, phylogenetic threat-relevant stimuli are biologically predisposed to elicit heightened fear reactions; however, such preexisting response tendencies require specific conditions to emerge, such as a state of arousal or threat. During Pavlovian aversive conditioning, the mere threat of electric stimulation would hence be sufficient to selectively sensitize and boost responding to threat-relevant stimuli to a larger extent than to threat-irrelevant stimuli. In this context, Lovibond et al. (1993) suggested that the effects of enhanced resistance to extinction to evolutionarily threat-relevant stimuli originate from increased responding to threat-relevant stimuli from evolutionary origin due to selective sensitization and the fact that such heightened responding is maintained to the threat-relevant CS+ during acquisition through its systematic pairing with the aversive US. Inversely, sensitized responding to the threat-relevant CS- is reduced during acquisition as the threat-relevant CS- acquires inhibitory properties through the prediction of the US absence, thus amplifying the CS+/CS- differentiation to threat-relevant stimuli prior to the extinction phase. On the other hand, responding to threat-irrelevant stimuli is not sensitized; the CS+/CS- differentiation is consequently lower than to threat-relevant stimuli before extinction, which may in turn result in faster extinction comparatively (Lovibond et al., 1993).

Despite the fact that selective sensitization can evidently occur during Pavlovian conditioning (e.g., Lipp et al., 2015; Öhman, Eriksson, et al., 1975), it has, however, been suggested to be a relatively short-lived phenomenon (e.g., Lipp et al., 2015). Thus, selective sensitization has been argued to be insufficient to explain the long-lasting and lingering effects

typically observed in Pavlovian conditioning studies using threat-relevant stimuli in humans and non-human primates (Öhman & Mineka, 2001).

Expectancy bias model

Davey (1992, 1995) proposed an alternative model that starkly opposes to the biological preparedness perspective as well as the fear module encapsulation hypothesis. According to this alternative model, preferential Pavlovian aversive conditioning to threat-relevant stimuli relies on cognitive biases rather than biological preparedness. Specifically, Davey holds that preferential associations between threat-relevant stimuli and aversive events arise from an expectancy bias, which consists of heightened expectation of aversive outcomes following threat-relevant stimuli. This expectancy bias would be notably a key determinant of Pavlovian CRs (Davey, 1992). In line with this view, Davey (1992) showed that human participants had enhanced a priori expectancies that aversive events would follow threat-relevant stimuli compared with threat-irrelevant stimuli, even when they were instructed that no aversive stimulus would be delivered. Such differential a priori expectancies between threat-relevant and threat-irrelevant stimuli have been shown to extinguish with continued nonreinforcement and to be reinstated by CS-US pairings or explicit threat of an aversive US (Davey, 1992). Differential expectancies have furthermore been reported to be translated into differential SCRs through exposure to the actual US either before or during the experiment (Davey, 1992), thereby highlighting that enhanced US expectancy after threat-relevant CSs is strongly associated with enhanced SCR to these stimuli under certain conditions (see also Dawson, Schell, & Banis, 1986). Importantly, expectancy bias is hypothesized to be essentially determined by ontogenetic, cultural factors, such as the CS dangerousness, without however excluding the possibility that such bias may reflect the complex interplay of evolutionary and cultural influences (Davey, 1995). Accordingly, the expectancy bias model can flexibly accommodate preferential Pavlovian aversive conditioning to both phylogenetic and ontogenetic threat-relevant stimuli (Flykt et al., 2007; Hugdahl & Johnsen, 1989).

Nevertheless, whereas Davey's (1992, 1995) expectancy bias model has potential in explaining a wide range of behavioral data generated in the context of Pavlovian aversive conditioning to threat-relevant stimuli in humans, Öhman and Mineka (2001) raised several problems with this model. In particular, it is inconsistent with findings that have demonstrated persistent differential SCRs to evolutionarily threat-relevant stimuli, but not to threat-irrelevant stimuli, even though US expectancies to both of these stimulus categories were already extinguished (Schell, Dawson, & Marinkovic, 1991), thus suggesting a dissociation between

expectancy bias and SCRs (Öhman & Mineka, 2001). Additionally, the expectancy bias model has difficulty in explaining the results from observational conditioning in monkeys showing that lab-reared monkeys with no ontogenetic experience with snakes readily developed fear reactions to these stimuli in a selective manner (M. Cook & Mineka, 1989, 1990; Mineka & Öhman, 2002). For these reasons, it has been argued that expectancies are not sufficient to account for the body of evidence supporting the preparedness model, but may conversely be consequences rather than causes of fear responding (see Mineka & Öhman, 2002; Öhman & Mineka, 2001).

Conditioned stimulus salience

Given that physical properties of the CS amplifying its salience can enhance its conditionability (e.g., Mackintosh, 1975; Rescorla & Wagner, 1972), CS salience has been suggested as a putative alternative mechanism to biological preparedness to explain the preferential Pavlovian aversive conditioning to threat-relevant stimuli (McNally, 1987). According to this view, threat-relevant stimuli are preferentially conditioned to threat not because of their threat-relevance, but due to their high salience.

Although more salient or intense stimuli – in the sense of physical or perceptual salience (see Footnote 1 here above) – have been reported to be more readily conditioned than less salient stimuli (e.g., Mackintosh, 1975; Pearce & Hall, 1980; Rescorla, 1988a; Rescorla & Wagner, 1972), it has been demonstrated that neutral stimuli that are highly perceptually salient (i.e., with a high visual complexity) do not induce enhanced resistance to extinction relative to neutral stimuli with a lower perceptual salience (i.e., low visual complexity; Öhman et al., 1976, Experiment 2). These findings thereby indicate that physical salience alone does not provide a satisfactory explanation for the effects of preferential Pavlovian aversive conditioning observed with threat-relevant stimuli (McNally, 1987; Öhman & Mineka, 2001). As mentioned previously (see chapter 2.3), the physical salience hypothesis also appears inconsistent with classical models of Pavlovian conditioning (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972), which predict that the CR to more salient stimuli should extinguish more rapidly than the CR to less salient stimuli (see, e.g., Siddle et al., 1988).

Nonetheless, it is important to note that salience can be equally conceptualized as not being limited only to the stimulus' physical characteristics but also encompassing the stimulus' relative importance to motivational contingencies that relate to the organism's needs and goals (Cunningham & Brosch, 2012; Öhman & Mineka, 2001; Rescorla, 1988a). From this

perspective, it could therefore be suggested that threat-relevant stimuli benefit from enhanced Pavlovian aversive conditioning because of their high motivational salience. To the best of our knowledge, such motivational salience hypothesis has not been investigated to date. Interestingly, a motivational salience account is however closely related to the alternative model that we propose hereafter, as it likewise suggests that preferential Pavlovian aversive conditioning stems from the interaction between the stimulus at stake and the organism's motivational state.

2.3.4. Relevance detection as a key mechanism underlying preferential emotional learning in humans

Appraisal theories of emotion

The conception that organisms have developed an evolved module of fear and fear learning implemented in the amygdala has been further challenged by several contemporary theories of emotion, which postulate that emotions are not modular but are underlain by common and shared dimensions or mechanisms (e.g., Russell, 2003; Sander et al., 2003, 2005, 2018). Among these theories, appraisal theories of emotion (e.g., Moors, Ellsworth, Scherer, & Frijda, 2013; Sander et al., 2003, 2005, 2018; Scherer & Moors, 2019; Scherer, Schorr, & Johnstone, 2001) propose that emotions are elicited and differentiated according to appraisal processes that continuously detect and evaluate the significance of stimulus events or situations in the environment for the organism's well-being (Moors et al., 2013). Such appraisal processes are characterized by their subjectivity, being a function of the individual and the specific situation or context, thus highlighting the crucial influence of intra- and inter-individual differences on emotional processes (Brosch, Pourtois, & Sander, 2010). Moreover, appraisal processes have been suggested to occur at different levels of processing and to be automatic in most cases, although automaticity is not diagnostic of them (e.g., Leventhal & Scherer, 1987; Moors, 2010).

Importantly, appraisal theories suggest that a key mechanism involved in emotion elicitation is relevance detection. Relevance detection is conceptualized as a rapid and flexible mechanism that serves to detect and establish whether a stimulus in the environment is relevant to the organism's concerns (Frijda, 1986, 1988; Pool, Brosch, et al., 2016; Sander et al., 2003, 2005). A stimulus is detected and appraised as affectively relevant in the event "it increases the probability of satisfaction or dissatisfaction toward a major concern of the individual"

(Sander, 2013, p. 22). Affective relevance hence reflects the interaction between the stimulus at hand and the organism's current concerns. Concerns refer to affective representations of psychological and physiological motives, needs, goals, and values that are of central importance to the organism (Frijda, 1986, 1988; Pool, Brosch, et al., 2016) and can be broadly defined as "disposition[s] to desire the occurrence or nonoccurrence of a given kind of situation" (Frijda, 1986, p. 335).

In contrast to the fear module theory, appraisal theories do not assign a special role for the emotion of fear in comparison with other emotions, or for evolutionarily threat-relevant stimuli relative to other affectively relevant stimuli. In fact, whereas the notion of relevance detection captures the dimension of biological and evolutionary significance, it is not limited to it and also refers to other types of concerns (Sander, 2013; Sander et al., 2018). According to appraisal theories, affectively relevant stimuli are preferentially processed; this preferential processing being considered to function in many similar ways across negative and positive stimuli regardless of their evolutionary history as such.

Consistent with this view, a growing body of research has demonstrated that both negative and positive stimuli that are relevant to the organism's concerns can benefit from enhanced processing relative to neutral stimuli with less relevance. For instance, empirical results in the domain of emotional attention have shown that rapid spatial orienting is not only observed for negative or threat-related stimuli, but also evident and equally strong for positive relevant stimuli (e.g., Brosch, Sander, Pourtois, & Scherer, 2008; Brosch, Sander, & Scherer, 2007; Pool, Brosch, et al., 2016). Brosch and colleagues (2008) notably showed that spatial attention orienting processes were modulated in a highly similar manner at the behavioral and brain levels by angry faces and baby faces, indicating that threat-relevant and positive biologically relevant stimuli benefited from similar prioritization during attention selection. Additionally, Pool and colleagues (Pool, Brosch, et al., 2014) demonstrated across two experiments that (a) attentional resources were rapidly oriented toward an initially neutral stimulus that acquired affective relevance through association with a primary reward (i.e., a chocolate odor) during Pavlovian conditioning, and (b) when the reward was devaluated, attention was no longer preferentially oriented toward the stimulus previously paired with the reward, thereby reflecting the fact that rapid attention orienting was highly dynamic and modulated on the basis of the stimulus' affective relevance. Relatedly, stimuli that are relevant to the individual's concerns have also been reported to facilitate memory (see Montagrin, Brosch, & Sander, 2013; Montagrin & Sander, 2016; Montagrin et al., 2018).

Furthermore, converging evidence in neuroimaging has shown that the amygdala is not specifically involved in the processing of threat-relevant stimuli, but is more broadly involved in the processing of stimuli that are affectively relevant to the organism (Cunningham & Brosch, 2012; Pessoa & Adolphs, 2010; Sander et al., 2003; Sergerie, Chochol, & Armony, 2008), including positive or rewarding stimuli (Gottfried et al., 2003; Sergerie et al., 2008). In addition, the amygdala is considered a core brain structure of the motivational circuits underlying reinforcement learning, directly contributing to both aversive and appetitive reinforcement learning (Averbeck & Costa, 2017). Taken together, these results suggest that the amygdala's domain of specificity extends beyond fear-related processes, thereby challenging the view that this brain structure mostly operates as a fear module (Öhman, 2005; Öhman & Mineka, 2001). Instead, the amygdala's computational profile seems to more closely correspond to that of a relevance detector, subserving the rapid detection of the stimulus' relevance to the organism's concerns (Cunningham & Brosch, 2012; Pessoa & Adolphs, 2010; Sander et al., 2003, 2018).

Relevance detection model of emotional learning

Based on appraisal theories, we here suggest an alternative model to the preparedness and fear module theories, which holds that preferential emotional learning in humans is determined by a general mechanism of relevance detection as opposed to a threat-specific mechanism (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). According to this model, preferential emotional learning is not selective to threat-relevant stimuli but extends to all stimuli that are relevant to the organism's concerns (Stussi, Pourtois, et al., 2018). From this perspective, evolutionarily threat-relevant stimuli are preferentially conditioned to threat not because they have been associated with fear across evolution, but because they are likely to be detected and automatically appraised as highly relevant to the organism's survival and well-being. This alternative model thereby allows for incorporating and reinterpreting existing evidence on preferential Pavlovian aversive conditioning to threat-relevant stimuli from evolutionary origin in the Pavlovian conditioning literature. It also aligns with studies reporting the existence of learning biases to evolutionarily novel threat-relevant stimuli (e.g., Flykt et al., 2007; Hugdahl & Johnsen, 1989), as well as evidence suggesting that preferential Pavlovian aversive conditioning to social threat-relevant stimuli may rely on sociocultural rather than genetic factors alone (Mallan et al., 2013). Although the relevance detection model is similar to the expectancy bias model in that regard, these accounts differ in terms of the hypothesized psychological mechanisms underlying preferential emotional learning. Whereas the relevance

detection model accommodates the role of cognitive biases in preferential emotional learning, it states that the key determinant thereof corresponds to the stimulus' affective relevance to the organism, which is neither limited to nor necessarily dependent on expectancy biases. By suggesting that preferential emotional learning is not specific to threat-related stimuli, the relevance detection approach further contrasts with the expectancy bias model, according to which preferential associations selectively occur to threatening stimuli due to a putative expectancy bias primarily resulting from the stimulus' appraised dangerousness (Davey, 1995). Of note, the relevance detection account additionally posits that the occurrence of preferential emotional learning to affectively relevant stimuli primarily relies on an associative learning process, which facilitates the preferential associations of these stimuli with aversive and appetitive outcomes, rather than mere preexisting reponse tendencies, hence departing from the selective sensitization hypothesis (e.g., Lovibond et al., 1993).

Critically, the relevance detection model also generates new testable hypotheses. At the heart of this model lies the core prediction that stimuli that are detected as relevant to the organism's concerns benefit from preferential emotional learning independently of their valence and evolutionary status per se. More specifically, the model predicts that stimuli detected as highly relevant are more readily and persistently associated with aversive and appetitive events during Pavlovian conditioning, as reflected by the faster acquisition of a CR and enhanced resistance to extinction of that CR, respectively.

On the basis of this prediction, five main assumptions can be derived. First, stimuli varying along their relevance to the organism should be differentially conditioned, with stimuli with higher relevance inducing faster and more persistent emotional learning than stimuli with lower relevance. Second and importantly, the relevance detection model suggests the existence of a general mechanism underlying preferential emotional learning in humans that is shared across negative and positive emotional stimuli. This notably implies – even if it might appear somewhat counterintuitive at first sight – that positive stimuli with high relevance to the organism should be likewise readily and persistently associated with an aversive outcome, as is the case for threat-relevant stimuli. Evidently, this prediction starkly contrasts with the preparedness and fear module theories, according to which only threat-relevant stimuli are preferentially conditioned to threat, the CR to positive stimuli being similarly, or even more swiftly, extinguished than the CR to neutral stimuli (Öhman & Dimberg, 1978; Öhman & Mineka, 2001). Third, the relevance detection hypothesis posits a shared mechanism of emotional learning not only across negative and positive emotional stimuli, but also across

aversive and appetitive contingencies. Affectively relevant stimuli are thus assumed to benefit from preferential emotional learning not only in aversive contexts, but also in appetitive ones. Fourth, preferential emotional learning is considered to expand to affectively relevant stimuli to the organism beyond purely biological and evolutionary considerations. Fifth, the relevance detection model further asserts that emotional learning is largely affected by individual differences given that relevance detection is both individual- and situation-specific. Indeed, the process of relevance detection is inextricably tied to the organism's concerns, the salience and priority of which may rapidly and flexibly change according to environmental contingencies and which are likely to vary substantially across individuals (Cunningham & Brosch, 2012; Frijda, 1986; Sander et al., 2005). Correspondingly, the same stimulus may potentially produce preferential emotional learning for a given individual, but not for another one, in the event these individuals differ as a function of their current concerns, and hence the way in which they appraise the relevance of the stimulus at stake.

In a first attempt to explore relevance detection as a general mechanism underlying emotional learning in humans, we (Stussi et al., 2015) tested whether threat-related stimuli with a high level of relevance to the organism could be preferentially conditioned to threat relative to threat-related stimuli with a lower level of relevance. Specifically, we examined the impact of self-relevance on Pavlovian aversive conditioning by manipulating the interaction between emotion and gaze direction in facial expressions of anger and fear (see, e.g., Sander, Grandjean, Kaiser, Wehrle, & Scherer, 2007; Sander et al., 2003). According to appraisal theories of emotion, the processing of gaze direction modulates the self-relevance appraisal of a facial expression (Sander et al., 2003, 2007). In particular, appraisal theories posit that angry faces are more self-relevant with direct than averted gaze because they signal danger of being attacked, whereas fearful faces are more self-relevant with averted than direct gaze as they signal a danger in the proximal environment (Sander et al., 2003). Congruent with these predictions, results showed (a) a faster acquisition of a CR to angry faces with direct compared with averted gaze, and (b) a greater resistance to extinction of the CR to fearful faces with averted relative to direct gaze, which indicates that threat-related stimuli higher in self-relevance can induce faster and more persistent Pavlovian aversive conditioning than threat-related stimuli with lower self-relevance (Stussi et al., 2015). Whereas these findings provided initial and preliminary evidence that relevance detection may represent a general mechanism determining preferential emotional learning in humans, and suggested that the relevance detection framework provides a credible alternative to the biological preparedness models, the

role of relevance detection in emotional learning remained to be extended to stimuli with no inherent threat value to confirm some of its main predictions. In the present thesis, we thus aim to establish whether relevance detection constitutes a general mechanism underlying preferential emotional learning in humans.

2.4. THESIS OBJECTIVES

According to the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, living organisms have been biologically predisposed to selectively learn to preferentially associate stimuli that have threatened survival across evolution with aversive events. In contrast, we here seek to challenge this dominant view by suggesting that enhanced or preferential emotional learning is driven by a general mechanism of relevance detection that is not specific to threat (Stussi, Pourtois, et al., 2018). Accordingly, the purpose of this thesis is to determine whether relevance detection constitutes a general mechanism underlying preferential emotional learning in humans. More precisely, we aim to investigate the role of relevance detection in emotional learning by systematically testing the theoretical prediction deriving from appraisal theories of emotion (e.g., Sander et al., 2003, 2005, 2018) and stating that stimuli that are detected as highly relevant to the organism's concerns can benefit from preferential (e.g., faster and/or more persistent) Pavlovian conditioning, independently of their valence and evolutionary status per se.

To this end, the present thesis has five objectives (see Table 2.2). The first and primary objective is to assess whether, similar to threat-relevant stimuli, positive stimuli with biological relevance to the organism are likewise preferentially associated with a naturally aversive event during Pavlovian aversive conditioning. The second objective is to characterize at the computational level the influence of the stimulus' affective relevance on Pavlovian aversive conditioning. Two studies were conducted to achieve these objectives. In Study 1, we investigated in a series of three experiments whether both angry faces or snake images and baby faces or erotic images are more readily and persistently associated with an aversive outcome (electric stimulation) than neutral faces or colored squares, respectively, in Pavlovian aversive conditioning. In Study 2, we assessed whether, like angry faces, preferential Pavlovian aversive conditioning may be observed to happy compared with happy faces using a larger sample size than commonly used in the extant literature. In both studies, we additionally performed computational analyses using simple reinforcement learning models to characterize the impact of the stimulus' affective relevance on Pavlovian aversive learning.

The third objective of the thesis consists in examining whether preferential Pavlovian aversive conditioning extends to stimuli that are detected as relevant to the organism's concerns beyond biological and evolutionary considerations. This objective is addressed in Study 3, in

which we tested whether initially neutral stimuli without any inherent biological threat value but that acquired goal-relevance for participants could be preferentially conditioned to threat.

As an important assumption of the relevance detection model is that preferential emotional learning varies depending on individual differences in the organism's concerns, the fourth thesis objective is to investigate the role of inter-individual differences in the organism's concerns in Pavlovian aversive conditioning. To do so, in Study 2, we examined whether enhanced Pavlovian aversive conditioning to happy faces was modulated by inter-individual differences in extraversion and in happy faces' affective evaluation. In Study 3, we assessed whether inter-individual differences in achievement motivation exerted a modulatory influence on Pavlovian aversive conditioning to goal-relevant versus goal-irrelevant stimuli.

Finally, although the present thesis primarily focuses on Pavlovian aversive conditioning, the relevance detection model critically suggests that the involvement of a relevance detection mechanism is not restricted to aversive contingencies, but expands to appetitive contingencies as well. However, Pavlovian appetitive conditioning processes have been rarely investigated systematically in humans compared with Pavlovian aversive conditioning processes (e.g., Martin-Soelch et al., 2007), possibly due to a lack of existing appropriate psychophysiological measures commonly used to detect physiological changes induced by appetitive conditioning (Stussi, Delplanque, et al., 2018). For these reasons, the final objective of this thesis is methodologically oriented and aims at testing and validating a novel psychophysiological measure of human Pavlovian appetitive conditioning that could be used in further research to investigate whether the role of relevance detection can also generalize to appetitive, and not only aversive, contingencies. In light of the evidence suggesting that the postauricular reflex may constitute a reliable index of appetitive processing (Benning, Patrick, & Lang, 2004; Sandt, Sloan, & Johnson, 2009), we therefore examined in Study 4 whether the postauricular reflex could provide a valid and sensitive physiological indicator of Pavlovian appetitive conditioning in humans.

Table 2.2

Overview of the thesis objectives and the studies in which they are addressed

	Objective	Study #
Objective 1	Examining whether, similar to threat-relevant stimuli, positive stimuli with biological relevance are preferentially associated with an aversive event during Pavlovian aversive conditioning.	1, 2
Objective 2	Characterizing the influence of the stimulus' affective relevance on Pavlovian aversive conditioning	1, 2
Objective 3	Assessing whether preferential Pavlovian aversive conditioning extends to stimuli detected as relevant to the organism's concerns beyond biological and evolutionary considerations.	3
Objective 4	Investigating the role of inter-individual differences in the organism's concerns in preferential Pavlovian aversive conditioning.	2, 3
Objective 5	Testing and validating the postauricular reflex as a new psychophysiological indicator of Pavlovian appetitive conditioning in humans.	4

3. EMPIRICAL PART

3.1. STUDY 1:

ENHANCED PAVLOVIAN AVERSIVE CONDITIONING TO POSITIVE EMOTIONAL STIMULI[2]

Abstract

Pavlovian aversive conditioning is an evolutionarily well-conserved adaptation enabling organisms to learn to associate environmental stimuli with biologically aversive events. However, mechanisms underlying preferential (or enhanced) Pavlovian aversive conditioning remain unclear. Previous research has suggested that only specific stimuli that have threatened survival across evolution (e.g., snakes and angry faces) are preferentially conditioned to threat. Here, we challenge this view by showing that positive stimuli with biological relevance (baby faces and erotic stimuli) are likewise readily associated with an aversive event (electric stimulation) during Pavlovian aversive conditioning, thereby reflecting a learning bias to these stimuli. Across three experiments, our results reveal an enhanced persistence of the conditioned response to both threat-relevant and positive relevant stimuli compared with the conditioned response to neutral stimuli. These findings support the existence of a general mechanism underlying preferential Pavlovian aversive conditioning that is shared across negative and positive stimuli with high relevance to the organism, and provide new insights into the basic mechanisms underlying emotional learning in humans.

[2] Reprint of: Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147*, 905-923. https://doi.org/10.1037/xge0000424. The data reported in this article are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/2FYHQ).

3.1.1. Introduction

In Pavlovian conditioning, a conditioned stimulus acquires a predictive and emotional value through a single or repeated contingent pairing with a biologically potent stimulus. This learning process represents a fundamental evolutionarily well-conserved adaptation enabling organisms to predict and detect stimuli in the environment, and shape appropriate responses to them. Pavlovian conditioning has substantially contributed to our understanding of the psychological and neurobiological underpinnings of learning, memory, and emotion (e.g., Büchel, Morris, Dolan, & Friston, 1998; LaBar & Cabeza, 2006; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; LeDoux, 2000, 2012, 2014; Phelps, Delgado, Nearing, & LeDoux, 2004; Phelps & LeDoux, 2005; Schiller et al., 2010). Research on Pavlovian conditioning has essentially focused on unveiling the general principles of learning (Pavlov, 1927), delineating in particular the central role of prediction error (i.e., the discrepancy between the predicted and the actual outcome) and stimulus' associability (i.e., the degree to which the stimulus reliably predicts and easily enters into association with the outcome) in associative learning (see, e.g., Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla & Wagner, 1972). However, this line of research has generally omitted to consider the relative importance of the stimuli at stake for the organism. Apart from this trend, preparedness theory (Seligman, 1970, 1971) posits that certain classes of "evolutionarily prepared" threat stimuli are preferentially associated with aversive events based on biological predispositions shaped by evolution. Consistent with this view, a series of empirical studies have shown that evolutionary threat-relevant stimuli – such as snakes, angry faces, or outgroup faces – are more readily associated with an aversive outcome than threat-irrelevant stimuli – such as flowers, happy faces, or ingroup faces (e.g., Öhman & Dimberg, 1978; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001; Olsson, Ebert, Banaji, & Phelps, 2005; but see Mallan, Lipp, & Cochrane, 2013, for a review of evidence showing that threat conditioned to social threat-relevant stimuli is more malleable than threat conditioned to animal threat-relevant stimuli). Extending preparedness theory, Öhman and Mineka (2001) proposed the existence of an evolved fear module centered on the amygdala in the human brain dedicated to processing threat-relevant stimuli from phylogenetic origin, thus subserving the preferential processing of, and the learning bias to, evolutionarily prepared threat stimuli.

In contrast, we suggest that preferential emotional learning is not specific to threat-related stimuli but extends to all stimuli that are relevant to the organism's concerns (Frijda, 1988). This alternative model holds that such preferential learning is driven by a general

mechanism of relevance detection that is not specific to threat. Relevance detection is conceptualized as a rapid process, which enables the organism to detect and continuously appraise stimuli as a function of their affective relevance in relation to the organism's concerns (Pool, Brosch, Delplanque, & Sander, 2016; Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005). A stimulus is therefore detected and appraised as relevant if "it increases the probability of satisfaction or dissatisfaction toward a major concern of the individual" (Sander, 2013, p. 22). Concerns refer to affective representations of psychological and physiological motives, needs, goals, and values that are of major importance to the organism (Frijda, 1988; Pool, Brosch, et al., 2016). According to this model, phylogenetically threat-relevant stimuli lead to preferential processing and learning because they are highly relevant to the organism's survival. More specifically, the relevance detection hypothesis predicts that stimuli detected as relevant to the organism benefit from enhanced processing (Brosch, Sander, Pourtois, & Scherer, 2008; Pool, Brosch, et al., 2016) and preferential learning regardless of their valence. If the organism does preferentially learn associations involving highly relevant stimuli irrespective of their valence, this implies – even if it might seem counterintuitive – that positive stimuli with high relevance to the organism should be likewise readily associated with an aversive outcome, as is the case for threat-relevant stimuli.

Here, we therefore assessed whether positive relevant stimuli are readily associated with a biologically significant stimulus in Pavlovian aversive conditioning, thus reflecting a learning bias. Such learning bias can be characterized by a faster acquisition of a conditioned response, the acquisition of a larger conditioned response, and/or enhanced resistance to extinction of that conditioned response (Öhman & Mineka, 2001). Although all of these different indicators are considered as inherently valid, preferential emotional learning has been most consistently evidenced in humans as an enhanced persistence of the learned threat response to threat-relevant stimuli, whereas the learned threat response to threat-irrelevant stimuli generally extinguishes rapidly (Öhman & Mineka, 2001). According to preparedness and fear module theories, evolutionarily prepared threat-relevant – but not positive relevant – stimuli are readily associated with an aversive event. These theories would therefore imply that a conditioned response to positive relevant stimuli should hence be similarly, or even more quickly, extinguished than a conditioned response to neutral stimuli (Öhman & Dimberg, 1978; Öhman & Mineka, 2001). Conversely and congruently with the predictions of the relevance detection model, we predicted that the conditioned response to both threat-relevant and positive

relevant stimuli would be more persistent than the conditioned response to neutral stimuli with less relevance.

To test this competing hypothesis, we conducted three experiments examining whether, similar to threat-relevant stimuli, positive stimuli with biological relevance to the organism likewise induce a learning bias during Pavlovian aversive conditioning. In each experiment, we manipulated the conditioned stimuli's valence in a differential aversive conditioning paradigm by using three distinct conditioned stimulus categories: negative biologically relevant stimuli (angry faces in Experiments 1 and 2, and snakes in Experiment 3), positive biologically relevant stimuli (baby faces in Experiments 1 and 2, and erotic stimuli in Experiment 3), and neutral, less relevant stimuli (neutral faces in Experiments 1 and 2, and neutral colored squares in Experiment 3). This set of experiments thereby is key in order to test the hypothesis that preferential emotional learning is driven by a relevance detection mechanism, without being selective to negative threatening stimuli.

3.1.2. Experiments 1 and 2

In Experiments 1 and 2, we investigated whether angry faces and baby faces are preferentially conditioned to threat relative to neutral faces. Experiment 2 consisted of a direct replication of Experiment 1 with the aim of establishing the observed effects' reproducibility and robustness within an even more highly powered experiment. Baby faces were selected as positive relevant conditioned stimuli (CSs) because they represent a prototypical instance of stimuli being positive and highly biologically relevant for the survival of the species (Brosch et al., 2008; Kringelbach, Stark, Alexander, Bornstein, & Stein, 2016; Pool, Brosch, et al., 2016; see also Lorenz, 1943). In agreement with this view, baby faces have been shown to elicit positive evaluations (e.g., Brosch, Sander, & Scherer, 2007), to be readily prioritized for access to attentional resources (Brosch et al., 2007, 2008; Kringelbach et al., 2016; Pool, Brosch, et al., 2016), and to hold high motivational salience and a high reward value (Parsons, Young, Kumari, Stein, & Kringelbach, 2011), all of these characteristics serving as evolutionarily adaptive traits for promoting caregiving behaviors in adults and ultimately infant survival (Kringelbach et al., 2016; Lorenz, 1943). In both experiments, the differential aversive conditioning procedure comprised three contiguous phases, following standard methodology (see Lonsdorf et al., 2017). During the initial habituation phase, all CSs were presented without being reinforced. In the subsequent acquisition phase, one stimulus (reinforced stimulus [CS+])

from each CS category was systematically paired with a mild electric stimulation (unconditioned stimulus [US]) using a partial reinforcement schedule, whereas the other stimulus (unreinforced stimulus [CS-]) from each category was never associated with the electric stimulation. During the extinction phase that followed, no electric stimulation was delivered. Skin conductance responses (SCRs) were measured during all the phases. The conditioned response (CR) was operationalized as the differential SCR to the CS+ minus CS- from the same CS category (see, e.g., Olsson et al., 2005) and used as an index of learning. Our prediction was that the CR to both angry faces and baby faces would be more resistant to extinction than the CR to neutral faces.

3.1.2.1. Method

Participants

In Experiment 1, 52 participants were recruited at the University of Geneva. They provided informed consent prior to the start of the experiment, which was approved by the Faculty of Psychology and Educational Sciences Ethics committee at the University of Geneva, and they received either partial course credit or monetary compensation (20 Swiss francs) for their participation. Twelve participants were excluded from the analyses due to technical problems (n = 8), for displaying virtually no SCRs (n = 2), or for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (n = 2). These exclusion criteria are commonly applied in the contemporary human conditioning literature (e.g., Olsson et al., 2005; Olsson & Phelps, 2004; Phelps et al., 2004; Stussi, Brosch, & Sander, 2015) and were determined prior to data collection. The final sample comprised 40 participants (31 women, 9 men), aged between 18 and 52 years old (mean age = 23.85 ± 6.26 years). The sample size was determined based on a power analysis conducted with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). The analysis revealed that a total sample of 34 participants would be required to obtain a power of 80% to detect a moderate effect (d = 0.5) as reported in a previous study (Stussi et al., 2015). For counterbalancing purposes, we aimed to recruit a sample of 40 participants exhibiting differential conditioning to at least one of the three CS categories and stopped collecting data when we ascertained that the required number of participants had been reached.

In Experiment 2, 88 undergraduate psychology students from the University of Geneva were tested. None of them took part in Experiment 1. They provided informed consent prior to

the start of the experiment, which was approved by the Faculty of Psychology and Educational Sciences Ethics committee at the University of Geneva, and received partial course credit for their participation. Twenty-eight participants were excluded from the analyses due to technical problems (n = 7), for displaying virtually no SCRs (n = 8), or for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (n = 13). The final sample consisted of 60 participants (46 women, 14 men), aged between 19 and 50 years old (mean age = 23.03 ± 6.25 years). The sample size was determined based on a power analysis, which indicated that at least 54 participants would be required to achieve a power of 95% to detect a moderate effect (d = 0.5). We therefore aimed to recruit a sample of 60 participants who were conditioned to at least one of the three CS categories and stopped data collection once this sample had been reached.

Stimuli and apparatus

The CSs consisted of six different (male) faces divided into three categories: two adult faces with an angry expression, two adult faces with a neutral expression, and two baby faces. The four adult faces were taken from the Radboud Faces Database (model numbers 23 and 46 for the angry faces, and model numbers 15 and 25 for the neutral faces; Langner et al., 2010). The baby faces were selected from a set of infant faces used in previous studies (Coppin et al., 2014; Van Duuren, Kendell-Scott, & Stark, 2003). The selected faces were cut out from their original background and placed on a solid, gray background. All stimulus images were grayscale-transformed. Quantitative analyses (see Delplanque, N'diaye, Scherer, & Grandjean, 2007) confirmed that the angry, neutral, and baby stimulus images did not differ statistically in terms of luminance, apparent contrast, or mean energy in spatial-frequency bands. Each face served both as a CS+ and a CS-, counterbalanced across participants. An independent rating study (N = 63; see 3.1.5. Supplementary materials) in which the stimuli used in Experiments 1 and 2 were evaluated on a visual analog scale (VAS) ranging from 0 (*very unpleasant*) to 100 (*very pleasant*) substantiated that the angry faces were evaluated as negative (M = 30.17, SE = 2.07), the neutral faces as neutral (M = 50.71, SE = 1.53), and the baby faces as positive (M = 72.12, SE = 2.08). In Experiment 1, the US consisted of a mild electric stimulation (200-ms duration, 50 pulses/s) delivered to the participants' right wrist through a Grass SD9 stimulator (Grass Medical Instruments, West Warwick, RI) charged by a stabilized current. In Experiment 2, the US was a mild electric stimulation (10-ms duration) delivered to the participants' right wrist through a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA).

In Experiment 1, the CR was assessed through SCR measured with two pre-gelled disposable Ag-AgCl electrodes (11-mm contact diameter). In Experiment 2, the CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. In both experiments, the electrodes were attached to the distal phalanges of the second and third digits of the participants' left hand. The SCR data was continuously recorded with a sampling rate of 1000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). SCR was analyzed offline with AcqKnowledge software (version 4.2 in Experiment 1, and version 4.4 in Experiment 2; BIOPAC Systems Inc. Goleta, CA).

Procedure

Before conditioning, a work-up procedure was conducted to individually set the stimulation intensity (M = 36.75 V, SE = 1.27 in Experiment 1, and M = 34.75 V, SE = 0.98 in Experiment 2) to a level reported as "uncomfortable, but not painful" by the participant (e.g., Lonsdorf et al., 2017; Olsson et al., 2005). The initial habituation phase of the differential aversive conditioning procedure comprised two unreinforced presentations of each of the six CSs. During the acquisition phase, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Five of the seven presentations of each CS+ coterminated with the US delivery, whereas the presentations of each CS- were never paired with the US. We used a partial reinforcement schedule to potentiate the CR resistance to extinction, with the aim of optimizing the investigation of the differences in the persistence of learned emotional responses between the three CS categories used. The final extinction phase consisted of six unreinforced presentations of each CS. During all the conditioning phases, the CSs were presented for 6 s with an intertrial interval ranging from 12 to 15 s. The CSs' order of presentation was pseudorandomized into eight different orders to systematically counterbalance the associations between the face stimuli and CS type (CS+ vs. CS-) across the three CS categories (anger vs. baby vs. neutral).

After the extinction phase, participants completed subjective ratings of CS-US contingency and CS liking as manipulation checks in order to assess their awareness of the reinforcement contingencies and the CSs' pleasantness, respectively. In this procedure, the CSs were presented again, accompanied by a VAS. For the CS-US contingency ratings, participants were asked to rate to what extent the CS was predictive of the delivery of an electric stimulation, the VAS ranging from 0 (*never*) to 100 (*always*). For the CS liking ratings, participants were asked to rate to what extent the CS was unpleasant or pleasant, the VAS

ranging from 0 (*very unpleasant*) to 100 (*very pleasant*). The order of the CS presentations and the questions was randomized across participants.

Response definition

SCR was measured for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5 to 4.5 s temporal window following CS onset. The minimal response criterion was 0.02 μS. Responses below this criterion were scored as '0' and remained in the analyses. The SCR data was low-pass filtered (Blackman -92 dB, cutoff frequency = 1 Hz). SCRs were detected automatically with AcqKnowledge software as well as checked manually for artifacts and response detection. Trials containing artifacts influencing the coding of event-related SCRs or containing loss of SCR signal (1.78% in Experiment 1, and 0.003% in Experiment 2) were removed from the analyses. The raw SCR scores were square-root-transformed to normalize the distributions and scaled according to each participant's mean square-root-transformed unconditioned response (UR). The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5 to 4.5 s temporal window following the US delivery, and the mean UR was calculated across all USs for each participant (see 3.1.5. Supplementary materials). The habituation means included the first two presentations of each CS (see Figure 3.1.1). To examine the CR acquisition speed, the acquisition means were separated into an early (i.e., the first three presentations of each CS following the first association of the CS+ with the US; Trials 4 to 6, see Figure 3.1.1) and a late (the subsequent three presentations of each CS; Trials 7 to 9, see Figure 3.1.1) phase (see, e.g., Lonsdorf et al., 2017; Stussi et al., 2015). The first acquisition trial for each CS was omitted from the analyses because the CSs+ were predictive of the US only after their first association with the electric stimulation. The extinction means comprised the last six presentations of each CS (i.e., Trials 10 to 15, see Figure 3.1.1). The analyses of the conditioning data were performed on the CR, which was calculated by subtracting the SCR to the CS- from the SCR to the CS+ from the same CS category (e.g., Olsson et al., 2005). This procedure permits to reduce the confounding role of preexisting differences in the CS categories' emotional salience (Olsson et al., 2005) and to specifically control for learning within participant.

Statistical analyses

As it is standardly done in the human conditioning literature (see, e.g., Lonsdorf et al., 2017), the SCR data was analyzed separately for the habituation, acquisition, and extinction

phases. One-way repeated measures analyses of variance (ANOVAs) with CS category (anger vs. baby vs. neutral) as a within-participant factor were used to analyze the habituation and extinction data, whereas a two-way repeated measures ANOVA with CS category (anger vs. baby vs. neutral) and time (early vs. late) as within-participant factors was used for the acquisition data. One-sample *t*-tests were conducted to assess whether differential conditioning occurred to angry, baby, and neutral faces across the whole acquisition phase. To specifically test our a priori hypothesis, we performed a planned contrast analysis comparing the CR to both angry (contrast weight: +1) and baby (contrast weight: +1) faces versus neutral faces (contrast weight: -2) in extinction. Following this main contrast, three further contrasts were conducted to examine more closely whether the CR would be more persistent to (a) angry (contrast weight: +1) versus neutral (contrast weight: -1) faces and (b) baby (contrast weight: +1) versus neutral (contrast weight: -1) faces, and to assess the possible differences between (c) angry (contrast weight: +1) and baby (contrast weight: -1) faces. Because these contrasts were non-orthogonal, a Holm-Bonferroni sequential procedure (Holm, 1979) was applied to correct for multiple comparisons. Specifically, the alpha level of the contrast with the lowest *p* value was set as $\alpha = .05/4 = .0125$, the alpha level of the contrast with the second lowest *p* value as $\alpha = .05/3 = .0167$, the alpha level of the contrast with the second highest *p* value as $\alpha = .05/2 = .025$, and the alpha level of the contrast with the highest *p* value as $\alpha = .05$. An alpha level of $\alpha = .05$ was adopted for all the other statistical analyses performed. For each contrast, we additionally computed the Bayes factor (BF_{10}) quantifying the likelihood of the data under the alternative hypothesis relative to the likelihood of the data under the null hypothesis (see, e.g., Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009), using a Cauchy prior width of 0.5. For instance, a BF_{10} of 4 indicates that the data is four times more likely to be observed under the alternative hypothesis than under the null hypothesis. A BF_{10} larger than 3 (moderate evidence), larger than 10 (strong evidence), or larger than 30 (very strong evidence) is considered to provide evidence in favor of the alternative hypothesis relative to the null hypothesis, whereas a BF_{10} smaller than 0.333 (moderate evidence), smaller than 0.100 (strong evidence), or smaller than 0.033 (very strong evidence) is considered to provide evidence in favor of the null hypothesis over the alternative hypothesis (Jeffreys, 1961). We performed one-sided testing to test our a priori, theory-driven directional hypotheses (one-sample *t*-tests, main contrast and contrasts a and b), whereas two-sided testing was used when we did not have a directional prediction (contrast c).

The CS-US contingency and CS liking ratings were each analyzed with a two-way repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (anger vs. baby vs. neutral) as within-participant factors. Significant effects were followed up with a multiple comparison procedure using Tukey's HSD tests when applicable.

We report either partial η^2 or Hedges' g_{av} as estimates of effect size (see Lakens, 2013) and their 90% or 95% confidence interval (CI), respectively. Huynh-Feldt adjustments of degrees of freedom were applied when appropriate.

3.1.2.2. Results

Figure 3.1.1 displays the mean SCR magnitudes to angry, baby, and neutral faces throughout the habituation, acquisition, and extinction phases separately for the CS+ and the CS-. The conditioned response to angry, baby, and neutral faces during acquisition and extinction is depicted in Figure 3.1.2.

Experiment 1

Skin conductance response. In the habituation phase, no preexisting difference in differential SCRs to the CS categories was found, $F(2, 78) = 0.64$, $p = .533$, partial $\eta^2 = .016$, 90% CI [.000, .069]. Similarly, no statistical difference between the CS categories emerged during acquisition, $F(2, 78) = 0.44$, $p = .643$, partial $\eta^2 = .011$, 90% CI [.000, .057]. Moreover, the CR did not statistically differ between the early and late phases of acquisition, $F(1, 39) = 0.05$, $p = .816$, partial $\eta^2 = .001$, 90% CI [.000, .054]. No statistically significant interaction effect of CS category and time was observed, $F(2, 78) = 1.75$, $p = .180$, partial $\eta^2 = .043$, 90% CI [.000, .120], which indicates that there was no statistical difference in the speed of the CR acquisition across the CS categories. Further analyses revealed however a reliably greater SCR to the CS+ than CS- for angry, $t(39) = 2.31$, $p = .013$ (one-tailed), $g_{av} = 0.507$, 95% CI [0.061, 0.967], baby, $t(39) = 3.05$, $p = .002$ (one-tailed), $g_{av} = 0.669$, 95% CI [0.214, 1.141], and neutral faces, $t(39) = 2.61$, $p = .006$ (one-tailed), $g_{av} = 0.571$, 95% CI [0.122, 1.036], indicating successful differential conditioning to all three CS categories (see Figure 3.1.2a). Central to our hypothesis, analysis of the extinction phase showed that the CS categories differentially affected the persistence of the CR, $F(2, 78) = 4.51$, $p = .014$, partial $\eta^2 = .104$, 90% CI [.012, .204]. As predicted by the relevance detection hypothesis, the CR to both angry and baby faces was more resistant to extinction than the CR to neutral faces, $t(39) = 3.04$, $p = .002$ (one-tailed), $g_{av} = 0.598$, 95% CI [0.191, 1.021], $BF_{10} = 19.154$ (see Figure 3.1.2a). Direct comparisons

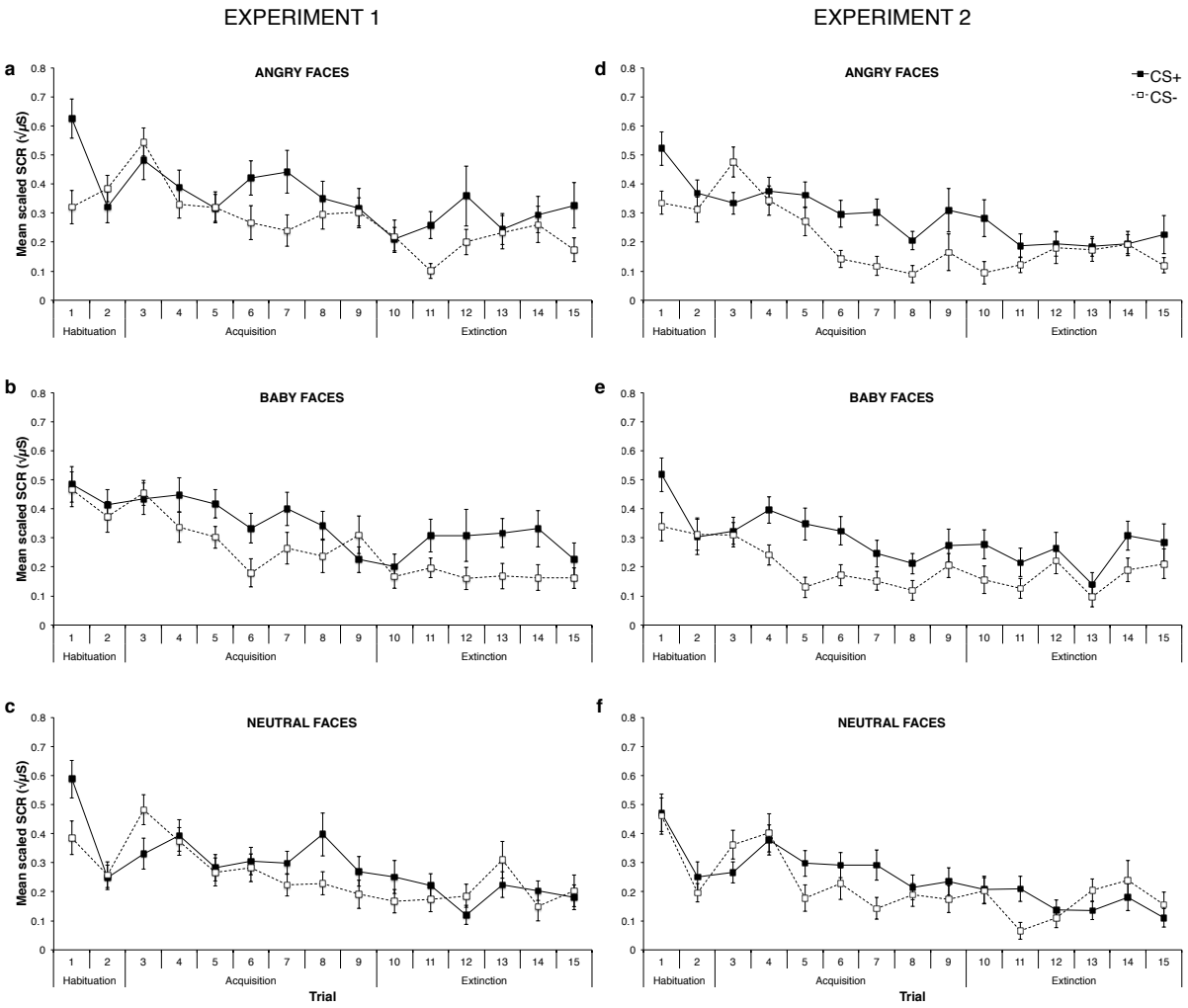

Figure 3.1.1. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials in (a-c) Experiment 1 and (d-f) Experiment 2. Mean scaled SCR to (a, d) angry faces, (b, e) baby faces, and (c, f) neutral faces. Errors bars indicate ± 1 SEM adjusted for within-participant designs (Morey, 2008).

revealed a more persistent CR to angry faces compared with neutral faces, $t(39) = 2.43$, $p = .010$ (one-tailed), $g_{av} = 0.472$, 95% CI [0.076, 0.881], $BF_{10} = 5.348$ (see Figure 3.1.2a). Importantly, the CR to baby faces was likewise more persistent than the CR to neutral faces, $t(39) = 2.73$, $p = .005$ (one-tailed), $g_{av} = 0.569$, 95% CI [0.141, 1.014], $BF_{10} = 9.679$, whereas there was no statistical difference in the resistance to extinction of the CR to angry faces compared with baby faces, $t(39) = -0.64$, $p = .524$ (two-tailed), $g_{av} = -0.132$, 95% CI [-0.545, 0.278], $BF_{10} = 0.279$ (see Figure 3.1.2a).

Subjective ratings. The CS-US contingency ratings showed that the CSs+ were deemed more likely to be associated with the US than the CSs-, $F(1, 39) = 75.25$, $p < .001$,

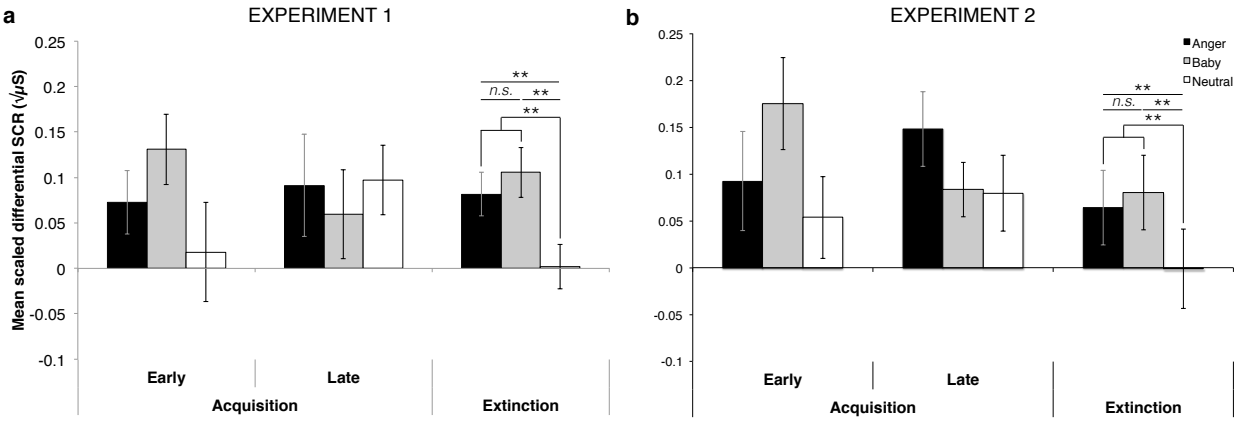

Figure 3.1.2. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (anger vs. baby vs. neutral) during (early and late) acquisition and extinction in (a) Experiment 1 and (b) Experiment 2. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (**$p < .01$, one-tailed) and *ns* indicates a statistically nonsignificant difference.

partial $\eta^2 = .659$, 90% CI [.494, .745], whereas there was no interaction between CS type and CS category, $F(2, 78) = 0.73$, $p = .485$, partial $\eta^2 = .018$, 90% CI [.000, .075]. Moreover, the CS categories differentially influenced the CS-US contingency ratings, $F(1.69, 66.00) = 7.97$, $p = .001$, partial $\eta^2 = .170$, 90% CI [.045, .291]. Follow-up analyses revealed that angry faces were rated as more likely to be predictive of the US than both baby faces ($p = .011$, $g_{av} = 0.621$, 95% CI [0.108, 1.151]) and neutral faces ($p < .001$, $g_{av} = 0.878$, 95% CI [0.399, 1.381]), whereas there was no statistical difference in the CS-US contingency ratings for baby faces relative to neutral faces ($p = .681$, $g_{av} = 0.225$, 95% CI [-0.196, 0.652]; see Figure 3.1.3a).

The CS liking ratings revealed that the CSs- were more liked than the CSs+, $F(1, 39) = 5.75$, $p = .021$, partial $\eta^2 = .128$, 90% CI [.011, .289], a significant main effect not qualified by an interaction with CS category, $F(2, 78) = 0.25$, $p = .780$, partial $\eta^2 = .006$, 90% CI [.000, .040]. The CS liking ratings were also modulated by the CS categories, $F(1.78, 69.23) = 68.92$, $p < .001$, partial $\eta^2 = .639$, 90% CI [.514, .710]. Follow-up analyses showed that baby faces were rated as more pleasant than angry faces ($p < .001$, $g_{av} = 2.505$, 95% CI [1.792, 3.302]) and neutral faces ($p < .001$, $g_{av} = 1.386$, 95% CI [0.918, 1.898]), and that neutral faces were rated as more pleasant than angry faces ($p < .001$, $g_{av} = 1.310$, 95% CI [0.796, 1.863]; see Figure 3.1.3b).

Experiment 2

Skin conductance response. During habituation, there was no statistical difference in differential SCRs to the different CS categories, $F(1.80, 105.96) = 0.76$, $p = .459$, partial $\eta^2 = .013$, 90% CI [.000, .057]. Likewise, the CR did not statistically differ across the three CS categories during the acquisition phase, $F(1.84, 108.67) = 1.72$, $p = .186$, partial $\eta^2 = .028$, 90% CI [.000, .087]. No statistically significant main effect of time was found, $F(1, 59) = 0.02$, $p = .881$, partial $\eta^2 = .0004$, 90% CI [.000, .016]. The interaction between CS category and time did not yield statistical significance either, $F(1.78, 104.89) = 1.53$, $p = .222$, partial $\eta^2 = .025$, 90% CI [.000, .083], which suggests that the CR acquisition speed did not differ across the CS categories. As in Experiment 1, one-sample *t* tests showed a greater SCR to the CS+ than CS- for angry, $t(59) = 4.80$, $p < .001$ (one-tailed), $g_{av} = 0.865$, 95% CI [0.482, 1.264], baby, $t(59) = 4.45$, $p < .001$ (one-tailed), $g_{av} = 0.801$, 95% CI [0.422, 1.195], and neutral faces, $t(59) = 1.96$, $p = .027$ (one-tailed), $g_{av} = 0.353$, 95% CI [-0.007, 0.720][3], reflecting successful differential conditioning to all three CS categories (see Figure 3.1.2b). Analysis of the extinction phase revealed that the CS categories differentially modulated the CR resistance to extinction, $F(2, 118) = 4.93$, $p = .009$, partial $\eta^2 = .077$, 90% CI [.012, .153]. Replicating results from Experiment 1, the CR to both angry and baby faces was more persistent than the CR to neutral faces, $t(59) = 3.21$, $p = .001$ (one-tailed), $g_{av} = 0.444$, 95% CI [0.162, 0.735], $BF_{10} = 31.123$ (see Figure 3.1.2b). Direct comparisons showed that the CR to angry faces was more resistant to extinction relative to neutral faces, $t(59) = 2.45$, $p = .009$ (one-tailed), $g_{av} = 0.352$, 95% CI [0.063, 0.647], $BF_{10} = 5.363$ (see Figure 3.1.2b). Critically, the CR to baby faces was also more resistant to extinction than the CR to neutral faces, $t(59) = 2.99$, $p = .002$ (one-tailed), $g_{av} = 0.451$, 95% CI [0.144, 0.765], $BF_{10} = 17.861$, whereas the CR persistence to angry faces did not statistically differ from the CR persistence to baby faces, $t(59) = -0.57$, $p = .571$ (two-tailed), $g_{av} = -0.094$, 95% CI [-0.423, 0.233], $BF_{10} = 0.225$ (see Figure 3.1.2b)[4].

Subjective ratings. The CS-US contingency ratings indicated that the CSs+ were rated as being more predictive of the US than the CSs-, $F(1, 59) = 108.15$, $p < .001$, partial $\eta^2 = .647$,

[3] The descriptively less robust aversive conditioning to neutral faces across the acquisition phase in Experiment 2 was mainly driven by the presence of an outlier (-4.77 *SD* from the mean conditioned response to neutral faces), who strongly conditioned to the neutral face CS-. The one-sample *t* test excluding this outlier indeed revealed a stronger differential conditioning to neutral faces, $t(58) = 3.26$, $p < .001$ (one-tailed), $g_{av} = 0.593$, 95% CI [0.221, 0.975]. However, since we had no a priori reason to exclude this outlier, we kept it in the analyses.

[4] Given the nature of the stimuli used, we also analyzed the SCR data of Experiments 1 and 2 including a gender factor (men vs. women) to explore potential gender differences during conditioning. In Experiment 1, this analysis revealed that men exhibited a greater conditioned response than women across CS categories during the

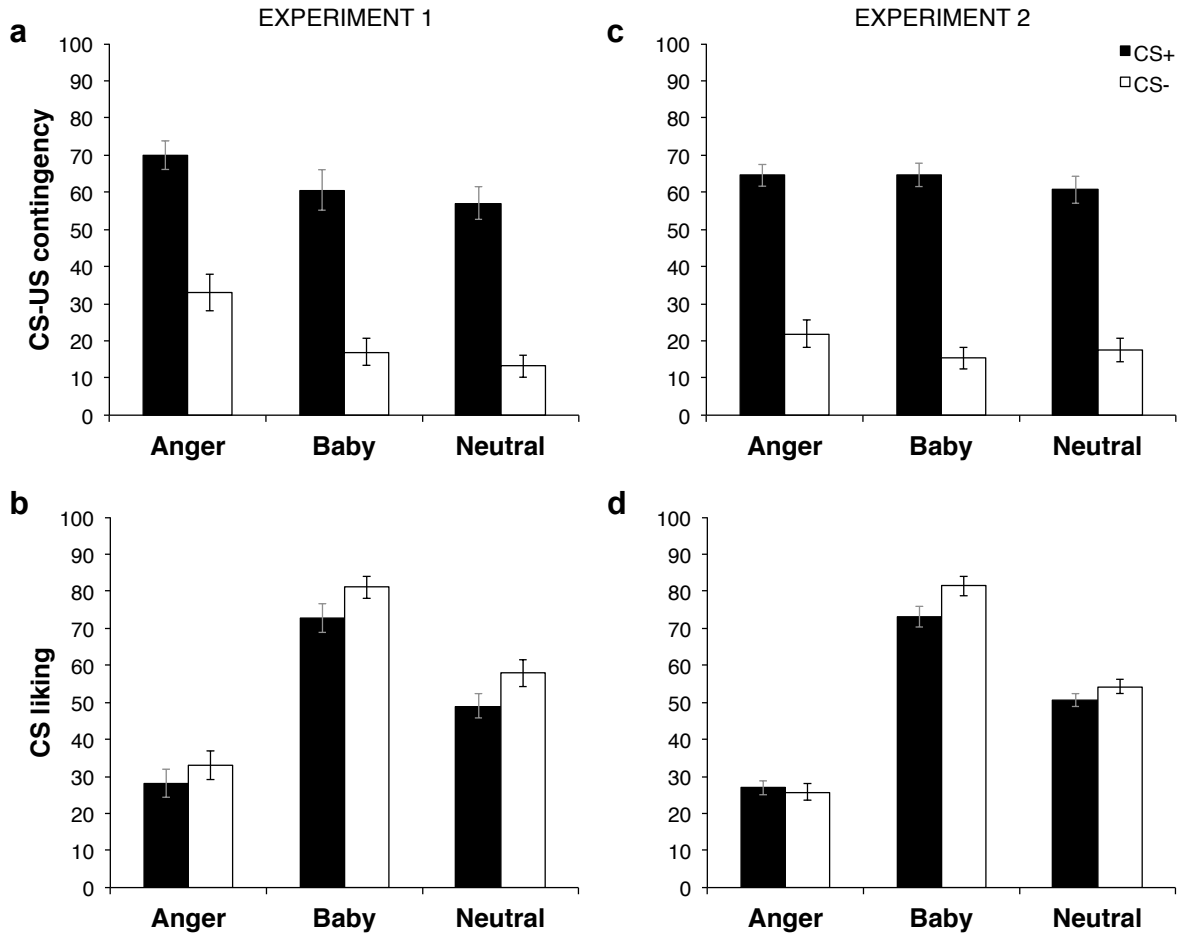

Figure 3.1.3. Mean subjective ratings as a function of the conditioned stimulus type (CS+ vs. CS-) and the conditioned stimulus category (anger vs. baby vs. neutral) in (a-b) Experiment 1 and (c-d) Experiment 2. Mean (a, c) CS-US contingency ratings and (b, d) CS liking ratings. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

90% CI [.518, .724] (see Figure 3.1.3c), whereas the interaction between CS type and CS category did not reach statistical significance, $F(2, 118) = 1.12$, $p = .331$, partial $\eta^2 = .019$, 90% CI [.000, .065]. In contrast to Experiment 1, no main effect of CS category was found, $F(2, 118) = 1.47$, $p = .235$, partial $\eta^2 = .024$, 90% CI [.000, .076].

The CS liking ratings revealed a main effect of CS type, $F(1, 59) = 4.55$, $p = .037$, partial $\eta^2 = .072$, 90% CI [.002, .191], and a main effect of CS category, $F(1.66, 98.16) = 196.77$, $p < .001$, partial $\eta^2 = .769$, 90% CI [.701, .810]. These main effects were however

habituation phase, as shown by a main effect of gender, $F(1, 38) = 5.03$, $p = .031$, partial $\eta^2 = .117$, 90% CI [.006, .278]. No other main effect or interaction effect of gender reached statistical significance (all Fs < 2.65, all ps > .07). In Experiment 2, no statistically significant main effect or interaction effect of gender was found (all Fs < 0.86, all ps > .42). These results thus suggest that no gender difference emerged among the CS categories during conditioning.

qualified by the higher-order interaction between CS type and CS category, $F(2, 118) = 3.37$, $p = .038$, partial $\eta^2 = .054$, 90% CI [.002, .122]. Follow-up analyses showed that baby faces were rated as more pleasant than angry faces (all ps < .001, 2.41 < g_{av}s < 2.96) and neutral faces (all ps < .001, 1.02 < g_{av}s < 1.80), while neutral faces were rated as more pleasant than angry faces (all ps < .001, 1.59 < g_{av}s < 1.80). Furthermore, whereas the CS- was evaluated as more pleasant than the CS+ for baby faces ($p = .021$, $g_{av} = 0.397$, 95% CI [0.068, 0.734]), there was no statistical difference in rated pleasantness between the CS- and the CS+ for angry faces ($p = .997$, $g_{av} = -0.072$, 95% CI [-0.323, 0.179]) and neutral faces ($p = .711$, $g_{av} = 0.270$, 95% CI [-0.080, 0.626]; see Figure 3.1.3d).

3.1.2.3. Discussion

In line with the relevance detection model's prediction, Experiments 1 and 2 revealed that both angry faces and baby faces produced a learning bias during Pavlovian aversive conditioning, as shown by the enhanced conditioned response persistence to angry faces and baby faces compared with neutral faces. Whereas the results for angry faces replicate previous findings (e.g., Öhman & Dimberg, 1978; Öhman & Mineka, 2001), the greater resistance to extinction of the conditioned response to baby faces expands the existing human conditioning literature, and suggests that positive stimuli with biological relevance can likewise be preferentially conditioned to threat, thereby demonstrating that preferential Pavlovian aversive conditioning is not specific to threat-related stimuli.

In contrast, we found no evidence for faster or stronger acquisition of the conditioned response to angry or baby faces relative to neutral faces. Such absence of differences across conditioned stimulus categories during acquisition is however not surprising when considering the human conditioning literature, which has generally shown a lack of experimental support for faster or stronger aversive conditioning to specific stimulus classes, such as threat-relevant stimuli (see McNally, 1987; Öhman & Mineka, 2001, for reviews). Although enhanced resistance to extinction has been frequently demonstrated to threat-relevant stimuli (Öhman & Mineka, 2001), evidence for faster or larger aversive conditioning to threat-relevant stimuli remains by comparison very scarce (Ho & Lipp, 2014; Öhman, Eriksson, & Olofsson, 1975). A potential explanation for this absence of significant effect relates to the use of a relatively high reinforcement rate whereby the CSs+ reliably predicted the US, which may have entailed rapid aversive conditioning to all the conditioned stimulus categories within a few pairings

between the CSs+ and the US, and consequently led to ceiling effects in the conditioned response acquisition readiness, thereby potentially obscuring the emergence of differences in learning patterns among the stimulus categories (see Ho & Lipp, 2014; Lissek, Pine, & Grillon, 2006).

Further, it should also be noted that the pattern of skin conductance responses in Experiment 1 was somewhat unusual at the descriptive level in comparison with what is generally observed in human aversive conditioning studies. Whereas the difference between the CS+ and the CS- is usually evident at the end of acquisition and at the onset of extinction, there seemed to be no such difference at the last acquisition trial and first extinction trial for angry faces (see Figure 3.1.1a) and baby faces (see Figure 3.1.1b). It could be speculated that this pattern may be due to the use of a within-participant design using six different CSs, instead of a between-participant design (e.g., Öhman & Dimberg, 1978; Öhman et al., 1976) or a within-participant design including only two to four conditioned stimuli (e.g., Ho & Lipp, 2014; Olsson et al., 2005), which might have entailed a stronger habituation of skin conductance responses to the CS+ than commonly observed. The subsequent reemergence of differences between the CS+ and the CS- could then have been induced by the change of contingency between the CS+ and the US, thus possibly leading to dishabituation effects. However, it remains unclear why this relative lack of evident CS+/CS- differentiation at the last acquisition trial and first extinction trial was observed for angry faces and baby faces but not for neutral faces, and why it was observed in Experiment 1, but not in Experiment 2, which suggests that it may otherwise simply reflect noise in the data.

It is also noteworthy that the observed enhanced resistance to extinction effects might be interpreted as reflecting selective sensitization, a nonassociative process, in addition to – or rather than – a conditioning process (Lovibond, Siddle, & Bond, 1993). Selective sensitization has been proposed as a putative mechanism responsible for enhanced responding to threat-relevant CSs+ during extinction, emerging as a result of the activation of preexisting response tendencies to these stimuli under certain conditions, such as threat or a state of arousal (e.g., Lovibond et al., 1993). In the present case, it could then be argued that the angry and the baby face CSs+ may have led to a greater resistance to extinction of the conditioned response than the neutral face CS+ because of their inherent potential to elicit enhanced responses in a state of arousal (i.e., induced by threat of electric stimulation). Even though we cannot completely rule out this possibility, it is unlikely that selective sensitization was the sole factor accounting for our results. Selective sensitization, as a relatively short-lived phenomenon (e.g., Lipp,

Cronin, Alhadad, & Luck, 2015), has been suggested to be insufficient to explain the long-lasting effects classically observed in human aversive conditioning studies using threat-relevant stimuli (Öhman & Mineka, 2001). Furthermore, analyses of the SCRs during the habituation phase in Experiments 1 and 2 provided no support for a selective sensitization to angry and baby faces compared with neutral faces[5], thereby suggesting that the enhanced resistance to extinction to angry and baby faces primarily resulted from an associative learning process.

In Experiments 1 and 2, subjective ratings showed that the CS+ was evaluated as being more likely to be predictive of the US delivery than the CS- across the three stimulus categories, indicating that, overall, participants were aware of the contingencies. In Experiment 1, angry faces were deemed more predictive of the US than baby and neutral faces, which might suggest that negative threat-relevant stimuli are more likely to be associated with an aversive outcome at the explicit level irrespective of the actual contingencies (Davey, 1992; Tomarken, Mineka, & Cook, 1989). However, this interpretation should be considered with caution as subjective ratings were collected exclusively after extinction but not after acquisition. Moreover, this effect did not replicate in Experiment 2, highlighting that the boundary conditions of such potential expectancy or covariation bias remain to be determined. As anticipated, baby faces were evaluated as more pleasant than neutral and angry faces, and neutral faces were rated as more pleasant than angry faces after the extinction phase in both experiments, thus reflecting an efficient manipulation of the CSs' valence. In Experiment 1, aversive conditioning had a similar effect on the CS+'s and the CS-'s rated pleasantness across the three stimulus categories; however, the CS- was evaluated as statistically significantly more pleasant than the CS+ only for baby faces in Experiment 2. Although not central to the present study's aims, these results likely stem from the fact that the electric stimulation was shorter in Experiment 2 than in Experiment 1 (10-ms vs. 200-ms duration), thus being less aversive and perceived as

[5] In order to examine whether angry and baby faces elicited enhanced sensitization in comparison with neutral faces, we performed a repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (angry vs. baby vs. neutral) as within-participant factors on SCR during the habituation phase both in Experiment 1 and 2. Although our experiments were not explicitly designed to assess selective sensitization effects, such analysis allows for a test thereof when an electric stimulation workup procedure preceding habituation is included, this workup procedure being supposedly sufficient to induce sensitization (see Lipp et al., 2015). The outcome of these analyses revealed no main effect of CS category either in Experiment 1, $F(2, 78) = 1.41$, $p = .250$, partial $\eta^2 = .035$, 90% CI [.000, .107], or in Experiment 2, $F(2, 118) = 0.77$, $p = .468$, partial $\eta^2 = .013$, 90% CI [.000, .053], thus failing to provide evidence for the occurrence of selective sensitization to angry and baby faces.

less intense[6], which might have induced less robust evaluative conditioning effects (see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010).

In sum, the occurrence of a Pavlovian learning bias to both angry faces and baby faces supports the view that preferential emotional learning is underlain by a relevance detection mechanism rather than a threat- or valence-specific mechanism, such as a fear module (Öhman & Mineka, 2001). Nonetheless, we only used a single instance of positive relevant stimuli in both experiments, thus entailing the possibility that the observed effects are selective to baby faces. The relevance detection model however predicts that positive biologically relevant stimuli induce a learning bias during Pavlovian aversive conditioning, this learning bias thereby not being confined to baby faces. Findings showing that other categories of positive relevant stimuli are preferentially conditioned to threat as well would hence provide additional empirical evidence in favor of this model. Therefore, we tested in Experiment 3 whether an enhanced Pavlovian aversive conditioning to positive relevant stimuli also occurs in response to another category of positive emotional stimuli that are relevant to the organism, namely erotic stimuli (see, e.g., Bradley, Codispoti, Cuthbert, & Lang, 2001; Panksepp, 1998; Sennwald et al., 2016).

3.1.3. Experiment 3

In Experiment 3, we aimed to replicate and extend the findings from Experiments 1 and 2 with different categories of stimuli. More specifically, we investigated whether both snakes and erotic stimuli are preferentially conditioned to threat in comparison with neutral stimuli. To this end, we used a differential aversive conditioning procedure, in which snake images, erotic images, and colored squares were presented as CSs. Erotic stimuli were selected as positive biologically relevant CSs because they are typically positive and rewarding, and hold high relevance for the species' reproduction and survival, thereby being biologically and motivationally relevant to the organism (Berridge & Kringelbach, 2015; Bradley et al., 2001; Georgiadis & Kringelbach, 2012; Panksepp, 1998; Pool, Brosch, et al., 2016; Sander et al., 2003; Schultz, 2015; Sennwald et al., 2016). Snakes were selected as negative biologically relevant CSs because they constitute the prototypical instance of negative threat-relevant

[6] A Welch's *t* test for unequal sample sizes supported this interpretation by showing that the mean square-root-transformed unconditioned response in Experiment 2 (M = 0.72, SE = 0.04) was overall smaller than in Experiment 1 (M = 1.48, SE = 0.08), $t(62.04) = 8.78$, $p < .001$, $g_s = 1.923$, 95% CI [1.451, 2.418], suggesting that the unconditioned stimulus was indeed less intense in Experiment 2 than in Experiment 1.

stimuli from phylogenetic origin that have threatened the survival of the species (see, e.g., Öhman & Mineka, 2001). The differential aversive conditioning procedure was identical to the one used in Experiments 1 and 2. After the habituation phase, during which all CSs were presented without being reinforced, the CS+ from each CS category was systematically paired with a mild electric stimulation (US) using a partial reinforcement schedule during acquisition, whereas the CS- from each category was never associated with the electric stimulation. In the subsequent extinction phase, the electric stimulation was no longer delivered. As in Experiments 1 and 2, the CR was operationalized as the differential SCR to the CS+ minus CS- from the same CS category (see, e.g., Olsson et al., 2005) and used as an index of learning. Our prediction was that the CR to both snake images and erotic images would be more resistant to extinction than the CR to neutral colored squares.

3.1.3.1. Method

Participants

Fifty-five male volunteers were recruited at the University of Geneva. They provided informed consent prior to the start of the experiment, which was approved by the Regional Research Ethics Committee in Geneva, and received monetary compensation (20 Swiss francs) for their participation. As visual sexual stimuli are primarily tailored for men, who are accordingly thought to be generally more interested in such stimuli than women (e.g., Hamann, Herman, Nolan, & Wallen, 2004; but see, e.g., Rupp & Wallen, 2008, for a discussion of the role of the stimulus materials used), only men were included in the experiment. Fifteen participants were excluded from the analyses due to technical problems (n = 2), for displaying virtually no SCRs (n = 4), for failing to acquire a CR to at least one of the three CSs predictive of the US delivery (n = 6), or for withdrawing from the experiment early (n = 3). The final sample consisted of 40 men aged between 19 and 42 years old (mean age = 24.80 ± 5.43 years). The sample size was established on the basis of a power analysis (see Experiment 1) with the aim of recruiting a sample of 40 participants exhibiting differential conditioning to at least one of the three CS categories. We stopped collecting data when the required number of participants had been reached.

Stimuli and apparatus

The CSs were selected individually for each participant among a set of 12 snake images taken from the International Affective Picture System[7] (IAPS; Lang, Bradley, & Cuthbert, 2008), 24 erotic images (12 images of nude or partially nude men and 12 images of nude or partially nude women; Sennwald et al., 2018), and 12 colored squares. Based on each participant's ratings, the two most disliked snake images, the two most liked erotic images, and the two most neutral colored squares were used as CSs. In the event that several images had identical liking ratings within a CS category, the two most arousing images were selected for the snake and erotic CS categories, respectively, whereas the two least arousing colored squares were selected for the neutral CS category. If the liking and arousal ratings were identical for several images within a CS category, the images that had been the most recently presented were chosen. The attribution of the CS+ and CS- roles to the two selected stimuli for each CS category was counterbalanced across participants. The rationale for the CSs' selection procedure was to take into account individual differences in response to erotic stimuli, the responses to such stimuli being notoriously highly variable, by adequately considering individual preferences (see Kagerer et al., 2014; Sennwald et al., 2018). This way we could ensure that the erotic stimuli were rewarding, thereby increasing the chances of these stimuli to be motivationally relevant for the participants' sexual concerns (see Sennwald et al., 2018). The selection procedure was likewise applied to the snake and neutral CSs to ensure the equal treatment of each CS category, as well as to ensure that the snake CSs were deemed negative and the neutral CSs neutral. The US was a mild electric stimulation (200-ms duration, 50 pulses/s) delivered to the participants' dominant wrist through a Grass SD9 stimulator (Grass Medical Instruments, West Warwick, RI) charged by a stabilized current.

The CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants' non-dominant hand. The SCR data was continuously recorded with a sampling rate of 1000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). SCR was analyzed offline with AcqKnowledge software (version 4.2; BIOPAC Systems Inc., Goleta, CA).

[7] IAPS numbers of the snake images used in Experiment 3: 1022, 1026, 1033, 1040, 1050, 1051, 1052, 1070, 1090, 1113, 1114, 1120.

Questionnaires

The Sexual Desire Inventory 2 (SDI-2; Spector, Carey, & Steinberg, 1996) and a questionnaire on sexual orientation were used in this experiment. The SDI-2 consists of a 14-item inventory indexing dyadic (summed score from 0 to 62) and solitary sexual desire (summed score from 0 to 23), as well as general sexual desire (summed score from 0 to 109). It was used to examine whether there might be an association between participants' sexual desire and their CR to erotic stimuli during the acquisition and extinction phases of the aversive conditioning procedure (see 3.1.5. Supplementary materials). Participants reported a mean dyadic sexual desire of 42.05 (*SE* = 1.02, range = 27-60), a mean solitary sexual desire of 10.70 (*SE* = 0.88, range = 0-23), and a mean general sexual desire of 66.08 (*SE* = 1.69, range = 47-93). The sexual orientation questionnaire was used to establish participants' sexual orientation using the Kinsey scale (Kinsey, Pomeroy, & Martin, 1948) on four different aspects of sexual orientation (i.e., sexual attraction, sexual behavior, sexual fantasies, and sexual identity).

Procedure

Prior to the experiment, participants were asked to fill out the SDI-2 and the sexual orientation questionnaire. Subsequently, they were asked to rate the 48 stimulus images according to their liking and felt arousal. The liking ratings measured how much participants liked seeing the displayed image on a VAS ranging from 0 (*not at all*) to 100 (*extremely*), whereas the arousal ratings measured how much participants felt physiologically aroused by the displayed image on a VAS ranging from 0 (*very weakly*) to 100 (*very strongly*). The stimulus images' presentation order was randomized across participants.

Once the CSs' selection procedure was completed, participants first underwent a work-up procedure in order to individually set the electric stimulation intensity (*M* = 29.75 V, *SE* = 1.16), and then the differential aversive conditioning procedure. Finally, participants completed subjective ratings of CS-US contingency and CS liking as manipulation checks to assess their awareness of the reinforcement contingencies and the CSs' pleasantness, respectively. All these procedures were identical to the ones used in Experiments 1 and 2.

Response definition

Response definition was strictly the same as in Experiments 1 and 2. Trials containing artifacts influencing the coding of event-related SCRs (0.005%) were removed from the analyses.

Statistical analyses

We performed repeated measures ANOVAs with CS type (CS+ vs. CS-) and CS category (snake vs. erotic vs. neutral) as within-participant factors on the liking and arousal ratings collected during the CSs' selection procedure to ensure (a) that there were no preexisting differences in the liking and arousal ratings between the selected CS+ and CS- within each CS category, and (b) that the selected erotic images were more liked than the selected snake images and the selected neutral colored squares, and that the selected neutral colored squares were more liked than the selected snake images. A multiple comparison procedure using Tukey's HSD tests was applied to follow up significant effects when applicable. Statistical analyses of the SCR data and the subjective ratings (i.e., CS-US contingency and CS liking ratings) were identical to the ones used in Experiments 1 and 2.

As in Experiments 1 and 2, we report either partial η^2 or Hedges' g_{av} as estimates of effect size (see Lakens, 2013) and their 90% or 95% CI, respectively. Huynh-Feldt adjustments of degrees of freedom were applied when appropriate.

3.1.3.2. Results

Figure 3.1.4 displays the mean SCR magnitudes to snake, erotic, and neutral stimuli across the habituation, acquisition, and extinction phases separately for the CS+ and the CS-. The conditioned response to snake, erotic, and neutral stimuli during acquisition and extinction is shown in Figure 3.1.5.

CS's evaluation. Table 3.3.1 shows the mean liking and arousal ratings of the CSs selected for each CS category. No main effect of CS type was found for the liking ratings of the selected CSs, $F(1, 39) = 0.73$, $p = .397$, partial $\eta^2 = .018$, 90% CI [.000, .132]. Likewise, the interaction between CS type and CS category was not statistically significant, $F(1.79, 69.77) = 0.31$, $p = .710$, partial $\eta^2 = .008$, 90% CI [.000, .053]. These results indicate that the selected CS+ and CS- did not statistically differ in terms of rated liking within each CS category. As expected, a significant main effect of CS category for the liking ratings was observed, $F(2, 78) = 284.71$, $p < .001$, partial $\eta^2 = .880$, 90% CI [.835, .902]. Follow-up analyses confirmed that the selected erotic images were more liked than the selected snake images ($p < .001$, $g_{av} = 5.769$, 95% CI [4.494, 7.260]) and the selected neutral colored squares ($p < .001$, $g_{av} = 3.560$, 95% CI [2.699, 4.548]), whereas the selected colored squares were more liked than the selected snake images ($p < .001$, $g_{av} = 1.932$, 95% CI [1.329, 2.598]).

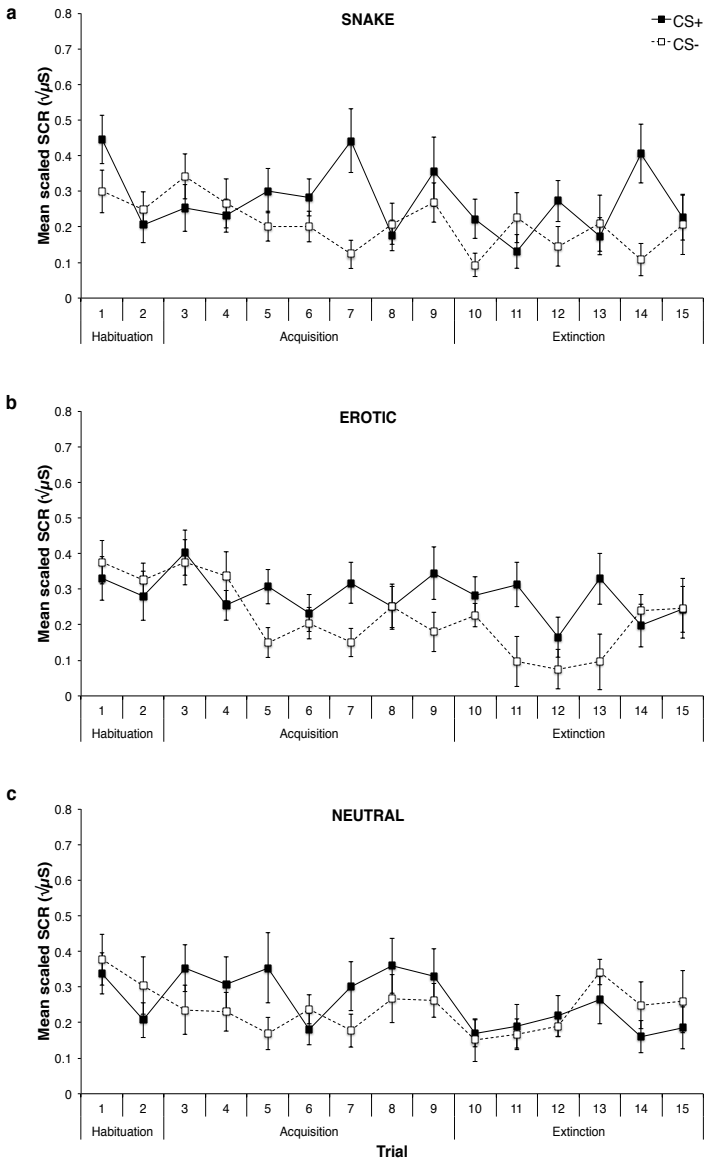

Figure 3.1.4. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials in Experiment 3. Mean scaled SCR to (a) snake stimuli, (b) erotic stimuli, and (c) neutral stimuli. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Similarly to the liking ratings, the main effect of CS type for the arousal ratings of the selected CSs was not statistically significant, $F(1, 39) = 1.03$, $p = .316$, partial $\eta^2 = .026$, 90% CI [.000, .148], and no interaction effect between CS type and CS category was found, $F(2, 78) = 0.25$, $p = .779$, partial $\eta^2 = .006$, 90% CI [.000, .040], reflecting that the selected CS+ and CS- did not statistically differ in terms of rated arousal within each CS category. As expected, the CS categories differentially influenced the arousal ratings of the selected CSs, $F(2, 78) = 75.45$, $p < .001$, partial $\eta^2 = .659$, 90% CI [.548, .723]. Follow-up tests showed that

the selected snake images were rated as more arousing than the selected neutral colored squares ($p < .001$, $g_{av} = 0.843$, 95% CI [0.410, 1.301]), and that the selected erotic images were rated as more arousing than the selected colored squares ($p < .001$, $g_{av} = 3.249$, 95% CI [2.441, 4.172]). In addition, the selected erotic images were evaluated as more arousing than the selected snake images ($p < .001$, $g_{av} = 1.523$, 95% CI [1.017, 2.076])[8].

Table 3.1.1

Mean ratings (and standard errors) of the selected conditioned stimuli (CSs) in Experiment 3.

	Snake		Erotic		Neutral	
CS type	Liking	Arousal	Liking	Arousal	Liking	Arousal
CS+	13.66 (2.48)	47.36 (5.33)	93.21 (1.75)	86.85 (2.22)	43.72 (2.70)	22.76 (4.11)
CS-	12.53 (2.58)	49.35 (5.30)	91.99 (1.87)	86.93 (2.13)	43.84 (2.56)	24.97 (3.89)

Skin conductance response. In the habituation phase, no preexisting difference in differential SCRs to the CS categories was observed, $F(2, 78) = 1.06$, $p = .353$, partial $\eta^2 = .026$, 90% CI [.000, .091]. In the acquisition phase, the CR did not statistically differ across the CS categories either, $F(2, 78) = 0.03$, $p = .967$, partial $\eta^2 = .001$, 90% CI [.000, .017], and there was no statistically significant main effect of time, $F(1, 39) = 1.41$, $p = .243$, partial $\eta^2 = .035$, 90% CI [.000, .164]. Similarly, no statistically significant interaction effect of CS category and time was found, $F(1.73, 67.50) = 0.20$, $p = .789$, partial $\eta^2 = .005$, 90% CI [.000, .043], reflecting that there was no statistical difference in the CR acquisition speed among the CS categories. Further analyses revealed that the SCR to the CS+ was greater than to the CS- for snake images, $t(39) = 2.50$, $p = .008$ (one-tailed), $g_{av} = 0.547$, 95% CI [0.099, 1.010], erotic images, $t(39) = 2.29$, $p = .014$ (one-tailed), $g_{av} = 0.502$, 95% CI [0.056, 0.962], and neutral

[8] A repeated measures ANOVA with CS type (CS+ vs. CS-) and CS category (snake vs. erotic vs. neutral) as within-participant factors on SCR during the habituation phase however showed no main effect of CS category, $F(1.54, 59.96) = 0.31$, $p = .676$, partial $\eta^2 = .008$, 90% CI [.000, .064], indicating there was no statistical difference between the different CS categories in terms of physiological arousal as measured by SCR. Similarly, no main effect of CS type, $F(1, 39) = 0.41$, $p = .528$, partial $\eta^2 = .010$, 90% CI [.000, .111], or interaction effect between CS type and CS category, $F(2, 78) = 1.06$, $p = .353$, partial $\eta^2 = .026$, 90% CI [.000, .091], were found. Of note, the absence of a statistically significant main effect of CS category also did not provide evidence for the occurrence of selective sensitization to snakes and erotic stimuli relative to neutral colored squares.

colored squares, $t(39) = 2.46$, $p = .009$ (one-tailed), $g_{av} = 0.540$, 95% CI [0.092, 1.002], indicating successful differential conditioning to all three CS categories (see Figure 3.1.5). Analysis of the extinction phase showed that the CR persistence was differentially affected by the CS categories, $F(1.73, 67.62) = 4.68$, $p = .016$, partial $\eta^2 = .107$, 90% CI [.012, .218]. As predicted by the relevance detection model, the CR to both snake and erotic images was more persistent than the CR to neutral colored squares, $t(39) = 2.62$, $p = .006$ (one-tailed), $g_{av} = 0.496$, 95% CI [0.109, 0.898], $BF_{10} = 7.777$ (see Figure 3.1.5). Pairwise comparisons revealed that the CR to snake images was more resistant to extinction than colored squares, $t(39) = 2.52$, $p = .008$ (one-tailed), $g_{av} = 0.432$, 95% CI [0.082, 0.794], $BF_{10} = 6.397$. The CR to erotic images was likewise more resistant to extinction compared with the CR to colored squares, $t(39) = 2.38$, $p = .011$ (one-tailed), $g_{av} = 0.504$, 95% CI [0.072, 0.950], $BF_{10} = 4.815$, whereas no statistical difference in CR resistance to extinction emerged between snake images and erotic images, $t(39) = -0.51$, $p = .610$ (two-tailed), $g_{av} = -0.095$, 95% CI [-0.466, 0.274], $BF_{10} = 0.261$ (see Figure 3.1.5).

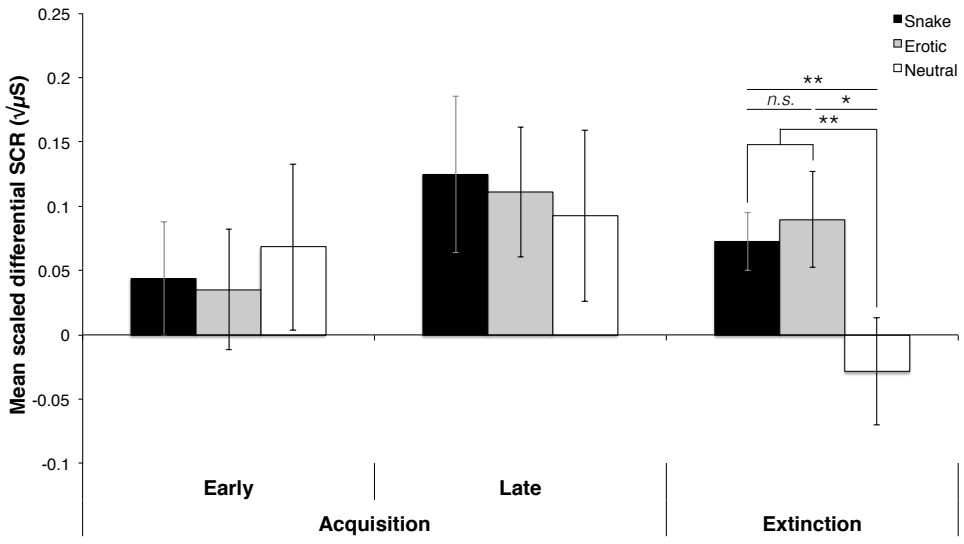

Figure 3.1.5. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (snake vs. erotic vs. neutral) during (early and late) acquisition and extinction in Experiment 3. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (**$p < .01$, *$p < .05$, one-tailed) and *ns* indicates a statistically nonsignificant difference.

Subjective ratings. The CS-US contingency ratings showed that the CSs+ were more likely to be associated with the US than the CSs-, $F(1, 39) = 26.62$, $p < .001$, partial $\eta^2 = .406$, 90% CI [.203, .547], while the interaction between CS type and CS category did not reach statistical significance, $F(2, 78) = 2.66$, $p = .076$, partial $\eta^2 = .064$, 90% CI [.000, .152]. Moreover, the CS-US contingency ratings were differentially modulated by the CS categories, $F(2, 78) = 3.55$, $p = .034$, partial $\eta^2 = .083$, 90% CI [.004, .178]. Follow-up tests indicated that erotic images were rated as being more predictive of the US compared with colored squares ($p = .038$, $g_{av} = 0.479$, 95% CI [0.055, 0.917]), but not relative to snake images ($p = .890$, $g_{av} = 0.093$, 95% CI [-0.309, 0.497]), whereas snake images were not evaluated as more predictive of the US than colored squares ($p = .109$, $g_{av} = 0.388$, 95% CI [0.037, 0.750]; see Figure 3.1.6a).

The CS liking ratings revealed that the CSs- were not deemed more pleasant than the CSs+ after the extinction phase, $F(1, 39) = 0.56$, $p = .459$, partial $\eta^2 = .014$, 90% CI [.000, .122]. Expectedly, a main effect of CS category was found, $F(2, 78) = 135.20$, $p < .001$, partial $\eta^2 = .776$, 90% CI [.697, .818]. This main effect was not qualified by an interaction with CS type, $F(2, 78) = 0.22$, $p = .801$, partial $\eta^2 = .006$, 90% CI [.000, .037]. Follow-up analyses showed that erotic images were evaluated as more pleasant than snake images ($p < .001$, $g_{av} = 3.801$, 95% CI [2.879, 4.860]) and colored squares ($p < .001$, $g_{av} = 2.654$, 95% CI [1.963, 3.438]), while colored squares were rated as more pleasant than snake images ($p = .001$, $g_{av} = 0.797$, 95% CI [0.337, 1.279]; see Figure 3.1.6b).

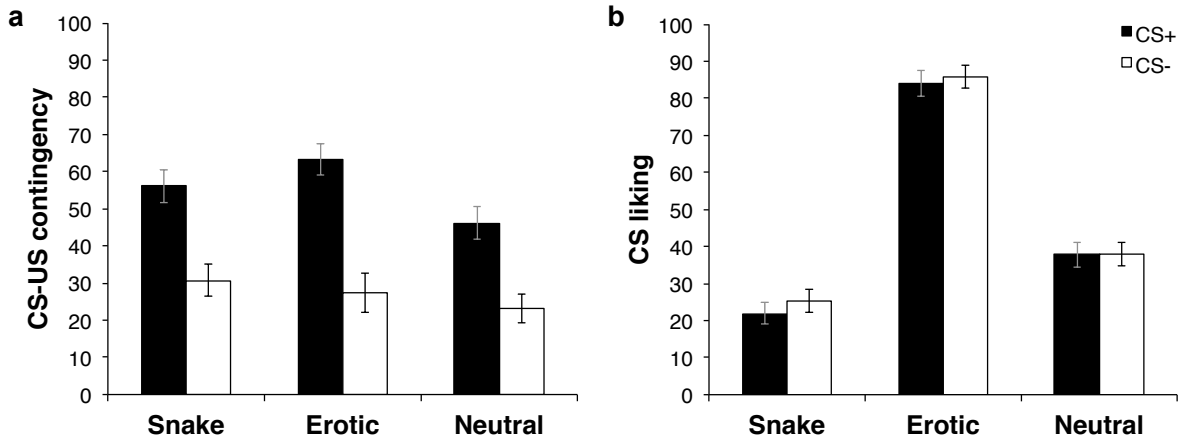

Figure 3.1.6. Mean subjective ratings as a function of the conditioned stimulus type (CS+ vs. CS-) and the conditioned stimulus category (snake vs. erotic vs. neutral) in Experiment 3. Mean (a) CS-US contingency ratings and (b) CS liking ratings. Errors bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

3.1.3.3. Discussion

Experiment 3 replicated and extended the key findings of Experiments 1 and 2 by demonstrating that, like threat-relevant stimuli, positive stimuli with biological relevance to the organism are preferentially conditioned to threat, and, in particular, that these findings generalize beyond baby faces. Results indeed showed that the conditioned response to snake images was more resistant to extinction than the conditioned response to neutral colored squares, which concurs with previous research in the human conditioning literature (e.g., Öhman et al., 1976; Öhman & Mineka, 2001). Of critical importance, the conditioned response to erotic images was likewise more resistant to extinction relative to neutral colored squares, thereby reflecting that both snake and erotic stimuli induced a learning bias during Pavlovian aversive conditioning.

Of note, previous studies by Hamm and colleagues (Hamm, Greenwald, Bradley, & Lang, 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) have also used erotic stimuli as CSs in a differential aversive conditioning procedure. Although these studies showed a greater responding in SCR to the CS+ than the CS- across the various stimulus categories used (e.g., threatening animals, mutilations, household objects, and nature scenes) during extinction, none of them seemed to suggest an enhanced resistance to extinction to erotic stimuli, thus contrasting with the current findings. Nonetheless, it is important to note that these studies did not take into account individual preferences for erotic stimuli, and thereby did not directly consider erotic stimuli's affective relevance for the individual's sexual concerns, which may potentially account for the discrepancy between their results and ours.

In line with prior reports in the human conditioning literature (see McNally, 1987; Öhman & Mineka, 2001, for reviews), we observed no reliable differences among the conditioned stimulus categories during the acquisition phase, thus providing no evidence for faster or larger acquisition of a conditioned response to snake images and erotic stimuli compared with neutral stimuli. As for Experiments 1 and 2, this absence of effect might be explained by the specifics of the experimental paradigm used here, in which the various CSs+ predicted relatively unambiguously the US, thereby possibly masking the emergence of differences in the conditioned response acquisition readiness across the conditioned stimulus categories (Ho & Lipp, 2014; Lissek et al., 2006).

Overall, the CSs' ratings during the CSs' selection procedure confirmed that the selected snake stimuli were deemed negative, the selected neutral stimuli neutral, and the

selected erotic stimuli positive. The selected erotic and snake stimuli were additionally rated as more arousing than the selected neutral stimuli, whereas the erotic stimuli were also rated as more arousing than the snake stimuli. This latter effect might have occurred because some participants may have misinterpreted the notion of physiological arousal as sexual arousal, thus entailing a possible undervaluation of the actual snake stimuli's arousal value. Importantly, there was however no statistical difference between the selected CS+ and the selected CS- within each stimulus category in the liking and arousal ratings, thereby reflecting an appropriate selection of the conditioned stimuli for each stimulus category.

Subjective ratings collected after extinction revealed that the CSs+ were evaluated as more predictive of the US than the CSs- across the three stimulus categories, indicating that, overall, participants were aware of the contingencies. Moreover, erotic stimuli were deemed more likely to be associated with the US than neutral stimuli regardless of the actual contingencies. This might suggest that expectancy (Davey, 1992) and/or covariation (Tomarken et al., 1989) biases are not selective to associations involving negative threat-relevant stimuli, but can also encompass certain associations between positive biologically relevant stimuli and aversive outcomes. However, this interpretation should be considered with caution because we collected subjective ratings only after extinction, but not after acquisition. In addition, the fact that we did not find such an effect either in Experiment 1 or 2 highlights that further research is needed to explore its determinants, along with its reproducibility and robustness. The CS liking ratings confirmed that erotic stimuli were still evaluated as more pleasant than neutral and snake stimuli after extinction, whereas neutral stimuli were still rated as more pleasant than snake stimuli. In contrast to Experiments 1 and 2 as well as previous reports in the human conditioning literature (e.g., Hamm et al., 1993; Hamm & Vaitl, 1996), no resistant-to-extinction evaluative effects were observed in this experiment. A potential explanation for this discrepancy could be that the addition of CSs' prior ratings during the CSs' selection procedure may have biased participants' postextinction ratings of the same CSs, leading to reduced evaluative conditioning effects (see Lipp & Purkis, 2006).

In brief, Experiment 3 aligns with Experiments 1 and 2 in suggesting that preferential aversive conditioning is not selective to threat-related stimuli, but extends to positive biologically relevant stimuli as well. Experiment 3 thus provides further evidence supporting the hypothesis that stimuli that are relevant to the organism's concerns benefit from preferential emotional learning independently of their valence.

3.1.4. General discussion

In the present study, we aimed at directly testing the predictions of two competing models of emotion with respect to emotional learning; more specifically, we aimed to test the appraisal-based hypothesis that preferential emotional learning is driven by a relevance detection mechanism that is not selective to threat, an hypothesis that is opposed to the fear module hypothesis according to which preferential emotional learning is driven by a fear-specific mechanism that is selective to threat. In order to do so, we investigated whether, similar to threat-relevant stimuli, positive stimuli that are biologically relevant to the organism are likewise preferentially conditioned to threat. In three experiments, we used a differential aversive conditioning paradigm, in which negative biologically relevant stimuli (angry faces, snakes), positive biologically relevant stimuli (baby faces, erotic stimuli), and neutral, less relevant stimuli (neutral faces, colored squares) were used as conditioned stimuli. Taken together, results demonstrate a preferential Pavlovian aversive conditioning to both threat-relevant and positive relevant stimuli.

The enhanced persistence of the learned threat response to threat-relevant stimuli compared with neutral stimuli replicates the basic finding of preferential emotional learning to threat-relevant stimuli consistently reported in the human conditioning literature (e.g., Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005; see also Mallan et al., 2013). More importantly, our findings showing an enhanced persistence of the conditioned response to positive relevant stimuli relative to neutral stimuli reflect that positive stimuli with biological relevance are likewise readily associated with a biologically significant event during Pavlovian aversive conditioning, even if this event is naturally aversive. In contradiction to the fear module theory, and somewhat counterintuitively, our hypotheses-driven findings therefore demonstrate that preferential aversive conditioning is not limited to negative stimuli carrying threatening information, but can be extended to positive stimuli that are biologically relevant to the organism. In this respect, our results concur with prior empirical findings in the field of emotional attention, which have shown that attention is not exclusively biased toward negative threatening stimuli, but also orients preferentially and quickly toward positive relevant stimuli (Brosch et al., 2008; Pool, Brosch, et al., 2016). In addition, our data also align with neurobiological evidence suggesting the existence of shared mechanisms across negative and positive valence. Indeed, the encoding and processing of negative and positive stimulus' values has been shown to rely on overlapping brain structures (e.g., Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002; Janak & Tye, 2015; Jin, Zelano, Gottfried, & Mohanty,

2015; Namburi et al., 2015; Paton, Belova, Morrison, & Salzman, 2006; Seymour, Daw, Dayan, Singer, & Dolan, 2007; Shabel & Janak, 2009) and neurotransmitter systems (e.g., M. Matsumoto & Hikosaka, 2009). However, the occurrence of a learning bias to threat-relevant and positive relevant stimuli strongly contrasts with previous research suggesting that preferential aversive conditioning is restricted to specific classes of stimuli that have provided threats to the survival of our ancestors across evolution (Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001; Olsson et al., 2005; Seligman, 1970, 1971). Our findings challenge the view that threat-relevant stimuli are readily associated with an aversive event because they have been correlated with threat through evolution, and alternatively suggest that the key factor underlying preferential emotional learning to threat-relevant stimuli in humans is their high affective relevance to the organism. Our study thereby provides strong support for the existence of a general relevance detection mechanism underlying emotional learning in humans that is common across negative and positive stimuli with biological relevance to the organism.

Nonetheless, it might be proposed that the enhanced persistence of the conditioned response to both threat-relevant and positive relevant stimuli was driven by their a priori negative and positive valence, respectively. Such an account appears nevertheless unlikely because learned threat to happy faces, which represent a typical instance of highly positive stimuli with a relatively low level of general relevance to the organism (Brosch et al., 2008; Pool, Brosch, et al., 2016) and the processing of which is likely to be sensitive to individual differences (Canli et al., 2002), has been shown to rapidly extinguish (e.g., Öhman & Dimberg, 1978; Rowles, Lipp, & Mallan, 2012).

As negative and positive biologically relevant stimuli are typically highly arousing, it could be possible that our findings were mediated by the stimuli's arousal value, the respective contributions of relevance detection and arousal to enhanced aversive conditioning being difficult to disentangle from one another (Montagrin & Sander, 2016; Pool, Brosch, et al., 2016; Sander, 2013). In fact, appraisal theories (e.g., Sander et al., 2003, 2005) posit that stimuli that are appraised as relevant to the organism's concerns also very often elicit a motivational state, which is reflected in a consequent physiological state of arousal that may be felt consciously (Pool, Brosch, et al., 2016). However, the relevance detection and arousal accounts fundamentally differ in terms of the hypothesized psychological mechanisms underlying preferential emotional learning. Whereas the arousal account suggests that the stimulus' arousal value directly drives learning bias, the relevance detection hypothesis

explicitly states that the stimulus' affective relevance to the organism's concerns determines learning bias. Accordingly, the mechanism responsible for enhanced emotional learning lies in the emotion elicitation process for the relevance detection account; by contrast, it lies in one component of the emotional response for the arousal account. Indirect evidence in favor of the relevance detection hypothesis comes from a recent meta-analysis on attentional bias for positive stimuli (Pool, Brosch, et al., 2016), which has demonstrated that, whereas both arousal and affective relevance modulated the attentional bias magnitude, only affective relevance remained a significant predictor of the magnitude of the attentional bias when the contributions of arousal and affective relevance were tested by statistically controlling their respective variances, thus implying that relevance detection is more likely to constitute the key mechanism underlying biases in emotional attention than arousal. Additional evidence challenging the arousal account can also be found in studies by Hamm and colleagues (Hamm et al., 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996), which have shown that highly arousing positive and negative stimuli, without considering their affective relevance to the organism's concerns, did not lead to enhanced resistance to extinction compared with stimuli with a lower arousal level. These results hence indicate that arousal alone might not be sufficient for triggering enhanced Pavlovian aversive conditioning, thereby suggesting that relevance detection provides a more appropriate and plausible mechanism to account for our findings.

Alternatively, it could be argued that preferential emotional learning to threat-relevant stimuli relies on a fear module on the one hand, whereas preferential emotional learning to positive relevant stimuli is triggered by another module dedicated to processing positive, appetitive, or reward-related stimuli with high relevance on the other hand. However, increasing converging evidence shows that the amygdala, which plays a fundamental role in emotional learning (e.g., Büchel et al., 1998; Janak & Tye, 2015; LaBar et al., 1998; LeDoux, 2000, 2012; Phelps & LeDoux, 2005) and was historically conceived as a fear module (Öhman & Mineka, 2001), is not specifically involved in the processing of threat-relevant stimuli, but in the processing of stimuli that are relevant to the organism (Cunningham & Brosch, 2012; Pessoa & Adolphs, 2010; Sander et al., 2003; Sergerie, Chochol, & Armony, 2008), including positive or rewarding stimuli (Gottfried, O'Doherty, & Dolan, 2003; Sergerie et al., 2008). Furthermore, the amygdala has been shown to be a core brain structure of the motivational neural circuits underlying reinforcement learning, directly contributing not only to aversive but also to appetitive reinforcement learning (Averbeck & Costa, 2017). In particular, the

amygdala is implicated in the computation of both prediction error (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013) and stimulus' associability (Boll et al., 2013; Li, Schiller, Schoenbaum, Phelps, & Daw, 2011), which are fundamental determinants of associative learning in computational models of Pavlovian conditioning (e.g., Li et al., 2011; Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla & Wagner, 1972). In light of this evidence, we argue that relevance detection constitutes a parsimonious and plausible account of the learning bias to both threat-relevant and positive relevant stimuli during Pavlovian aversive conditioning in humans.

A wider consideration of computational models of Pavlovian conditioning (e.g., Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) however raises the question as to whether the existence of a learning bias to negative and positive stimuli with biological relevance is adequately captured, and can be characterized, by such Pavlovian learning models. Given the critical role of prediction error and stimulus' associability in associative learning, it could be hypothesized that stimulus' biological relevance may bias Pavlovian conditioning by altering such learning signals. A potential computational learning mechanism whereby the influence of stimulus' biological relevance may operate is stimulus salience, which constitutes a key parameter determining the learning rate and ultimately affecting the impact of prediction error and associability in a number of computational models of conditioning (e.g., Pearce & Hall, 1980; Rescorla & Wagner, 1972).

Stimulus salience traditionally refers to a bottom-up perceptual process based on the stimulus' physical properties (see, e.g., Öhman & Mineka, 2001; Parkhurst, Law, & Niebur, 2002; Pearce & Hall, 1980). Although more salient or intense stimuli – in the sense of physical or perceptual salience – have been shown to be more easily conditioned than less salient or intense stimuli (e.g., Pearce & Hall, 1980; Rescorla, 1988a; Rescorla & Wagner, 1972), it has been demonstrated that neutral stimuli with a high perceptual salience do not produce enhanced resistance to extinction compared with neutral stimuli with a low perceptual salience (Öhman et al., 1976), thereby reflecting that physical salience alone provides an insufficient and unlikely explanation for the effects observed in our three experiments (see also McNally, 1987; Öhman & Mineka, 2001). However, stimulus salience has not solely been discussed in the literature as a mere characteristic of the stimulus, but has also been discussed in terms of motivational contingencies relating to the organism's needs and goals (see Cunningham & Brosch, 2012; Öhman & Mineka, 2001; Rescorla, 1988a). In this respect, various stimuli can be considered as motivationally salient, such as the threat-relevant and positive relevant stimuli

used in our study (see, e.g., Öhman & Mineka, 2001; Parsons et al., 2011; Schultz, 2015). It has been argued that the process of incentive salience is conceptually very closely related to the construct of relevance detection as used in appraisal theories of emotion (see Pool, Sennwald, Delplanque, Brosch, & Sander, 2016; Sennwald, Pool, & Sander, 2017). For instance, it has been suggested that the human amygdala is the key brain system involved in relevance detection (Sander et al., 2003), an idea that is conceptually very similar to the proposal that the amygdala is the key region involved in motivational salience (Cunningham & Brosch, 2012). Of course, the constructs of relevance detection and motivational salience have different conceptual historical roots, and are used in different research traditions but share a fundamental aspect underlying why a post-hoc explanation of our results in terms of motivational salience would closely mirror our a priori prediction in terms of relevance detection: Both constructs suggest that the key factor responsible for our results stems from the interaction between the stimulus and the organism's current concerns.

Critically, our findings of enhanced resistance to extinction of the learned emotional response to both threat-relevant and positive relevant stimuli are however in stark contrast with the predictions of the influential Rescorla-Wagner (Rescorla & Wagner, 1972) and Pearce-Hall (Pearce & Hall, 1980) models of Pavlovian conditioning, as well as previous empirical data from animal research (e.g., Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002). Although these models predict and account for the accelerated acquisition of the conditioned response to more salient stimuli during conditioning (e.g., Pearce & Hall, 1980; Rescorla, 1988a; Rescorla & Wagner, 1972), they also predict that, all else being equal, the conditioned response to more salient stimuli will extinguish faster than the conditioned response to less salient stimuli (see Siddle & Bond, 1988; see also Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002, for studies in rats providing either direct or indirect support for this prediction). A salience parameter as implemented in the Rescorla-Wagner and Pearce-Hall models therefore does not seem to provide a plausible computational learning mechanism that is able to adequately capture and characterize the influence of the type of stimulus' biological relevance that we investigated in our series of experiments. In line with this view, additional computational analyses of our data using simple reinforcement learning models (Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see 3.1.5. Supplementary materials) suggest that the influence of both negative and positive biologically relevant stimuli, relative to neutral stimuli with less relevance, might be specifically characterized by a lower learning rate for negative prediction error (i.e., when the expected outcome is omitted or when the outcome is

less than predicted) that biases inhibitory learning – which includes, without being limited to, extinction learning (Dunsmoor, Niv, Daw, & Phelps, 2015) – through a reduced impact of negative prediction error on associative strength, thus potentially accounting for the enhanced persistence of the conditioned response. Nonetheless, the computational mechanisms by which the influence of stimulus' affective relevance on Pavlovian conditioning operates remain yet to be better elucidated and characterized.

In conclusion, this series of three experiments suggests that relevance detection drives Pavlovian aversive conditioning in humans. Relevance detection constitutes a rapid (e.g., Grandjean & Scherer, 2008) and flexible (e.g., Moors, 2010) mechanism that enables the organism to adaptively and dynamically trigger the preferential processing and learning of stimuli that are detected as highly relevant. Importantly, the relevance detection account also allows for the accommodation and reinterpretation of existing evidence on preferential aversive conditioning to evolutionary threat stimuli, as these stimuli are a highly relevant signal for the organism. However, a relevance detection mechanism should trigger preferential emotional learning not only to biologically relevant stimuli but also to stimuli that are relevant to the organism's concerns independently of their evolutionary status per se. Primary evidence of this point still remains inconclusive. Some studies have shown a similar persistence of learned threat to threatening stimuli from both phylogenetic (i.e., snakes) and ontogenetic (i.e., pointed guns) origin (Flykt, Esteves, & Öhman, 2007; Hugdahl & Johnsen, 1989), while other studies have reported a greater persistence of learned threat to phylogenetically threat-relevant stimuli compared with ontogenetically threat-relevant stimuli (E. W. Cook, Hodes, & Lang, 1986; Hugdahl & Kärker, 1981). Further research will thus have to pinpoint whether preferential emotional learning is limited to evolutionary relevant stimuli or extends to stimuli with high relevance to the organism beyond biological and evolutionary considerations. As neural circuits underlying threat-related responses and behaviors have been shown to respond differently to actual threats posed by predators as opposed to standard aversive conditioning paradigms commonly used in laboratory settings (Mobbs & Kim, 2015), another interesting and important avenue for future research will be to investigate whether the role of relevance detection generalizes across more ethologically valid paradigms (e.g., using virtual reality) mirroring the ecological conditions under which threats and rewards typically occur in the organism's natural environment. By postulating a common mechanism of emotional learning not only across negative and positive stimuli but also across aversive and appetitive contingencies, the relevance detection approach offers a new perspective that may contribute

to a better understanding of the functioning of human emotional learning, as well as its alteration in specific disorders. Although the generality of a relevance detection mechanism remains to be determined in appetitive conditioning, our study provides new insights into the basic mechanisms underlying emotional learning in humans.

Context of the research

The present set of experiments originates from a research program that aims to investigate the links between the appraisal processes involved in emotion elicitation and the basic mechanisms underlying learning in humans. In this research program, we seek to challenge the dominant view that only threat-related stimuli induce preferential emotional learning by offering an alternative theoretical framework based on appraisal theories of emotion (e.g., Sander et al., 2003, 2005), which holds that emotional learning is driven by a process of relevance detection that is not specific to threat. Our goal is therefore to systematically test the theoretical prediction that stimuli that are detected as highly relevant to the organism's concerns benefit from enhanced Pavlovian conditioning, independently of their intrinsic valence. In this perspective, the findings reported here provide initial evidence for the existence of a relevance detection mechanism underlying emotional learning in humans, and suggest that appraisal theories may offer a promising framework to foster better insights into the understanding of human emotional learning. Ultimately, this framework might also be valuable to account for the high flexibility and large inter-individual differences typically observed in emotional learning across varying contexts and situations, as well as some impairments in this process preceding or following the onset and maintenance of specific emotional disorders. Accordingly, future research will focus on expanding the current findings with the aim of further establishing and characterizing the role of relevance detection in emotional learning.

3.1.5. Supplementary materials

Supplementary method and results

Independent rating study

Sixty-three volunteers (49 women, 14 men) aged between 18 to 48 years old (M = 27.54 ± 5.73 years) participated in an independent rating study to ensure that the angry faces used in Experiments 1 and 2 were evaluated as negative, the baby faces as positive, and the neutral faces as relatively neutral.

The independent rating study consisted of an online study using qualtrics® (https://www.qualtrics.com), in which the six different stimuli used in Experiments 1 and 2 were presented to participants, accompanied by a visual analog scale (VAS). Participants were asked to rate to what extent the face displayed onscreen was unpleasant or pleasant, the VAS ranging from 0 (*very unpleasant*) to 100 (*very pleasant*). The order of the face presentations was randomized across participants. The stimulus liking ratings were analyzed with a one-way repeated measures analysis of variance (ANOVA) with stimulus category (anger vs. baby vs. neutral) as a within-participant factor. The main effect of stimulus category was followed up with a multiple comparison procedure using Tukey's HSD tests if applicable.

Table S3.1.1 reports the mean liking ratings for each stimulus separately. The one-way repeated measures ANOVA revealed that the liking ratings were modulated by the stimulus category, $F(1.66, 102.91) = 127.54$, $p < .001$, partial $\eta^2 = .673$, 90% CI [.583, .729]. Follow-up analyses showed that participants rated the baby faces (M = 72.12, SE = 2.08) as more pleasant than both the angry faces (M = 30.17, SE = 2.07; $p < .001$, g_{av} = 2.519, 95% CI [1.958, 3.133]) and the neutral faces (M = 50.71, SE = 1.53; $p < .001$, g_{av} = 1.462, 95% CI [1.062, 1.891]), while the neutral faces were evaluated as more pleasant than the angry faces ($p < .001$, g_{av} = 1.406, 95% CI [1.031, 1.810]). Overall, the independent rating study thus confirmed that the selected angry faces were evaluated as negative, the selected baby faces as positive, and the selected neutral faces as relatively neutral.

Table S3.1.1

Mean liking ratings (and standard errors) of the stimuli used in Experiments 1 and 2 in the independent rating study.

	Angry faces		Baby faces		Neutral faces	
	Face 1	Face 2	Face 1	Face 2	Face 1	Face 2
	34.79 (2.33)	25.54 (2.38)	74.30 (2.54)	69.94 (2.09)	49.37 (1.84)	52.06 (1.81)
Source	RaFD model 23 (Langner et al., 2010)	RaFD model 46 (Langner et al., 2010)	Coppin et al. (2014); Van Duuren et al. (2003)	Coppin et al. (2014); Van Duuren et al. (2003)	RaFD model 15 (Langner et al., 2010)	RaFD model 25 (Langner et al., 2010)

Note. RaFD = Radboud Faces Database.

Unconditioned response analysis

Across the three experiments, we analyzed the unconditioned response (UR) to the unconditioned stimulus (US; i.e., electric stimulation) using two-way repeated measures ANOVAs with CS category (anger vs. baby vs. neutral in Experiments 1 and 2, snake vs. erotic vs. neutral in Experiment 3) and US trial (US trial 1 vs. US trial 2 vs. US trial 3 vs. US trial 4 vs. US trial 5) as within-participant factors in order to explore whether the CS categories differentially modulated the UR, and to investigate the UR changes across trials. Due to missing values on some trials, 33 participants could be included in the UR analysis in Experiment 1, 52 in Experiment 2, and 38 participants in Experiment 3.

In Experiment 1, the UR was not differentially influenced by the CS categories, $F(2, 64) = 1.01$, $p = .369$, partial $\eta^2 = .031$, 90% CI [.000, .106], and did not significantly change across trials, $F(2.84, 91.02) = 1.80$, $p = .155$, partial $\eta^2 = .053$, 90% CI [.000, .120] (see Figure S3.1.1a). Similarly, no statistically significant interaction between CS category and US trial was observed, $F(5.64, 180.34) = 1.31$, $p = .259$, partial $\eta^2 = .039$, 90% CI [.000, .066].

In Experiment 2, we found no statistically significant main effect of the CS categories on the UR, $F(2, 102) = 0.27$, $p = .764$, partial $\eta^2 = .005$, 90% CI [.000, .033]. In contrast with Experiment 1, we observed a statistically significant main effect of US trial, $F(2.86, 145.95) = 20.18$, $p < .001$, partial $\eta^2 = .284$, 90% CI [.175, .365], reflecting that the UR decreased over trials (see Figure S3.1.1b). This main effect was not qualified by an interaction with the CS categories, $F(6.82, 348.07) = 0.70$, $p = .671$, partial $\eta^2 = .013$, 90% CI [.000, .018].

In Experiment 3, the UR was not modulated by the CS categories, $F(2, 74) = 1.69$, $p = .191$, partial $\eta^2 = .044$, 90% CI [.000, .123]. However, a statistically significant main effect of US trial emerged, $F(2.91, 107.51) = 11.55$, $p < .001$, partial $\eta^2 = .238$, 90% CI [.115, .330], indicating that the UR decreased across trials (see Figure S3.1.1c). This main effect was not qualified by a higher order interaction with CS category, $F(7.35, 272.03) = 1.07$, $p = .385$, partial $\eta^2 = .028$, 90% CI [.000, .040].

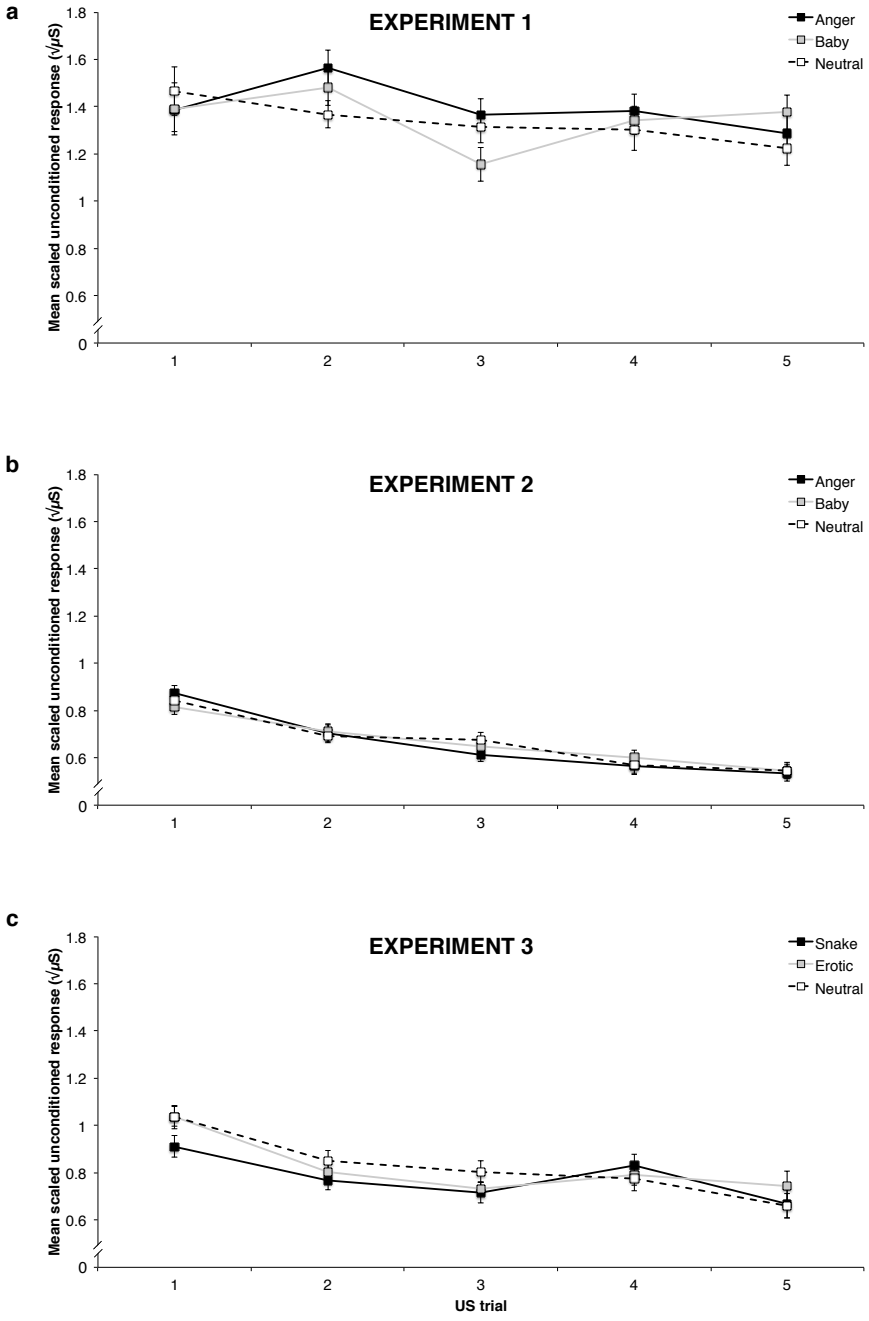

Figure S3.1.1. Mean scaled unconditioned response to the unconditioned stimulus (US; electric stimulation) as a function of the conditioned stimulus category and unconditioned stimulus trial in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Pavlovian learning models

We constructed simple reinforcement learning models (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) to characterize the influence of negative and positive stimuli with biological relevance (i.e., angry faces/snake images and baby faces/erotic stimuli, respectively), relative to neutral stimuli with less relevance (i.e., neutral faces/colored squares), on Pavlovian aversive conditioning. We fitted these models to the SCR data for each CS category separately for parameter estimation and model comparison, and we then compared the parameter estimates of the best-fitting model for each CS category.

Rescorla-Wagner model. The Rescorla-Wagner model (Rescorla & Wagner, 1972) is a classical and standard account of associative learning, in which learning is directly driven by the discrepancy between the actual and the predicted outcome, that is by prediction error. In this model, the value (or associative strength) V at trial $t + 1$ of a given conditioned stimulus j is updated based on the sum of the current expected value V_j at trial t, and the prediction error between the expected value V_j and the outcome R at trial t, weighted by a constant learning rate α:

$$V_j(t+1) = V_j(t) + \alpha \cdot (R(t) - V_j(t))$$

where the learning rate α is a free parameter within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model. The hybrid model introduced by Li et al. (2011) combines both the Rescorla-Wagner model and the Pearce-Hall model (Pearce & Hall, 1980), where the Rescorla-Wagner algorithm is implemented for error-driven value update, and the Pearce-Hall associability mechanism is substituted for the constant learning rate, thus acting as a dynamic learning rate. According to the Pearce-Hall algorithm, the conditioned stimulus' associability decreases when the conditioned stimulus correctly and reliably predicts the actual outcome, whereas it increases when the conditioned stimulus does not reliably predict the actual outcome. In the hybrid model, the value V of a given conditioned stimulus j is updated as follows:

$$V_j(t+1) = V_j(t) + \kappa \cdot \alpha_j(t) \cdot (R(t) - V_j(t))$$

$$\alpha_j(t+1) = \eta \cdot \left| R(t) - V_j(t) \right| + (1 - \eta) \cdot \alpha_j(t)$$

where the initial associability α_0, the learning rate κ, and the weighting factor η are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Rescorla-Wagner model with dual learning rates. As we predicted that both negative and positive biologically relevant stimuli would induce a learning bias, as reflected by an enhanced resistance to extinction and consequently a diminished inhibitory learning, we also implemented a dual-learning-rate model using the Rescorla-Wagner algorithm (see, e.g., Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012), where the learning rate differed as a function of whether the prediction error was positive (i.e., excitatory learning) or negative (i.e., inhibitory learning). To this end, we modified the Rescorla-Wagner model to allow for different learning rates for positive prediction error and for negative prediction error. In the dual-learning-rate Rescorla-Wagner model, the value V of a given conditioned stimulus j is updated as follows:

$$V_j(t+1) = \begin{cases} V_j(t) + \alpha^+ \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) > 0 \\ V_j(t) + \alpha^- \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) < 0 \end{cases}$$

where the learning rate for positive prediction error α^+ and the learning rate for negative prediction error α^- are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model with dual learning rates. We additionally considered a modified hybrid model implementing dual learning rates by allowing for different learning rates for positive prediction error and for negative prediction error. In this model, the value V of a given conditioned stimulus j is updated as follows:

$$V_j(t+1) = \begin{cases} V_j(t) + \kappa^+ \cdot \alpha_j(t) \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) > 0 \\ V_j(t) + \kappa^- \cdot \alpha_j(t) \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) < 0 \end{cases}$$

$$\alpha_j(t+1) = \eta \cdot \left| R(t) - V_j(t) \right| + (1 - \eta) \cdot \alpha_j(t)$$

where the initial associability α_0, the learning rate for positive prediction error κ^+, the learning rate for negative prediction error κ^-, and the weighting factor η are free parameters within the range [0, 1]. If the unconditioned stimulus was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Model and parameter fitting. The free parameters of the models were optimized using maximum a posteriori estimation, which found the set of parameters maximizing the probability of individual participant's trial-by-trial normalized (i.e., scaled and square-root-transformed) skin conductance response (SCR) measured following the conditioned stimulus (CS) given the model, constrained by regularizing priors (see Gershman, 2016; Niv et al., 2012). All the free parameters were constrained with a Beta (1.2, 1.2) prior distribution favoring a normal distribution of the parameter estimates. For the Rescorla-Wagner model (RW[V]) and the dual-learning-rate Rescorla-Wagner model (dual RW[V]), the trial-by-trial time series of CS values $V(t)$ was used to optimize the free parameters; for the Hybrid model and the dual-learning-rate Hybrid model, the free parameters were optimized separately for each possible combination using the trial-by-trial time series of CS values $V(t)$ (Hybrid[V] and dual Hybrid[V]), the trial-by-trial time series of CS associability $\alpha(t)$ (Hybrid[α] and dual Hybrid[α]), or the combination of both (Hybrid[$V+\alpha$] and dual Hybrid[$V+\alpha$]; see Li et al., 2011; Zhang et al., 2016). Initial values (V_0) for each CS were set to 0.5, as participants expected to receive electric stimulations due to the work-up procedure and the instructions. The models were fit using a separate set of free parameters for each participant (i) across all trials, and (ii) separately for each CS category (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013), thereby allowing for comparing the best-fitting parameter estimates among the different CS categories.

Model comparison. Model comparison was conducted using Bayesian information criterion (BIC; Schwarz, 1978; see also, e.g., Zhang et al., 2016), which quantitatively measures the models' goodness of fit, while taking into account and penalizing for the number of free parameters included in each model. The BIC value was calculated for each model averaged across participants using models with individual participant's parameter estimates. To ensure that the models outperformed a model with random predictions, we also compared the models against a baseline model, in which the value $V_f(t)$ and the prediction error were updated at each trial by adding random noise from a uniform random distribution within the range [-0.1, 0.1] (Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013). The BIC values for each model across the three experiments are reported in Table S3.1.2.

Table S3.1.2

Goodness of fit to skin conductance responses for individual models using the mean Bayesian Information Criterion (BIC) in Experiment 1 (N = 40), Experiment 2 (N = 59), and Experiment 3 (N = 40).

		Model								
Exp.	CS category	RW(*V*)	Dual RW(*V*)	Hybrid (*V*)	Hybrid (α)	Hybrid (*V*+α)	Dual Hybrid (*V*)	Dual Hybrid (α)	Dual Hybrid (*V*+α)	Baseline
1	All	41.84	38.22	49.37	40.74	41.45	45.40	43.90	44.42	51.78
	Anger	16.28	16.13	22.78	18.18	18.66	22.06	20.60	21.26	21.52
	Baby	13.65	13.50	19.49	16.47	17.30	19.32	17.93	18.38	19.51
	Neutral	7.93	7.51	14.36	9.09	9.23	13.02	11.89	12.13	13.11
2	All	41.49	35.84	48.77	39.46	39.82	42.88	42.58	42.61	51.34
	Anger	11.08	9.89	17.19	12.43	12.72	15.72	14.73	15.04	16.84
	Baby	12.02	10.86	18.42	14.21	14.72	16.91	15.93	16.21	17.40
	Neutral	12.24	11.46	18.28	13.72	13.75	16.64	15.68	15.93	18.14
3	All	53.75	53.74	61.46	57.27	57.38	61.55	60.35	60.63	61.70
	Snake	17.67	18.86	24.24	21.87	21.98	24.99	23.20	24.37	21.75
	Erotic	16.28	16.41	22.26	19.08	19.43	22.30	21.18	21.67	21.62
	Neutral	15.26	17.32	21.74	19.41	20.08	23.46	21.73	22.85	20.08

Note. Exp. = Experiment, RW = Rescorla-Wagner model, *V* = model values, α = associabilities, Dual = dual learning rates.

Relationship between modeled learning signals and participants' normalized skin conductance responses. To investigate whether and to what extent modeled learning signals from the optimized model predicted participants' trial-by-trial normalized SCRs, we computed a linear regression, in which we regressed value and prediction error time series generated using individual parameter estimates from the best-fitting model and averaged across participants against the trial-by-trial normalized SCRs averaged across participants. Across the three experiments, the best-fitting model based on all trials consisted of the Rescorla-Wagner model implementing dual learning rates (see Table S3.1.2). In Experiment 1, the results of the multiple linear regression analysis showed that value and prediction error signals generated from the Rescorla-Wagner model with dual learning rates explained a statistically significant amount of variance of trial-by-trial normalized SCRs (R^2 = .361, R^2_{adj} = .346, $F(2, 87) = 24.56$, $p < .001$). Value signals were found to statistically significantly predict trial-by-trial normalized SCRs, $\beta = .42$, $t(87) = 7.01$, $p < .001$ (see Figure S3.1.2a), which was not the case for prediction error signals, $\beta = .01$, $t(87) = 0.32$, $p = .751$.

In Experiment 2, value and prediction error signals generated from the Rescorla-Wagner model with dual learning rates likewise explained a statistically significant portion of variance of the trial-by-trial normalized SCRs (R^2 = .426, R^2_{adj} = .413, $F(2, 87) = 32.31$, $p < .001$). As in Experiment 1, trial-by-trial normalized SCRs were predicted by value signals, $\beta = .41$, $t(87) = 8.03$, $p < .001$ (see Figure S3.1.2b), but not by prediction error signals, $\beta = -.001$, $t(87) = -0.04$, $p = .969$.

In Experiment 3, the multiple linear regression indicated that value and prediction errors signals generated from the dual-learning-rate Rescorla-Wagner model explained a statistically significant, though considerably lower, amount of variance of trial-by-trial normalized SCRs (R^2 = .251, R^2_{adj} = .234, $F(2, 87) = 14.62$, $p < .001$). Value signals statistically significantly predicted trial-by-trial normalized SCRs, $\beta = .23$, $t(87) = 5.21$, $p < .001$ (see Figure S3.1.2c), while prediction error signals were only a marginally significant predictor, $\beta = .05$, $t(87) = 1.82$, $p = .072$.

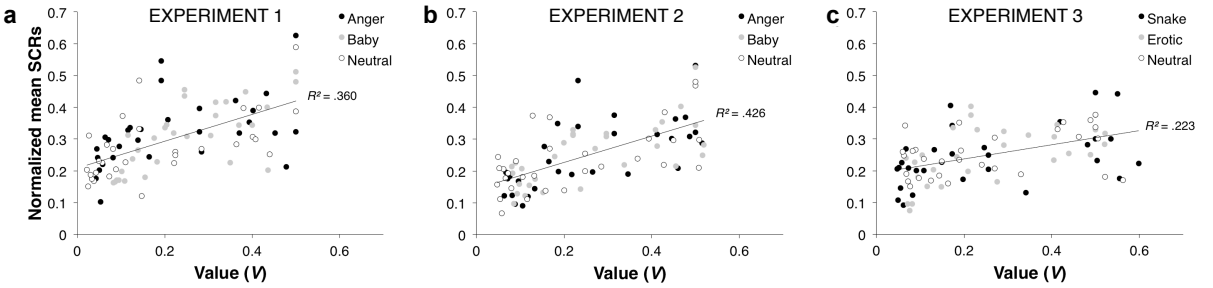

Figure S3.1.2. Relationship between modeled value (*V*) and trial-by-trial normalized skin conductance responses (SCRs) averaged across participants using the individual best-fitting parameters for the Rescorla-Wagner model implementing dual learning rates in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. The curve represents the best-fitting line using least squares estimation.

Parameter estimates analyses. As model comparison using the BIC indicated that the Rescorla-Wagner model implementing dual learning rates provided the best fit to the data based on all trials compared with the other models in all the three experiments (see Table S3.1.2), we therefore analyzed the estimated parameters from this model for each CS category. In Experiment 1, a one-way ANOVA with CS category (anger vs. baby vs. neutral) as a within-participant factor on the learning rate parameter estimates for positive prediction error showed no statistically significant difference between the CS categories, $F(2, 78) = 1.44$, $p = .243$, partial $\eta^2 = .036$, 90% CI [.000, .108] (see Figure S3.1.3a). In contrast, the CS categories differentially influenced the learning rate parameter estimates for negative prediction error, $F(2, 78) = 3.87$, $p = .025$, partial $\eta^2 = .090$, 90% CI [.006, .187]. A planned contrast analysis revealed that the learning rate for negative prediction error was lower for both angry (contrast weight: -1) and baby (contrast weight: -1) faces than for neutral faces (contrast weight: +2), $t(39) = 2.71$, $p = .005$ (one-tailed), $g_{av} = 0.596$, 95% CI [0.145, 1.064], $BF_{10} = 9.356$ (see Figure S3.1.3a), suggesting that angry and baby faces biased inhibitory learning through a diminished impact of negative prediction error. Further pairwise comparisons showed that the estimated learning rate for negative prediction error was lower for baby faces (contrast weight: -1) than for neutral faces (contrast weight: +1), $t(39) = 2.59$, $p = .007$ (one-tailed), $g_{av} = 0.611$, 95% CI [0.127, 1.112], $BF_{10} = 7.224$ (see Figure S3.1.3a), while it was marginally lower for angry faces (contrast weight: -1) relative to neutral faces (contrast weight: +1) with respect to the corrected alpha level for this contrast ($\alpha = .025$) using the Holm-Bonferroni sequential procedure (Holm, 1979), $t(39) = 1.91$, $p = .032$ (one-tailed), $g_{av} = 0.382$, 95% CI [-0.021, 0.794], $BF_{10} = 2.112$ (see Figure S3.1.3a). The estimated learning rate for negative prediction error did not statistically differ for angry faces (contrast weight: -1) compared with baby faces (contrast

weight: +1), $t(39) = -1.06$, $p = .293$ (two-tailed), $g_{av} = -0.257$, 95% CI [-0.746, 0.225], $BF_{10} = 0.381$ (see Figure S3.1.3a).

In Experiment 2, one participant was removed from the Pavlovian learning models analyses since their individual parameters for baby faces could not be estimated due to a lack of SCR to all the baby face CSs during the whole experiment. A one-way ANOVA with CS category (anger vs. baby vs. neutral) as a within-participant factor on the learning rate parameter for positive prediction error revealed no statistically significant difference between the CS categories, $F(2, 116) = 0.28$, $p = .757$, partial $\eta^2 = .005$, 90% CI [.000, .030] (see Figure S3.1.3b). However, the learning rate for negative prediction error parameter estimates were differentially modulated by the CS categories, $F(2, 116) = 4.23$, $p = .017$, partial $\eta^2 = .068$, 90% CI [.007, .142]. Both angry (contrast weight: -1) and baby (contrast weight: -1) faces exhibited a lower learning rate for negative prediction error than neutral faces (contrast weight: +2), $t(58) = 2.80$, $p = .003$ (one-tailed), $g_{av} = 0.433$, 95% CI [0.120, 0.754], $BF_{10} = 11.487$ (see Figure S3.1.3b), reflecting that angry and baby faces biased inhibitory learning. Further comparisons showed that the estimated learning rate for negative prediction error was lower for angry faces (contrast weight: -1) than for neutral faces (contrast weight: +1), $t(58) = 3.03$, $p = .002$ (one-tailed), $g_{av} = 0.465$, 95% CI [0.153, 0.786], $BF_{10} = 19.866$ (see Figure S3.1.3b), whereas it was marginally lower for baby faces (contrast weight: -1) compared with neutral faces (contrast weight: +1) with respect to the corrected alpha level for this contrast ($\alpha = .025$) (Holm, 1979), $t(58) = 1.92$, $p = .030$ (one-tailed), $g_{av} = 0.318$, 95% CI [-0.014, 0.656], $BF_{10} = 1.922$ (see Figure S3.1.3b). The estimated learning rate for negative prediction error for angry faces (contrast weight: -1) did not statistically differ from that for baby faces (contrast weight: +1), $t(58) = 0.77$, $p = .446$ (two-tailed), $g_{av} = 0.126$, 95% CI [-0.200, 0.455], $BF_{10} = 0.256$ (see Figure S3.1.3b).

In Experiment 3, analysis of the estimated learning rate for positive prediction error showed that the main effect of CS category did not reach statistical significance, $F(2, 78) = 2.40$, $p = .098$, partial $\eta^2 = .058$, 90% CI [.000, .143] (see Figure S3.1.3c). In contrast to Experiments 1 and 2, the estimated learning rate for negative prediction error was likewise not differentially modulated by the CS categories, $F(2, 78) = 0.50$, $p = .606$, partial $\eta^2 = .013$, 90% CI [.000, .061] (see Figure S3.1.3c).

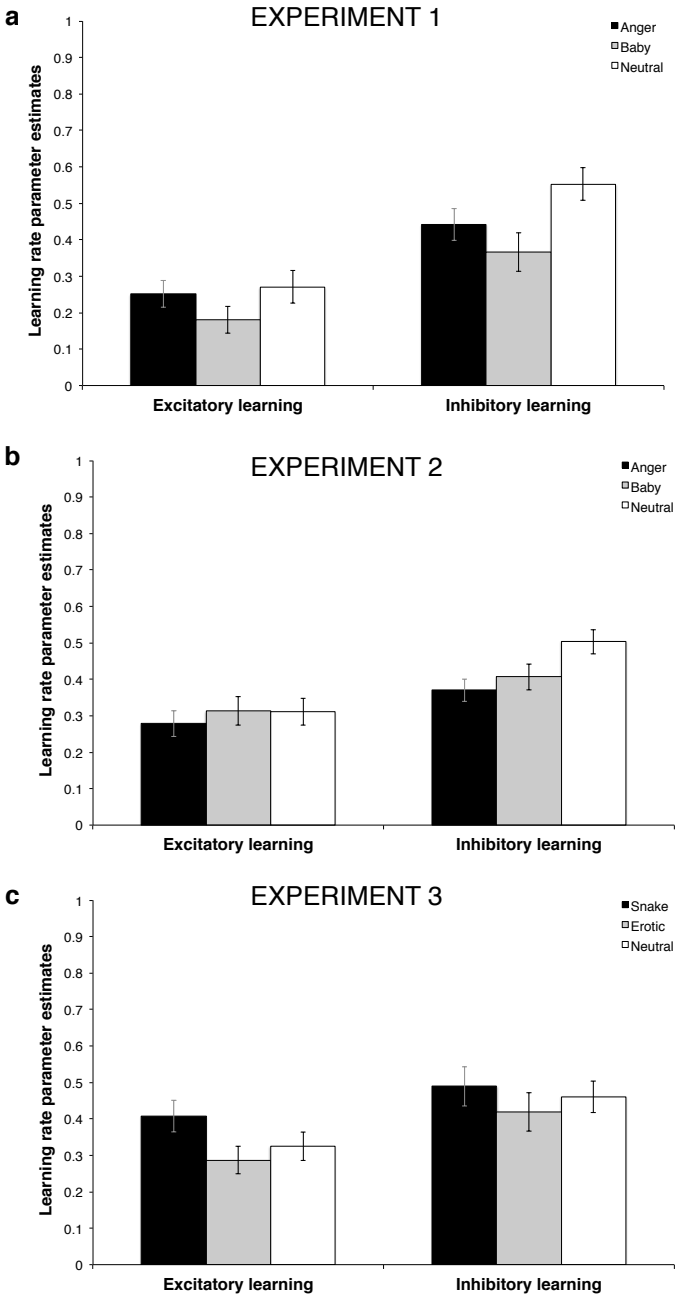

Figure S3.1.3. Learning rate parameter estimates of the Rescorla-Wagner model implementing dual learning rates using individual best-fitting parameters for positive prediction error (excitatory learning) and negative prediction error (inhibitory learning) as a function of the conditioned stimulus category in (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Altogether, the computational analyses using simple reinforcement learning models in Experiments 1 and 2 suggest that the influence of stimulus' biological relevance on Pavlovian aversive conditioning could be best characterized by a lower inhibitory learning rate diminishing the impact of negative prediction error on associative strength, thereby reflecting

a learning bias that may account for the enhanced persistence of the conditioned response to both negative and positive stimuli with biological relevance compared with the conditioned response to neutral stimuli with less relevance. These findings thus seem to provide further evidence for the existence of a shared mechanism underlying preferential Pavlovian aversive conditioning in humans that is common across negative and positive relevant stimuli, as predicted by the relevance detection hypothesis. However, these effects were not observed in Experiment 3, where no statistical difference in learning rate for either positive or negative prediction error across the CS categories was found, possibly because of somewhat noisier SCR data, as suggested by the reduced fit to the data observed in this experiment (see Figure S3.1.2). For these reasons and as the present findings represent only a first attempt to characterize at the computational level the influence of stimulus' biological relevance on Pavlovian conditioning in humans, it is important to highlight that further research is needed to better outline the computational characterization of the influence of stimulus' affective relevance on Pavlovian learning.

Exploratory correlational analysis in Experiment 3

In Experiment 3, we carried out an exploratory correlational analysis using Pearson's correlation coefficients to test whether the participants' CR to erotic stimuli during acquisition and extinction were associated with their dyadic, solitary, and general sexual desire measured with the Sexual Desire Inventory 2 (Spector, Carey, & Steinberg, 1996). One participant was excluded from the correlational analysis between participants' solitary and general sexual desire and their CR to erotic images during acquisition and extinction due to missing data preventing the computation of his solitary and general sexual desire score.

The correlational analysis did not show that participants' dyadic sexual desire was associated with their CR to erotic images during the acquisition ($r(38) = -.121$, $p = .457$, 95% CI [-0.416, 0.197]) or extinction ($r(38) = -.125$, $p = .441$, 95% CI [-0.420, 0.194]) phases. Similarly, no significant correlation was found between participants' solitary sexual desire and their CR to erotic images during acquisition ($r(37) = .172$, $p = .296$, 95% CI [-0.151, 0.462]) or extinction ($r(37) = -.042$, $p = .798$, 95% CI [-0.352, 0.277]). Furthermore, participants' general sexual desire did not correlate with their CR to erotic images in the acquisition phase ($r(37) = .019$, $p = .911$, 95% CI [-0.298, 0.332]) or in the extinction phase ($r(37) = -.116$, $p = .482$, 95% CI [-0.416, 0.207]).

3.2. STUDY 2: LEARNING BIASES TO ANGRY AND HAPPY FACES DURING PAVLOVIAN AVERSIVE CONDITIONING[9]

Abstract

Learning biases in Pavlovian aversive conditioning have been found in response to specific categories of threat-relevant stimuli, such as snakes or angry faces. This has been suggested to reflect a selective predisposition to preferentially learn to associate stimuli that provided threats to survival across evolution with aversive outcomes. Here, we contrast with this perspective by highlighting that both threatening (angry faces) and rewarding (happy faces) social stimuli can produce learning biases during Pavlovian aversive conditioning. Using a differential aversive conditioning paradigm, the present study (N = 107) showed that the conditioned response to angry and happy faces was more readily acquired and more resistant to extinction than the conditioned response to neutral faces. Whereas the effects of faster conditioning to angry and happy faces were of moderate size, the enhanced resistance to extinction to happy faces was of relatively small size and of lesser magnitude than that to angry faces. Strikingly, the conditioned response persistence to happy faces was influenced by inter-individual differences in happy faces' affective evaluation, as indexed by a Go/No-Go Association Task. These findings suggest that the occurrence of learning biases in Pavlovian aversive conditioning is not specific to threat-related stimuli and depends on the stimulus' affective relevance to the organism.

[9] Reprint of: Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2019). *Learning biases to angry and happy faces during Pavlovian aversive conditioning*. Manuscript in preparation.

3.2.1. Introduction

Learning to predict and anticipate impending threats in the environment holds a critical survival value to organisms (e.g., LeDoux & Daw, 2018). A basic form of learning whereby this skill is achieved is Pavlovian aversive conditioning. In this process and procedure, organisms learn to associate a stimulus from the environment (the conditioned stimulus) with a biologically aversive outcome (the unconditioned stimulus) through single or repeated contingent pairing (Pavlov, 1927; Rescorla, 1988), thereby endowing the conditioned stimulus with a predictive and emotional value eliciting an anticipatory response (the conditioned response).

Identifying the mechanisms underlying Pavlovian aversive conditioning has drawn a large interest in the animal and human literature (e.g., Delgado, Olsson, & Phelps, 2006; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Zhang, Mano, Ganesh, Robbins, & Seymour, 2016). Whereas this line of investigation has mainly focused on the general principles that apply across different types of stimuli irrespective of their nature (Pavlov, 1927), certain associations have, however, been revealed to be more easily formed and maintained than others (Garcia & Koelling, 1966; Öhman & Mineka, 2001; Seligman, 1970, 1971), reflecting the existence of learning biases in Pavlovian aversive conditioning. Surprisingly, mechanisms underlying such learning biases remain yet not well elucidated.

In humans, learning biases have been mostly demonstrated to stimuli from specific animal threat-relevant categories, such as snakes or spiders, these stimuli being more readily and persistently associated with aversive events than nonthreatening stimuli, such as birds or flowers (e.g., Ho & Lipp, 2014; Öhman, Eriksson, & Olofsson, 1975; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Öhman & Mineka, 2001; Olsson, Ebert, Banaji, & Phelps, 2005). Similar learning biases have also been observed in response to social threat-relevant stimuli, such as threatening or outgroup faces, in comparison with social nonthreatening stimuli, such as happy, neutral, or ingroup faces (e.g., Öhman & Dimberg, 1978; Olsson et al., 2005; see also Dimberg & Öhman, 1996). These findings have generally been interpreted as supporting the notions of preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001), according to which organisms have been biologically predisposed by evolution to preferentially associate stimuli that provided threats to the species' survival with naturally aversive events (but see Åhs et al. 2018, for a recent systematic review questioning the preparedness account). However, learning biases have also been reported to evolutionarily novel threat-relevant stimuli, such as pointed guns (Flykt, Esteves, & Öhman, 2007; Hugdahl

& Johnsen, 1989). Moreover, learned threat to social threat-relevant stimuli has been shown to be more malleable than to animal threat-relevant stimuli, indicating that learning biases to socially threatening stimuli may hinge on sociocultural rather than genetics factors alone (Mallan, Lipp, & Cochrane, 2013; Olsson et al., 2005). Altogether, these results suggest that the development of learning biases in Pavlovian aversive conditioning likely reflects the complex interplay of evolutionary and cultural factors (Davey, 1995; Lindström, Golkar, & Olsson, 2015).

In line with this view, an alternative framework to the preparedness and fear module theories, which derives from appraisal theories of emotion (e.g., Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005, 2018), proposes that evolutionarily threat-relevant stimuli are preferentially conditioned to threat not because they have been associated with threat through evolution, but because they are highly relevant to the organism's survival and well-being (Stussi, Brosch, & Sander, 2015; Stussi, Ferrero, Pourtois, & Sander, in press; Stussi, Pourtois, & Sander, 2018). Critically, this model predicts that preferential Pavlovian learning is not selective to threat-related stimuli, but extends to stimuli that are relevant to the organism's concerns, such as their psychological and physiological needs, goals, motives, values, or well-being (Frijda, 1986; Pool, Brosch, Delplanque, & Sander, 2016), beyond stimulus' valence or evolutionary status per se (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). Congruent with this hypothesis, a series of three experiments (Stussi, Pourtois, et al., 2018) has demonstrated that, similar to threat-relevant stimuli (angry faces or snakes), positive stimuli with high biological relevance to the organism (baby faces or erotic images) are likewise more persistently associated with an aversive unconditioned stimulus (electric stimulation) than neutral stimuli with less relevance (neutral faces or colored squares). These findings thereby suggest that learning biases during Pavlovian aversive conditioning not only occur in response to negative threat-relevant stimuli, but also to positive stimuli having heightened affective relevance.

If positive stimuli that are affectively relevant can produce learning biases during Pavlovian aversive conditioning, the question arises as to why such learning biases have not been observed to happy faces, the conditioned response thereto typically extinguishing swiftly (see, e.g., Bramwell, Mallan, & Lipp, 2014[10]; Esteves, Parra, Dimberg, & Öhman, 1994;

[10] Of note, Bramwell et al. (2014) reported resistance to extinction to outgroup race happy faces, thereby indicating that happy faces may lead to preferential aversive learning under certain circumstances. This effect was not due to negative evaluation of outgroup happy faces, which were evaluated as more pleasant than ingroup happy faces

Mazurski, Bond, Siddle, & Lovibond, 1996; Öhman & Dimberg, 1978; Öhman & Mineka, 2001; Rowles, Mallan, & Lipp, 2012; see also Dimberg & Öhman, 1996). Based on appraisal theories, we suggest that this apparent inconsistency stems from the fact that happy faces generally have a comparatively lower level of relevance to the organism than do other positive stimuli with enhanced biological relevance, such as baby faces, or threat-relevant stimuli, such as threat-related faces (Brosch, Pourtois, & Sander, 2010; Brosch, Sander, Pourtois, & Scherer, 2008; Pool et al., 2016). Threat-related and baby faces demand rapid in-depth processing for response preparation due to their high relevance to the organism and species' survival, and are thus likely to be consistently detected as highly relevant across individuals. Although happy faces can be considered as affectively relevant as a social reward, they do not require such immediate response preparation and can carry several meanings (e.g., Ambadar, Cohn, & Reed, 2009; Martin, Rychlowska, Wood, & Niedenthal, 2017), the processing thereof being likely varying as a function of the situation and individual differences (see, e.g., Canli, Silvers, Whitfield, Gotlib, & Gabrieli, 2002), hence resulting in weak learning biases. Consistent with this suggestion, fearful faces have been reported to consistently activate the amygdala, which plays a pivotal role in Pavlovian conditioning (Büchel, Morris, Dolan, & Friston, 1998; LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; LeDoux, 2012; LeDoux & Daw, 2018; Phelps & LeDoux, 2005) and relevance detection (Pessoa & Adolphs, 2010; Sander et al., 2003), across individuals; by contrast, amygdalar activation to happy faces appears to depend on inter-individual differences in extraversion (Canli et al., 2002). Moreover, a meta-analysis on attentional bias for positive emotional stimuli has revealed that, whereas baby and happy faces both led to a clear-cut attentional bias (i.e., facilitated orienting effect) compared with neutral faces, this bias was less clearly outspoken for happy faces where the corresponding effect size was small (Pool et al., 2016). Although these results suggest that learning biases to happy faces likely are of relatively small magnitude and sensitive to inter-individual differences, prior research investigating Pavlovian aversive conditioning to them has however mainly used small sample sizes, thereby undermining the possibility to explore the role of inter-individual differences in this process (typical *n* by group ranged between 15 and 25; see Bramwell et al., 2014; Esteves, Dimberg, & Öhman, 1994; Esteves, Parra, et al., 1994; Mazurski et al., 1996; Öhman & Dimberg, 1978; Rowles et al., 2012).

at the explicit level, whereas no difference in positive or negative evaluation was found between them at the implicit level. Nevertheless, no resistance to extinction was observed to ingroup happy faces, which suggests that the enhanced persistence of threat conditioned to outgroup happy faces was likely driven by the faces' race category.

Here, we sought to investigate whether learning biases may be observed to happy faces compared with neutral faces during Pavlovian aversive conditioning in a relatively large sample size (N = 107), and whether such learning biases are modulated by inter-individual differences in relevance appraisal of happy faces, thereby resulting in smaller effects than those found in response to angry faces at the group level. To this end, we used a differential Pavlovian aversive conditioning paradigm, in which two angry, happy, and neutral faces were presented as conditioned stimuli (CSs). One stimulus (CS+) from each CS category was systematically associated with a mild electric stimulation, whereas the other stimulus (CS-) from each CS category was never paired with the stimulation. Following standard practice, the conditioned response (CR) was operationalized as the differential skin conductance response (SCR) to the CS+ minus CS- from the same CS category, and used as an index of learning (see, e.g., Olsson et al., 2005; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). To assess inter-individual differences in relevance appraisal of happy faces, we examined the role of personality trait extraversion and happy faces' affective evaluation therein. We chose extraversion as this specific dimension has previously been related to the strength of Pavlovian aversive conditioning (e.g., Eysenck, 1965), even though the nature of this relation remains equivocal (Pineles, Vogt, & Orr, 2009). More importantly, individuals high in extraversion are typically characterized by high sociability, social attention, reward sensitivity, and positive affect (e.g., Ashton, Lee, & Paunonen, 2002; Lucas, Diener, Grob, Suh, & Shao, 2002; Smillie, 2013), and further show enhanced amygdala processing for happy faces (Canli et al., 2002). Accordingly, it has been suggested that they appraise happy faces as more relevant to their concerns than individuals lower in extraversion (Sander et al., 2003, 2005). We additionally assessed implicit associations between the different face categories and importance (see, e.g., Critcher & Ferguson, 2016) through a Go/No-go Association Task (GNAT; Nosek & Banaji, 2001) to indirectly measure individuals' affective evaluation of these faces. Specifically, we inferred that individuals appraising the various faces as more relevant to their concerns associate them more easily and rapidly with the attribute of importance (versus unimportance) than individuals who do not have this tendency.

Given that learning biases in Pavlovian aversive conditioning are generally reflected by a faster acquisition of a CR and/or an enhanced resistance to extinction of that CR (e.g., Öhman & Mineka, 2001), we predicted that (a) the CR to angry faces would be more readily acquired and more resistant to extinction than the CR to both happy faces and neutral faces across participants, whereas (b) the CR to happy faces would be acquired more readily and more

resistant to extinction than the CR to neutral faces. Moreover, we hypothesized that (c) participants' extraversion level, as well as the sensitivity and rapidity with which they associated happy faces with the attribute of importance versus unimportance, would predict the CR acquisition readiness and resistance to extinction to happy faces.

3.2.2. Method

Participants

One hundred and seventeen students from the University of Geneva participated in the experiment, which was approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva. They provided informed consent and received partial course credit for their participation. Ten participants were excluded from the analyses because of technical problems (n = 2), for displaying virtually no SCR (n = 2), for failing to acquire a CR to at least one of the CSs+ (n = 5), or for withdrawing from the study early (n = 1). These exclusion criteria were determined prior to data collection (see Olsson et al., 2005; Olsson & Phelps, 2004; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). The final sample size consisted of 107 participants (85 women, 22 men), aged between 19 and 34 years old (mean age = 21.85 ± 2.57 years). The sample size was established before data collection on the basis of the current heuristic suggesting a sample of at least 100 participants for studies considering inter-individual differences (see, e.g., Dubois & Adolphs, 2016). For counterbalancing purposes, we aimed to recruit a minimum sample size of 104 participants exhibiting differential conditioning to at least of one of three CS categories. We stopped collecting data at the end of the academic year and ascertained that the established sample size had been reached. A sensitivity power analysis performed with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size allowed for detecting a smallest population effect size of d_z = 0.242 with a power of 80% using a one-tailed paired-sample t test.

Apparatus and stimuli

The experiment took place in a sound-attenuated experimental chamber. The stimuli were presented using MATLAB (The MathWorks Inc., Natick, MA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) and displayed on a 23-inch LED monitor. Eight angry, eight happy, and eight neutral male face stimuli from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) were used either as targets or as distractors in the GNAT (see 3.2.5. Supplementary materials). Four word stimuli related to the

attribute of importance (i.e., important words; “important”, “relevant”, “significant”, “impactful”) and four word stimuli related to the attribute of unimportance (i.e., unimportant words; “unimportant”, “irrelevant”, “insignificant”, “secondary”) were also used both as targets and distractors.

In the differential Pavlovian aversive conditioning procedure, the CSs consisted of two male angry (model numbers AM10ANS, AM29ANS), two male happy (model numbers AM07HAS, AM22HAS), and two male neutral (model numbers AM11NES, AM31NES) faces taken from the KDEF (Lundqvist et al., 1998). These faces were selected based on the correct identification (hit rate range: 89.06%-100%) and intensity ratings (mean intensity range: 5.73-7.63) of their respective emotional expression (see Goeleven, De Raedt, Leyman, & Verschuere, 2008). Each face served both as a CS+ and as a CS-, counterbalanced across participants. Subjective ratings performed before the conditioning procedure (see 3.2.5. Supplementary materials) on a visual analog scale from 0 (*very unpleasant*) to 100 (*very pleasant*) indicated that the angry faces were evaluated as unpleasant ($M = 15.29$, $SD = 15.76$), the happy faces as pleasant ($M = 68.28$, $SD = 20.39$), and the neutral faces as relatively neutral ($M = 43.47$, $SD = 13.07$). The unconditioned stimulus (US) was a mild electric stimulation (200-ms duration) delivered to the participants’ right wrist through a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA). The CR was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants’ left hand. SCR was continuously recorded during the conditioning procedure with a sampling rate of 1000 Hz by means of a BIOPAC MP150 system (Santa Barbara, CA). The SCR data were analyzed offline with AcqKnowledge software (version 4.4; BIOPAC Sytems Inc., Goleta, CA).

Procedure

Between two to eight months prior to their participation in the study, participants completed the French version of the NEO-FFI (Costa & McCrae, 1992; Rolland, Parker, & Strumpf, 1998). Upon arrival at the laboratory, they were informed about the general layout of the experiment, provided written informed consent, and performed the GNAT. Participants were next asked to evaluate the to-be-CSs according to various dimensions (see 3.2.5. Supplementary materials) before undergoing the differential Pavlovian aversive conditioning procedure. Finally, they were asked again to provide subjective ratings of the CSs after conditioning (see 3.2.5. Supplementary materials) and were debriefed.

NEO-FFI. The NEO-FFI is a standard personality inventory measuring the Big Five personality traits consisting of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (Costa & McCrae, 1992). It comprises 60 items (12 per trait), each of which is measured on a 5-point Likert scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*). Given our a priori hypotheses, we focused here on extraversion (M = 28.23, SD = 5.69, range = 10-40, Cronbach's α = .76; see Figure S3.2.1 in the supplementary materials). Exploratory analyses including the other personality traits are reported in the supplementary materials.

Go/No-go Association Task. In the GNAT, participants were presented with faces from three emotional categories (angry vs. happy vs. neutral) and words from two categories (important vs. unimportant). In each trial, a face or a word was displayed at the center of the screen. Participants were instructed to press as quickly and accurately as possible on the "A" key if the stimulus was a member of a target category (go trials), but to withdraw from responding otherwise (no-go trials). Throughout the task, the labels of the target categories were continuously displayed at the top of the screen as a reminder. After each trial, feedback about participants' response was displayed at the bottom of the screen (i.e., a green check for correct or a red cross for incorrect) during a 150-ms inter-trial interval (see Figure 3.2.1).

The GNAT began with a practice session of five blocks in which there was only a single target category (see 3.2.5. Supplementary materials). The experimental session ensued and was composed of three parts, each divided into two blocks. Within each part, a specific face category was one of the two target categories with "important" words being the other target category in block 1, and "unimportant" words the other target category in block 2. The order of the three parts as a function of the face categories was counterbalanced between participants. Each block consisted of 96 trials: 16 training trials and 80 critical trials. Four faces from the target face category and two faces from each distractor face category were presented intermixed with the four "important" and the four "unimportant" words in a pseudorandom order. The response deadline was idiosyncratically adapted to the participants' reaction times and response accuracy (see, e.g., Coppin et al., 2016; Nosek & Banaji, 2001): When response was correct (for both go and no-go trials) and reaction time faster than the arbitrary response deadline (for go trials), the response deadline for the next trial was set as 500 ms or as 666 ms if reaction time was slower than 500 ms but faster than 666 ms (for go trials); otherwise, it was set as 800 ms.

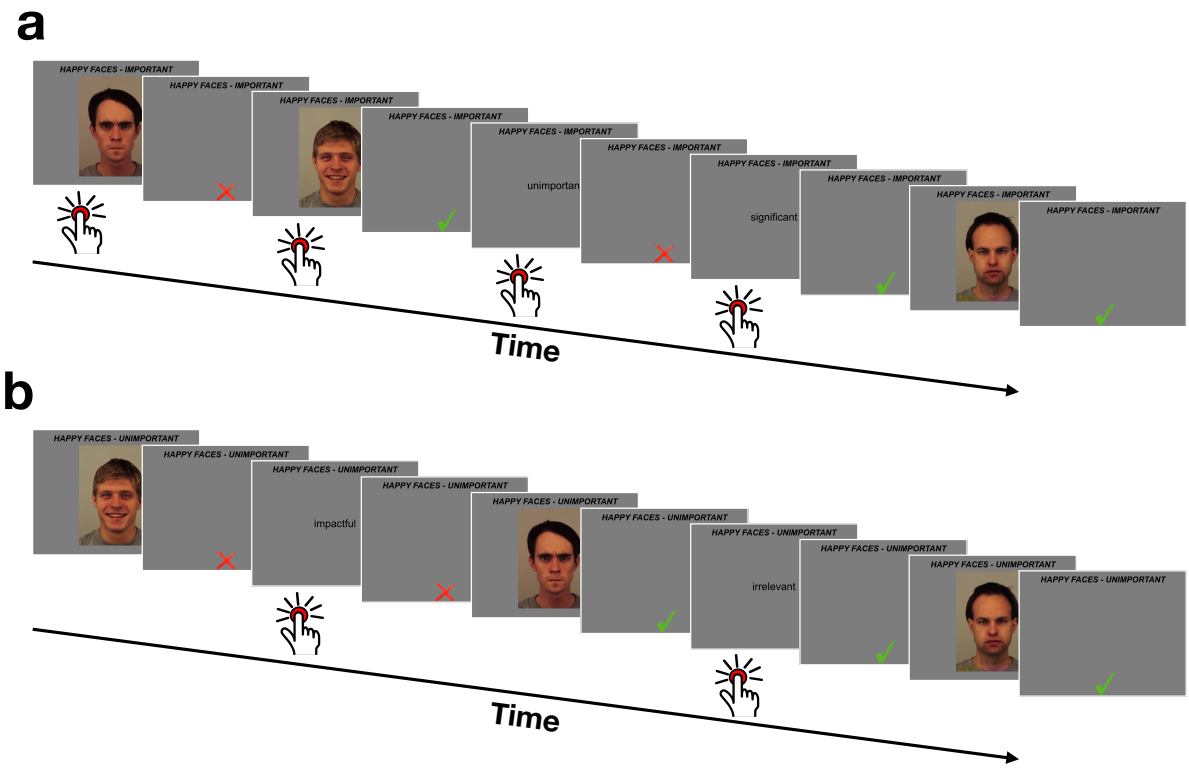

Figure 3.3.1. Illustration of the Go/No-go Association Task. (a) Example of five trials of a Go/no-go block in which participants had to detect whether the faces and the words belong to the target categories "Happy faces" or "Important words", respectively. If the face or word belonged to one of the two target categories, the correct response was to press 'A' on the keyboard, whereas if the face or word belonged to neither of the two target categories, the correct response was to withdraw from responding. After each response, participants received feedback consisting of either a green check for correct responses, or a red cross for incorrect responses. (b) Example of five trials of a Go/no-go block in which participants had to detect whether the faces and the words belong to the target categories "Happy faces" or "Unimportant words", respectively. If the face or word belonged to one of the two target categories, the correct response was to press 'A' on the keyboard, whereas the correct response was to withdraw from responding if the face or word belonged to neither of the two target categories.

Participants' reaction times and response accuracy were recorded for each trial. All trials with reaction times faster than 100 ms were excluded from analysis. According to signal detection theory, we calculated a d' score for each block within each part of the GNAT experimental session, considering only critical trials (Nosek & Banaji, 2001). We converted the proportions of hits (correct go-responses to targets) and false alarms (incorrect go-responses to distractors) to z scores before computing the difference between them, thereby obtaining d'. Hit and false-alarm rates equal to 0 or 1 were replaced with 1/(2N) and 1 – 1/(2N), respectively, where N is the number of trials (Macmillan & Creelman, 2005). A differential d' index was

then calculated by subtracting the d' scores of the second block (target face category + unimportant words) from those of the first block (target face category + important words; see, e.g., Coppin et al., 2016). Higher values on the differential d' index indicated higher sensitivity when faces from the target face category and "important" words were targets in comparison with when faces from the target face category and "unimportant" words were targets. The differential d' index was used to assess participants' sensitivity to the association between a given face category and importance versus unimportance. Additionally, we computed a differential index for reaction times by subtracting the mean reaction times of the first block to those of the second block, higher values thus reflecting faster responses when faces from the target face category and "important" words were targets relative to when faces from the target face category and "unimportant" words were targets. The differential reaction time index served as an indicator of the speed with which participants associated the face categories with the attribute of importance compared with that of unimportance.

Differential Pavlovian aversive conditioning. Prior to conditioning, the electrodes for measuring SCR and delivering the electric stimulation were attached to participants. A work-up procedure was then performed to individually calibrate the electric stimulation intensity (M = 34.55 V, SD = 7.57, range = 20-50 V) to a level reported as "uncomfortable, but not painful". The differential Pavlovian aversive conditioning procedure comprised three contiguous phases. In the initial habituation phase, the six CSs were each presented twice without being reinforced. During the subsequent acquisition phase, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Each CS+ was paired with the US with a partial reinforcement schedule, five of the seven CS+ presentations coterminating with the US delivery, whereas the CS- from each CS category was never associated with the US. The use of a partial reinforcement schedule aimed to potentiate the CR resistance to extinction, hence optimizing the examination of differences between the three CS categories used. The final extinction phase consisted of six unreinforced presentations of each CS. During all the conditioning phases, the CSs were presented for 6 s with an intertrial interval varying from 12 to 15 s. The CSs' presentation order was pseudorandomized into eight different orders to counterbalance the associations between the face stimuli and CS type (CS+ vs. CS-) across the three CS categories (angry vs. happy vs. neutral).

Response definition

SCR was scored for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window following CS

onset. The minimal response criterion was 0.02 μS, and responses below this criterion were scored as zero and remained in the analysis. A low-pass filter (Blackman -92 dB, 1 Hz) was applied on the SCR data before analysis. SCRs were detected automatically with AcqKnowledge software and manually checked for artifacts and response detection. Trials containing artifacts affecting the scoring of event-related SCRs (0.17%) were removed from the subsequent analyses. The raw SCRs were scaled according to each participant's mean unconditioned response (UR), and square-root-transformed to normalize the distributions. The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window after the US delivery, and the mean UR was calculated across all USs for each participant. The habituation means comprised the first two presentations of each CS (i.e., Trials 1 and 2). In order to tease apart effects of faster conditioning from those of larger conditioning, the acquisition means were split into an early (i.e., the first three presentations of each CS following the first pairing between the CS+ of a given CS category and the US; Trials 4 to 6) and a late (i.e., the following three presentations of each CS; Trials 7 to 9) phase (see, e.g., Lonsdorf et al., 2017; Olsson, Carmona, Downey, Bolger, & Ochsner, 2013; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). This allowed us to specifically examine the CR acquisition readiness during early acquisition. The first acquisition trial for each CS was removed from the analysis because the CSs+ became predictive of the US only after their first association therewith. The extinction means encompassed the last six presentations of each CS (i.e., Trials 10 to 15). The conditioning data analyses were performed on the CR, which was calculated as the SCR to the CS+ minus the SCR to the CS- from the same CS category (e.g., Olsson et al., 2005; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). This procedure allows for reducing possible preexisting differences in emotional salience between the different CS categories (Olsson et al., 2005).

Statistical analyses

The differential d' and the differential reaction time indices derived from the GNAT were each analyzed with a one-way repeated-measures analysis of variance (ANOVA) with face category (angry vs. happy vs. neutral) as a within-participant factor. Statistically significant main effects were followed up with a multiple comparison procedure using Tukey's HSD tests when applicable.

Following standard practice in the human conditioning literature (e.g., Lonsdorf et al., 2017; Olsson et al., 2005; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018), the SCR data was analyzed separately for each conditioning phase. The habituation and extinction

phases were each analyzed with a one-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) as a within-participant factor. The acquisition phase was analyzed with a two-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) and time (early vs. late) as within-participant factors. One-sample *t* tests were additionally performed to test whether differential conditioning occurred to the CS categories across the entire acquisition phase. To specifically test our a priori hypotheses, we conducted planned contrast analyses comparing the CR during early acquisition and during extinction to (a) angry (contrast weight: +1) versus neutral (contrast weight: -1) faces, (b) happy (contrast weight: +1) versus neutral (contrast weight: -1) faces, and (c) angry (contrast weight: +1) versus happy (contrast weight: -1) faces. As these contrasts were nonorthogonal, we applied a Holm-Bonferroni sequential procedure (Holm, 1979) to correct for multiple comparisons. The alpha level of the contrast with the lowest *p* value was set as $\alpha = .05/3 = .0167$, the alpha level with the second lowest *p* value as $\alpha = .05/2 = .025$, and the alpha level with the highest *p* value as $\alpha = .05$. For each planned contrast, we also calculated the Bayes factor (BF_{10}) quantifying the likelihood of the data under the alternative hypothesis compared with the likelihood of the data under the null hypothesis (e.g., Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Because we expected moderate effects for angry faces and relatively small effects for happy faces, we used a Cauchy prior width of 0.5 for the comparisons between angry and happy faces and between angry and neutral faces (see Stussi, Pourtois, et al., 2018), and of 0.25 for the comparison between happy and neutral faces. We performed one-sided testing to test our theory-driven directional hypotheses (one-sample *t* tests, contrasts a, b, and c).

To assess our a priori hypotheses that extraversion, as well as the sensitivity and the rapidity with which happy faces were associated with the attribute of importance predicted the CR acquisition readiness and resistance to extinction thereto, we conducted multiple linear regression analyses. These analyses tested whether the CR acquisition readiness (i.e., during early acquisition) and persistence (i.e., during extinction) to happy faces were predicted by participants' (a) extraversion level, (b) differential d' index for happy faces, and (c) differential reaction time index for happy faces. Further exploratory multiple linear regression analyses carried out on the CR to angry and neutral faces during early acquisition and extinction to investigate the specificity of these predictive effects are reported in the supplementary materials.

All statistical analyses were performed with R (R Core Team, 2018). Huynh-Feldt adjustments of degrees of freedom were applied for repeated-measures ANOVAs when

appropriate. Partial η^2 or Hedges' g_{av} and their 90% or 95% confidence interval (CI) were used as estimates of effect sizes (see Lakens, 2013) for the repeated-measures ANOVAs and the planned contrasts analyses, respectively, whereas the coefficient of determination R^2 along with its 90% CI was used for multiple linear regressions.

3.2.3. Results

Pavlovian aversive conditioning

Figure 3.2.2 depicts the mean SCR to angry, happy, and neutral faces across the habituation, acquisition, and extinction phases of the differential Pavlovian aversive conditioning separately for the CS+ and the CS-. In the habituation phase, no preexisting difference in differential SCR across the CS categories (angry vs. happy vs. neutral) was found, $F(2, 212) = 0.003$, $p = .997$, partial $\eta^2 = .00003$, 90% CI [.000, .0006].

Analysis of the acquisition phase revealed successful differential conditioning to all three CS categories, as reflected by larger SCRs to the CS+ than to the CS- for angry, $t(106) = 7.44$, $p < .001$ (one-tailed), $g_{av} = 1.010$, 95% CI [0.714, 1.316], happy, $t(106) = 8.10$, $p < .001$ (one-tailed), $g_{av} = 1.099$, 95% CI [0.798, 1.411], and neutral faces, $t(106) = 5.97$, $p < .001$ (one-tailed), $g_{av} = 0.811$, 95% CI [0.525, 1.105]. The CS categories however differentially influenced the CR acquisition as indicated by a statistically significant main effect of CS category, $F(2, 212) = 3.27$, $p = .040$, partial $\eta^2 = .030$, 90% CI [.001, .071], and a marginal trend for an interaction effect between CS category and time, $F(2, 212) = 2.60$, $p = .076$, partial $\eta^2 = .024$, 90% CI [.000, .062]. Congruent with our a priori hypothesis, a planned contrast analysis showed that the CR to angry faces was more readily acquired than the CR to neutral faces during early acquisition, $t(106) = 2.60$, $p = .005$ (one-tailed), $g_{av} = 0.358$, 95% CI [0.084, 0.636], $BF_{10} = 6.642$ (see Figure 3). Importantly, the CR to happy faces was likewise more readily acquired than to neutral faces, $t(106) = 3.25$, $p < .001$ (one-tailed), $g_{av} = 0.442$, 95% CI [0.169, 0.720], $BF_{10} = 41.237$, whereas there was no statistical difference in CR acquisition readiness to angry faces compared with happy faces, $t(106) = -0.58$, $p = .717$ (one-tailed), $g_{av} = -0.073$, 95% CI [-0.324, 0.177], $BF_{10} = 0.101$ (see Figure 3.2.3).

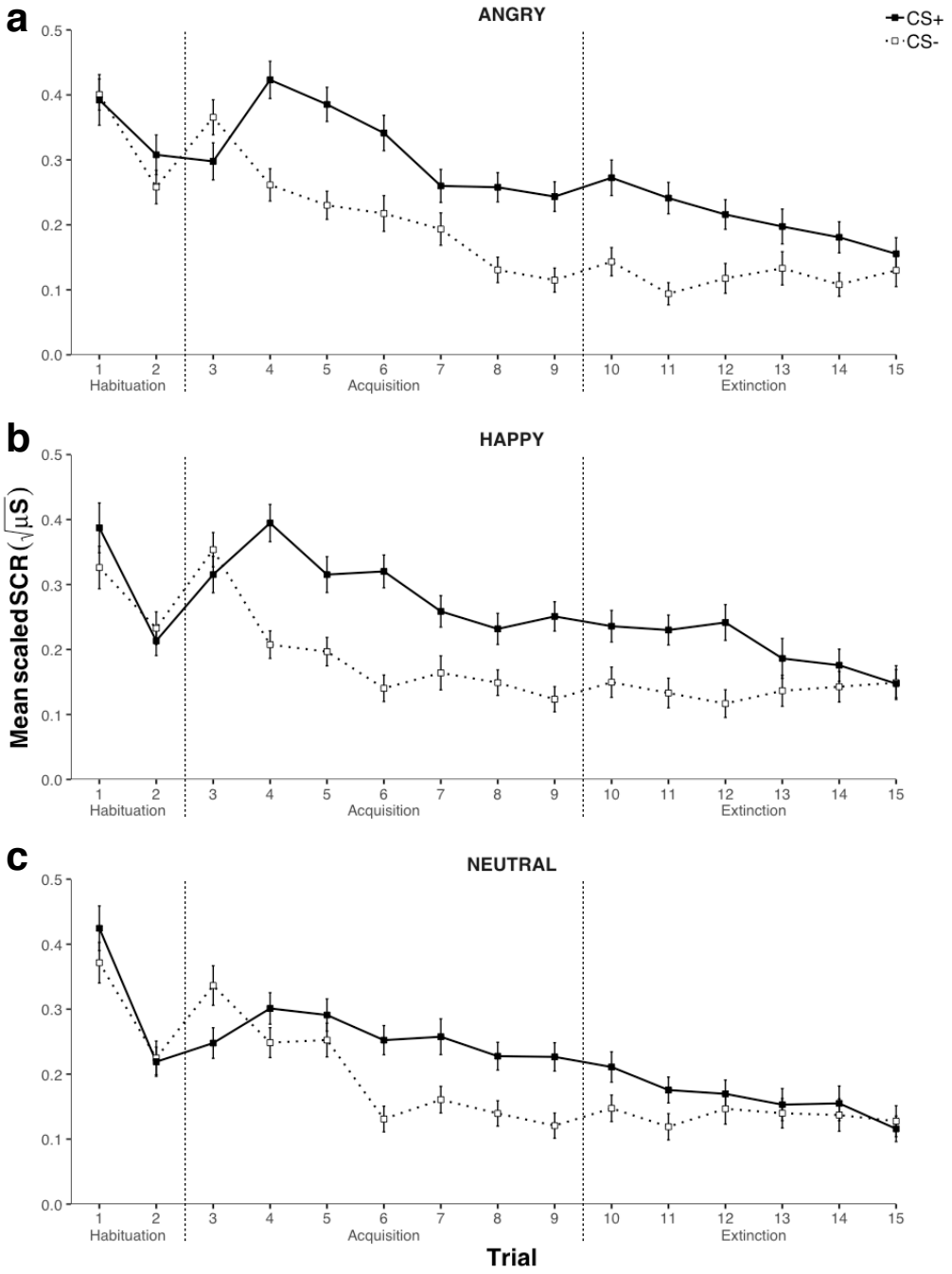

Figure 3.2.2. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of the conditioned stimulus type (CS+ vs. CS-) across trials. Mean scaled SCR to (a) angry faces, (b) happy faces, and (c) neutral faces. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Critically, the CR persistence was also modulated by the CS categories in extinction, $F(2, 212) = 5.97$, $p = .003$, partial $\eta^2 = .053$, 90% CI [.011, .104]. As predicted, the CR to angry faces was more resistant to extinction than the CR to neutral faces, $t(106) = 3.69$, $p < .001$ (one-tailed), $g_{av} = 0.432$, 95% CI [0.196, 0.672], $BF_{10} = 133.200$. Similarly, the CR to happy faces

was more persistent than to neutral faces, $t(106) = 2.01$, $p = .024$ (one-tailed), $g_{av} = 0.247$, 95% CI [0.003, 0.493], $BF_{10} = 2.777$ (see Figure 3.2.3). By comparison, we did not observe an enhanced CR persistence to angry faces relative to happy faces, $t(106) = 1.28$, $p = .102$ (one-tailed), $g_{av} = 0.133$, 95% CI [-0.072, 0.339], $BF_{10} = 0.573$.

Go/No-go Association Task

The analysis of the differential d' index showed a statistically significant main effect of face category (angry vs. happy vs. neutral), $F(2, 212) = 15.46$, $p < .001$, partial $\eta^2 = .127$, 90% CI [.061, .193]. The differential d' index was higher for happy faces ($M = 0.15$, $SD = 0.55$) than for angry ($M = -0.20$, $SD = 0.46$; $p < .001$, $g_{av} = 0.683$, 95% CI [0.407, 0.965]) and neutral faces ($M = -0.10$, $SD = 0.44$; $p < .001$, $g_{av} = 0.493$, 95% CI [0.222, 0.769]), whereas there was no statistical difference between angry and neutral faces ($p = .273$, $g_{av} = 0.219$, 95% CI [-0.030, 0.469]). These results showed that participants exhibited a greater sensitivity to the association between the attribute of importance versus unimportance with happy faces than either angry or

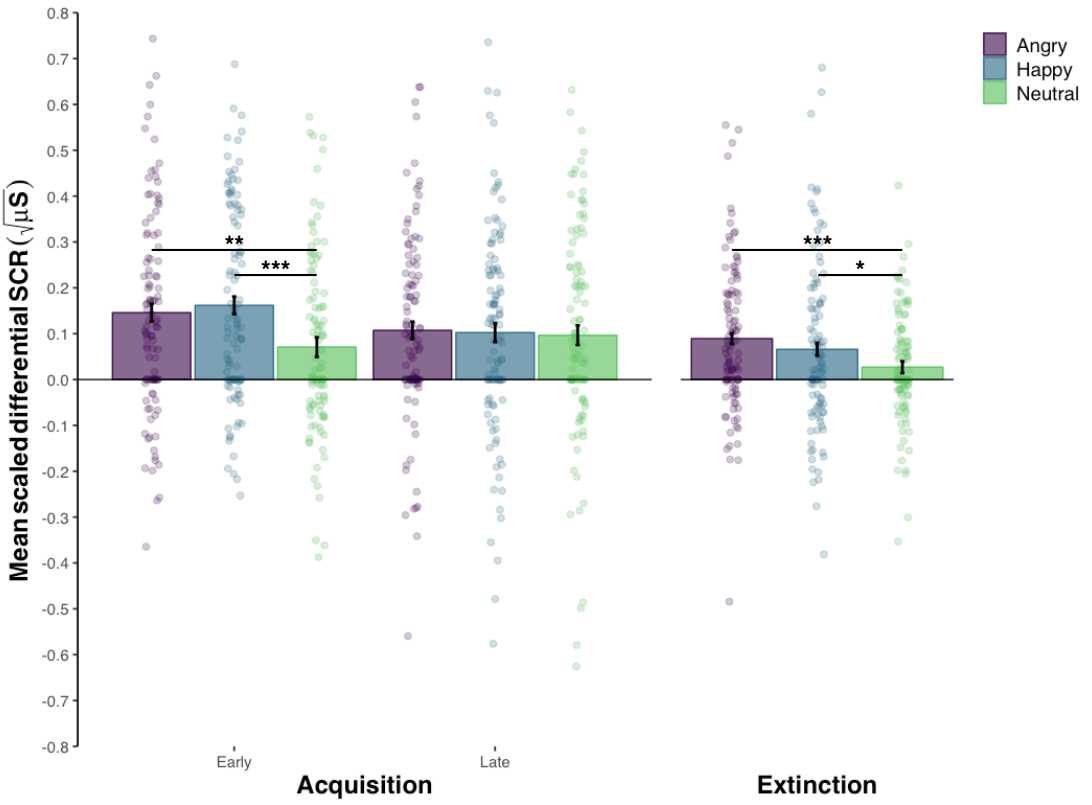

Figure 3.2.3. Mean conditioned response (scaled differential skin conductance response [SCR]) as a function of the conditioned stimulus category (angry vs. happy vs. neutral) during (early and late) acquisition and extinction. The dots indicate data for individual participants. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (***$p < .001$, **$p < .01$, *$p <. 05$, one-tailed, Holm-Bonferroni corrected).

neutral faces. Conversely, the differential reaction time index did not differ statistically across the face categories, $F(2, 212) = 2.45$, $p = .089$, partial $\eta^2 = .023$, 90% CI [.000, .059].

Regression analyses

The multiple linear regression analyses on the CR to happy faces (see Table 3.2.1) showed that participants' extraversion level, differential d' index for happy faces, and differential reaction time index for happy faces did not predict the CR acquisition readiness to happy faces during early acquisition (all *p*s > .34) where they only explained 1.51% of its variance ($R^2 = .015$, 90% CI [.000, .048], adjusted $R^2 = -.014$, $F(3, 103) = 0.53$, $p = .664$). However, these three predictors explained 13.06% of the variance of the CR to happy faces during extinction ($R^2 = .131$, 90% CI [.031, .224], adjusted $R^2 = .105$, $F(3, 103) = 5.16$, $p = .002$). Whereas extraversion and the differential d' index for happy faces did not predict the CR to happy faces (both *p*s > .38), the CR to happy faces was statistically significantly predicted by the differential reaction time index for these faces, $b = 0.002$, $SE = 0.0005$, $\beta = .360$, $t(103) = 3.83$, $p < .001$, reflecting that participants who were faster to associate happy faces with the attribute of importance than that of unimportance exhibited a larger CR to happy faces during extinction (see Figure 3.2.4). No statistically significant relationship was observed between the CR to angry and to neutral faces and participants' extraversion level, differential d' index for angry or neutral faces, and differential reaction time index for angry or neutral faces, respectively, either during early acquisition or extinction (all *p*s > .19; see 3.2.5. Supplementary materials).

3.2.4. Discussion

In this study, we investigated whether learning biases could occur to happy faces in comparison with neutral faces during Pavlovian aversive conditioning. We also aimed at testing whether such learning biases are influenced by inter-individual differences in extraversion and implicit evaluation of happy faces, thus entailing smaller effects than those found for angry faces. We first measured inter-individual differences in extraversion level, as well as in sensitivity and rapidity to associate happy faces with the attribute of importance versus unimportance. We then implemented a Pavlovian differential aversive conditioning paradigm, in which angry, happy, and neutral faces were used as conditioned stimuli.

Table 3.2.1

Results for the multiple linear regression analyses on the conditioned response to happy faces during early acquisition and extinction (N = 107)

	Early acquisition					Extinction				
	b	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (103)	*p*
Intercept	0.073	0.106		0.69	.494	0.027	0.087		0.31	.759
Extraversion	0.003	0.004	.085	0.87	.388	0.002	0.003	.046	0.50	.621
Differential d' index	-0.005	0.039	-.013	-0.13	.896	-0.028	0.032	-.081	-0.87	.386
Differential reaction time index	-0.001	0.001	-.096	-0.96	.341	0.002	0.0005	.360***	3.83	< .001
R^2			.015					.131		

Note. ***p < .001.

Taken together, our results indicate that both angry and happy faces were preferentially associated with an aversive outcome during Pavlovian conditioning relative to neutral faces, with the persistence of this association being somewhat weaker for happy than for angry faces and modulated by inter-individual differences in happy faces' affective evaluation.

The conditioned response to angry and happy faces was more readily acquired and more resistant to extinction than the conditioned response to neutral faces, thus reflecting the occurrence of learning biases to these stimuli. Whereas the greater persistence of the conditioned response to angry faces replicates well-established findings in the human conditioning literature (e.g., Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Mallan et al., 2013; Öhman & Mineka, 2001), the enhanced resistance to extinction of the conditioned response to happy faces aligns well with recent data showing that positive stimuli that are affectively relevant to the organism can likewise produce persistent Pavlovian aversive learning (Stussi, Pourtois, et al., 2018). The faster Pavlovian aversive conditioning to happy faces observed during early acquisition further expands these findings by demonstrating that, similar to threat-relevant stimuli such as angry faces, positive emotional

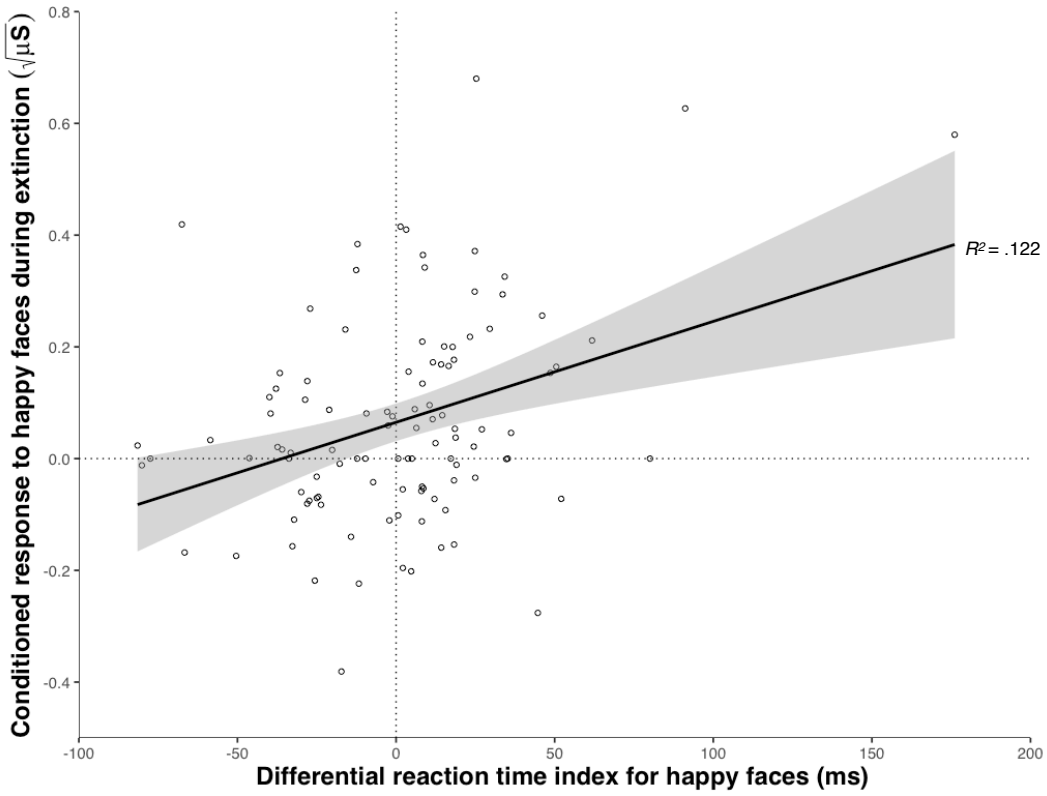

Figure 3.2.4. Relationship between the differential reaction time index for happy faces in the Go/No-go Association Task (mean reaction times in the block where happy faces and the attribute of importance were target categories minus mean reaction times in the block where happy faces and the attribute of unimportance were target categories) and the conditioned response to happy faces during extinction. The line represents the fitted regression line using least squares estimation and 95% confidence interval.

stimuli can also be more readily associated with a naturally aversive event than neutral, less relevant stimuli. In agreement with the affective relevance framework (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018), our study therefore provides additional evidence that preferential Pavlovian aversive learning is not specific to threat-related stimuli, but rather seems to occur to affectively relevant stimuli independently of their valence.

Whereas the effects of accelerated Pavlovian aversive conditioning to happy and angry faces compared with neutral faces during the early acquisition phase were of moderate size, the enhanced resistance to extinction to happy relative to neutral faces was of lesser magnitude than that to angry compared with neutral faces, as reflected by a smaller standardized effect size. This result thereby suggests that the resistance to extinction to happy faces was somewhat less robust than that to angry faces in comparison with neutral faces, which appears consistent with the notion that happy faces hold a general lower level of relevance to the organism than angry faces (Brosch et al., 2008, 2010; Pool et al., 2016). Nonetheless, we observed no

statistically significant difference between the conditioned response to angry faces and that to happy faces during both early acquisition and extinction, the differences between these two emotional categories being of relatively small size. Although this lack of preferential aversive learning for angry compared with happy faces contradicts earlier findings (e.g., Esteves, Parra et al., 1994, Experiment 2; Öhman & Dimberg, 1978, Experiment 2; Rowles et al., 2012), it should be noted that this effect has not been consistently reported in the literature. For instance, some earlier studies showed enhanced resistance to extinction to angry faces but not to happy faces, yet without the difference between angry and happy faces being statistically significant (e.g., Mazurski et al., 1996; Öhman & Dimberg, 1978, Experiment 1). Of importance, our study suggests that the seeming difference in enhanced aversive learning between angry and happy faces may reflect that the persistence of the conditioned response to happy – but not to angry – faces was related to inter-individual differences in their affective evaluation.

The fact that happy faces led to a relatively small learning bias during extinction could potentially account for failures to report a resistance-to-extinction effect for this specific emotional category in previous studies (see, e.g., Bramwell et al., 2014; Esteves, Parra, et al., 1994; Mazurski et al., 1996; Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Öhman & Mineka, 2001). Past studies have generally used between-participant designs (but see Bramwell et al., 2014) that are less sensitive than within-participant designs (see, e.g., Ho & Lipp, 2014), and importantly, often with modest sample sizes, typically varying from 15 to 25 participants by group. These two methodological factors likely contributed to hindering the possibility to reveal the existence of learning biases to happy faces in these previous studies given that, as our results suggest here with the use of a larger sample and stringent within-participant design, this bias has a small effect size. Additional post-hoc power analyses corroborated this assumption. They showed that achieved power to detect a small effect as reported in the present study ($g_{av} = 0.247$) using a one-tailed *t* test and an alpha level of .05 with a sample size ranging from 15 to 25 participants would vary between 23.14% and 32.83% for a within-participant design, and between 16.24% and 21.66% for a between-participant design. Given the relatively small effect size of the learning bias to happy faces during extinction, it is therefore highly desirable in future research to set up adequately-powered experiments when the goal is to explore differences in Pavlovian aversive learning to happy compared with neutral or angry faces.

Another potential explanation for the discrepancy between our study and previous findings available in the literature (e.g., Bramwell et al., 2014; Esteves, Parra, et al., 1994;

Mazurski et al., 1996; Öhman & Dimberg, 1978; Rowles et al., 2012; see also Dimberg & Öhman, 1996; Öhman & Mineka, 2001) could be related to other methodological factors, including the actual procedure used to establish Pavlovian aversive conditioning, and in particular, the distinction between a partial versus continuous (full) reinforcement schedule. Continuous reinforcement schedules have been reported to entail ceiling effects in the conditioned response acquisition readiness, thereby possibly masking the occurrence of differences between the emotional stimulus categories (Ho & Lipp, 2014; Lissek, Pine, & Grillon, 2006). In addition, continuous reinforcement schedules usually lead to faster extinction of the conditioned response than partial reinforcement schedules (e.g., Grady, Bowen, Hyde, Totsch, & Knight, 2016; Jenkins & Stanley, 1950). In light of these considerations, it is conceivable that the use of a full reinforcement schedule, as mostly employed in these previous studies, substantially facilitated the extinction of the conditioned response to happy faces, precluding in turn the detection of a learning bias for this specific emotion category at the statistical level.

Additional computational analyses (see 3.2.5. Supplementary materials) further revealed that the effects of enhanced resistance to extinction to angry and happy faces compared with neutral faces were characterized by lower learning rates for negative predictions errors (i.e., when the actual outcome is omitted or less than expected; Nv & Schoenbaum, 2008), thereby replicating and extending recent findings (Stussi, Pourtois, et al., 2018). More specifically, the learning rate for negative prediction errors to angry faces was lower than that to happy and neutral faces, and the learning rate for negative prediction errors to happy faces was lower than that to neutral faces, though the latter difference was only marginally significant after correction for multiple testing. This result suggests that the greater persistence of the conditioned response to angry and happy faces compared with neutral faces during extinction may have been underlain by a diminished impact of negative prediction errors to these stimuli (Stussi, Pourtois, et al., 2018). This lower impact of negative prediction errors likely contributed to weakening inhibitory learning underlying extinction (Dunsmoor, Niv, Daw, & Phelps, 2015). Of note, the lower learning rate for negative prediction errors to angry faces relative to happy faces also suggests that angry faces led to more resistant-to-extinction Pavlovian aversive conditioning than happy faces, even though this difference was not visible at the level of the skin conductance response. Furthermore, whereas the size of the difference in estimated inhibitory learning rates between angry and neutral faces was moderate, it was relatively small for angry relative to happy faces, and for happy compared with neutral faces,

thus mirroring the skin conductance response data. On the other hand, we did not find evidence that the faster acquisition of the conditioned response to angry and happy faces relative to neutral faces was driven by higher learning rates for positive prediction errors. This negative finding was possibly due to habituation effects in the skin conductance response affecting the conditioned response magnitude. This may have in turn biased the estimation of the excitatory learning rates, and mitigated the emergence of differences between the three emotional stimulus categories.

At odds with our predictions, we did not observe a modulatory effect of participants' extraversion level on the conditioned response during early acquisition and extinction. This null result provides no evidence that inter-individual differences in extraversion influenced the conditioned response readiness and persistence to happy faces. Given the relative homogeneous distribution of extraversion scores in the current sample (see Figure S3.2.1) relative to normative data from a similar student population (Rolland et al., 1998), additional studies based on large samples and wider ranges of extraversion level across individuals might be required to assess more appropriately and bring more conclusive evidence regarding the impact of this specific variable on Pavlovian aversive learning to happy faces.

Unlike in early acquisition, inter-individual differences in happy faces' affective evaluation were found to influence the resistance to extinction to them, as reflected by a greater persistence of the conditioned response to happy faces during extinction in participants who were faster to associate happy faces with the attribute of importance versus unimportance. By comparison, no such relationship was found for angry and neutral faces (see 3.2.5. Supplementary materials). In line with the affective relevance framework (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018), this result shows that the way happy faces are evaluated modulates the resistance to extinction to them. A caveat, however, is that the Go/No-go Association Task that we used probably did not provide a direct and pure measure of the affective relevance or importance value of the three different face categories used in our study. Results of this task showed that participants more easily associated happy faces with the attribute of importance versus unimportance than both angry and neutral faces, whereas no statistically significant difference was found between angry and neutral faces. This suggests the Go/No-go Association Task likely captured the stimuli's valence rather than their relevance, and may have reflected participants' implicit preferences or liking toward the face categories (see Nosek & Banaji, 2001). Accordingly, it is possible that differential implicit preferences toward happy faces actually drove the resistance to extinction seen for these faces

in the present study. Because we only found a clear relationship between the conditioned response to happy faces during extinction and the differential reaction time index but not with the differential d' index to happy faces, caution is needed in the interpretation of this relationship, and these findings await replication in future studies before stronger conclusions might be drawn.

As angry and happy faces are usually considered as more arousing than neutral faces, it could be argued that the enhanced Pavlovian aversive conditioning to them resulted from their higher arousal value rather than, or in addition to, their affective relevance. Appraisal theories (e.g., Sander et al., 2003, 2005, 2018) suggest that stimuli appraised as relevant to the organism's concerns often trigger a physiological state of arousal that can be felt consciously as a consequence of the elicitation of a motivational state (see Montagrin & Sander, 2016; Pool et al., 2016). Although we cannot completely rule out that arousal contributed to our findings, it seems unlikely that they were solely determined by felt and/or physiological arousal (see Stussi, Pourtois, et al., 2018, for a related discussion). In fact, previous studies (Hamm, Greenwald, Bradley, & Lang, 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) have reported that highly arousing negative and positive stimuli, without taking into account their affective relevance to the organism, did not produce preferential Pavlovian aversive conditioning relative to less arousing stimuli. Moreover, supplementary analysis of the habituation phase[11] revealed that (a) angry faces elicited larger skin conductance responses than happy faces before conditioning, whereas no difference emerged between angry and neutral faces, and between happy and neutral faces, and (b) the skin conductance responses to the various face categories during habituation did not correlate with the conditioned response to these stimuli during early acquisition and extinction. These considerations suggest that an explanation in terms of arousal alone does not satisfactorily account for the occurrence of differential learning biases to both angry and happy faces.

[11] A repeated-measures ANOVA with CS type (CS+ vs. CS-) and CS category (angry vs. happy vs. neutral) as within-participant factors performed on the skin conductance response data during habituation revealed a statistically significant main effect of CS category, $F(2, 212) = 4.20$, $p = .016$, partial $\eta^2 = .038$, 90% CI [.004, .083]. Further post-hoc comparisons using Tukey's HSD tests indicated that angry faces elicited larger skin conductance responses than happy faces ($p = .012$, $g_{av} = 0.215$, 95% CI [0.064, 0.369]), whereas no statistically significant difference was found between angry and neutral faces ($p = .190$, $g_{av} = 0.129$, 95% CI [-0.019, 0.279]) or between happy and neutral faces ($p = .497$, $g_{av} = -0.088$, 95% CI [-0.239, 0.062]). Pearson's correlation analyses moreover showed no statistically significant relationship between the skin conductance responses to the different faces during habituation and the conditioned response to these faces during the early acquisition phase ($-.129 <$ all rs$(105) < .100$, all ps $> .18$) or during the extinction phase ($.001 <$ all rs$(105) < .129$, all ps $> .18$).

Importantly, the results reported in this study lend support to the view that enhanced Pavlovian aversive conditioning depends on the stimulus' affective relevance to the organism. These findings align with the affective relevance framework according to which preferential Pavlovian aversive learning is underlain by a general and flexible mechanism that is shared across affectively relevant stimuli independently of their valence (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). Alternatively, our results could also be interpreted as reflecting the involvement of two different mechanisms: a specialized mechanism selectively acting on threat-related stimuli that is consistently engaged across individuals, and a more general one acting on affectively relevant stimuli that is more sensitive to individual differences. Future research should therefore aim to disentangle these two competing explanations by investigating whether learning biases in Pavlovian aversive conditioning occurring in response to threat-relevant stimuli are underpinned by a threat-specific mechanism that is functionally distinct from a mechanism of affective relevance, for instance through the use of neuroimaging techniques.

In conclusion, the present study highlights that, similar to angry faces, happy faces can lead to faster and more persistent Pavlovian aversive conditioning, with the effects of resistance to extinction to happy faces being smaller than to angry faces, and modulated by inter-individual differences in their affective evaluation. These findings replicate and extend recent work (Stussi, Pourtois, et al., 2018) by showing that learning biases in Pavlovian aversive conditioning are not specific to threat-relevant stimuli, but can also emerge in response to positive affectively relevant stimuli. They additionally suggest that inter-individual differences can play a key role in the occurrence of these learning biases (Stussi et al., in press). Accordingly, our study contributes to further elucidating the basic mechanisms underlying Pavlovian aversive learning in humans, and could ultimately provide insights into the understanding of impairments in this process that are typically associated with specific emotional disorders, including anxiety or phobia.

3.2.5. Supplementary materials

Supplementary method and results

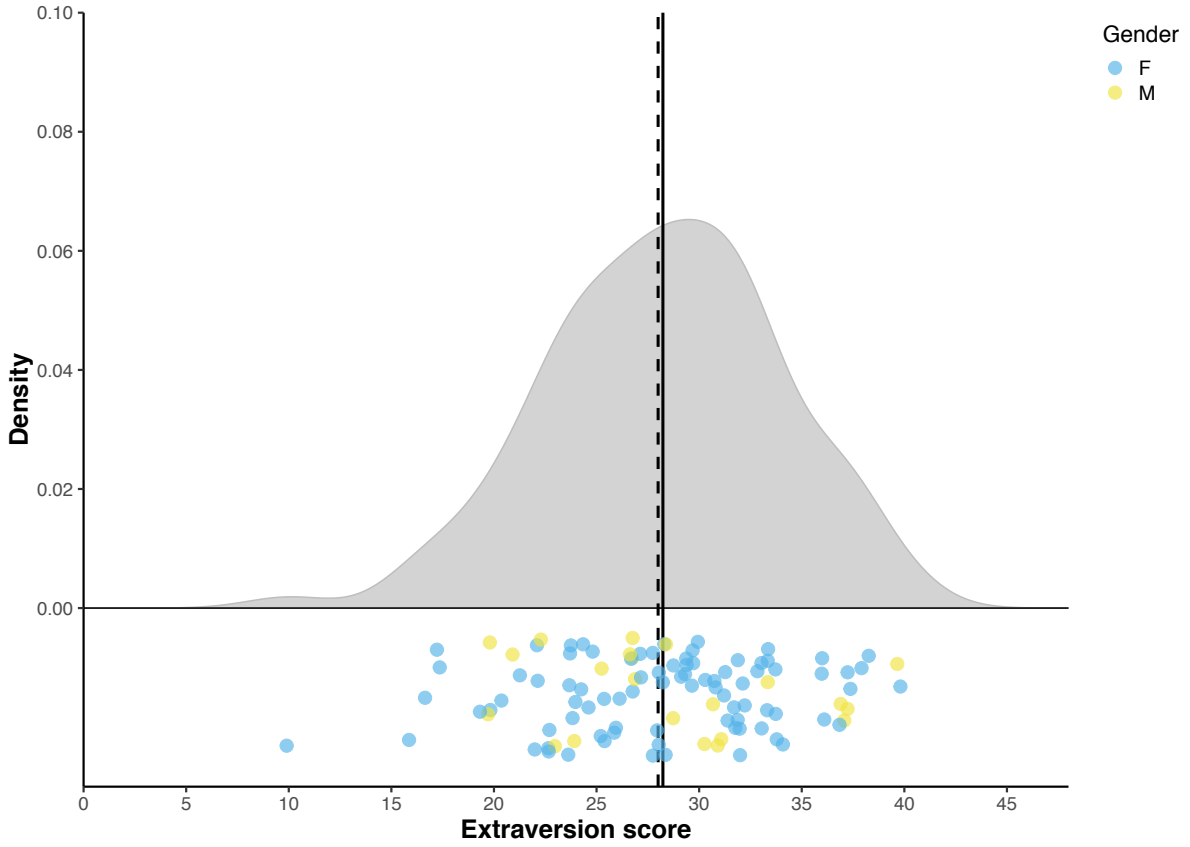

Figure S3.2.1. Distribution of extraversion scores as measured with the NEO-FFI (Costa & McCrae, 1992; Rolland, Parker, & Strumpf, 1998). The dots indicate data for individual participants. The solid line indicates the mean extraversion score, and the dashed line the median extraversion score.

Go/No-go Association Task

Face stimuli from the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) were used either as targets or as distractors in the Go/No-go Association Task (GNAT; Nosek & Banaji, 2001). They consisted of eight angry faces (model numbers for targets: AM05ANS, AM09ANS, AM17ANS, AM30ANS; model numbers for distractors: AM14ANS, AM19ANS, AM21ANS, AM24ANS), eight happy faces (model numbers for targets: AM20HAS, AM23HAS, AM25HAS, AM26HAS; model number for distractors: AM04HAS, AM12HAS, AM16HAS, AM32HAS), and eight neutral faces (models numbers for targets: AM01NES, AM06NES, AM08NES, AM13NES; model numbers for distractors: AM02NES, AM18NES, AM28NES, AM35NES).

The practice session of the GNAT included five blocks in which there was only a single target category. In the first three blocks, participants learned to discriminate between the different face categories, each of them being the target category in one of the blocks whereas the two other categories were distractors; the order being counterbalanced across participants. In these blocks, four faces from the target face category and two faces from each distractor face category were each presented twice in a pseudorandom order. In the last two practice blocks, participants were presented with the four "important" and the four "unimportant" words, which were each presented twice in a pseudorandom order. The "important" words were targets and the "unimportant" words distractors in the fourth block, which was reversed in the last block. Each practice block consisted of 16 trials and used a 666-ms response deadline.

Subjective ratings

Subsequent to the GNAT but before the differential Pavlovian aversive conditioning procedure, participants provided subjective ratings of the two angry face conditioned stimuli (CSs), the two happy face CSs, and the two neutral face CSs as a function of their pleasantness, subjective arousal, and subjective relevance. In this procedure, the faces were presented to participants along with a visual analog scale (VAS). For the pleasantness ratings, participants were asked to rate the degree to which the face was unpleasant or pleasant from 0 (*very unpleasant*) to 100 (*very pleasant*). For the arousal ratings, they were asked to rate the degree to which the face was arousing from 0 (*not at all arousing*) to 100 (*very arousing*). For the relevance ratings, participants were asked to rate the degree to which the face was important to them from 0 (*not at all important*) to 100 (*very important*). After the end of the conditioning procedure, participants completed again pleasantness, arousal, and relevance ratings of the CSs using the same procedure as for the preconditioning ratings. In addition, they were asked to rate how many electric stimulations they received to each CS on a Likert scale from 0 to 9 to assess their explicit awareness of the CS-US contingencies. The order of the CS presentations and the questions was randomized between participants for both the preconditioning and postconditioning ratings.

The pleasantness, arousal, and relevance ratings were analyzed with separate three-way repeated-measures analyses of variance (ANOVAs) with time (pre vs. post), CS category (angry vs. happy vs. neutral), and CS type (CS+ vs. CS-) as within-participant factors, whereas the CS-US contingency ratings were analyzed with a two-way repeated-measures ANOVA with CS category (anger vs. happy vs. neutral) and CS type (CS+ vs. CS-) as within-participant

factors. Statistically significant effects were followed up with more focused repeated-measures ANOVAs and/or a multiple comparison procedure using Tukey's HSD tests when applicable.

Analysis of the pleasantness ratings (see Figure S3.2.2a) showed a statistically significant three-way interaction between time, CS category, and CS type, $F(2, 212) = 5.29$, $p = .006$, partial $\eta^2 = .048$, 90% CI [.008, .096]. A follow-up 3 (CS category: angry vs. happy vs. neutral) × 2 (CS type: CS+ vs. CS-) repeated-measures ANOVA for the preconditioning ratings indicated that the CS categories modulated the CSs' rated pleasantness before conditioning, $F(1.73, 183.12) = 323.15$, $p < .001$, partial $\eta^2 = .753$, 90% CI [.707, .785]. As expected, happy faces were deemed more pleasant than angry faces ($p < .001$, $g_{av} = 2.887$, 95% CI [2.432, 3.378]) and neutral faces ($p < .001$, $g_{av} = 1.438$, 95% CI [1.149, 1.743]), whereas neutral faces were evaluated as more pleasant than angry faces ($p < .001$, $g_{av} = 1.933$, 95% CI [1.590, 2.298]). The follow-up repeated-measures ANOVA for the postconditioning ratings revealed a statistically significant interaction between CS category and CS type, $F(2, 212) = 6.40$, $p = .002$, partial $\eta^2 = .057$, 90% CI [.013, .109], reflecting that the difference in rated pleasantness between the CS+ and the CS- was higher for happy faces than angry and neutral faces. The CS+ was evaluated as less pleasant than the CS- for angry faces ($p = .015$, $g_{av} = 0.404$, 95% CI [0.160, 0.652]), happy faces ($p < .001$, $g_{av} = 0.811$, 95% CI [0.546, 1.085]), and neutral faces ($p < .001$, $g_{av} = 0.662$, 95% CI [0.401, 0.929]). Furthermore, happy faces were rated as more pleasant than angry faces (all ps $< .001$, $0.80 < g_{av}$s < 2.26), and neutral faces were deemed more pleasant than angry faces (all ps $< .04$, $0.39 < g_{av}$s < 1.59). The happy face CS- was likewise evaluated as more pleasant than the neutral face CS+ and CS- ($p < .001$, $g_{av} = 1.462$, 95% CI [1.133, 1.808], and $p < .001$, $g_{av} = 0.950$, 95% CI [0.669, 1.241], respectively), and the happy face CS+ as more pleasant than the neutral face CS+ ($p < .001$, $g_{av} = 0.479$, 95% CI [0.253, 0.710]), whereas there was no statistical difference in rated pleasantness between the happy face CS+ and the neutral face CS- ($p = .999$, $g_{av} = -0.044$, 95% CI [-0.326, 0.238]).

The arousal ratings analysis (see Figure S3.2.2b) revealed a statistically significant interaction between time and CS type, $F(1, 106) = 87.23$, $p < .001$, partial $\eta^2 = .451$, 90% CI [.335, .541]. Before conditioning, the CSs+ and the CSs- did not statistically differ in felt arousal ($p > .99$, $g_{av} = 0.004$, 95% CI [-0.155, 0.163]); by contrast, the CSs+ were rated as more arousing than the CSs- after conditioning ($p < .001$, $g_{av} = 1.149$, 95% CI [0.874, 1.436]). The CSs+ were also deemed more arousing after conditioning than before it ($p < .001$, $g_{av} = 0.872$, 95% CI [0.645, 1.108]), whereas the CSs- were deemed less arousing after than before conditioning ($p < .001$, $g_{av} = 0.382$, 95% CI [0.188, 0.581]). Moreover, the interaction between

time and CS category yielded statistical significance, $F(2, 212) = 22.81$, $p < .001$, partial $\eta^2 = .177$, 90% CI [.101, .248]. Angry and happy faces were evaluated as more arousing than neutral faces both in the preconditioning and postconditioning ratings (all ps $< .001$, $0.86 < g_{av}$s < 1.86), but did not differ statistically between each other (all ps $> .55$, $0.03 < g_{av} < 0.18$). In addition, neutral faces were evaluated as more arousing after than before conditioning ($p < .001$, $g_{av} = 0.837$, 95% CI [0.559, 1.123]), which was not the case for angry faces ($p = .949$, $g_{av} = 0.070$, 95% CI [-0.086, 0.226]) and happy faces ($p = .953$, $g_{av} = -0.077$, 95% CI [-0.253, 0.098]).

For the relevance ratings (see Figure S3.2.2c), the analysis showed a statistically significant interaction effect of time and CS type, $F(1, 106) = 38.56$, $p < .001$, partial $\eta^2 = .267$, 90% CI [.153, .371]. While there was no statistical difference in relevance ratings between the CSs+ and the CSs- prior to conditioning ($p = .843$, $g_{av} = 0.050$, 95% CI [-0.070, 0.171]), the CSs+ were deemed more relevant than the CSs- after conditioning ($p < .001$, $g_{av} = 0.786$, 95% CI [0.539, 1.042]). Furthermore, the CSs+ were rated as more relevant after than before conditioning ($p < .001$, $g_{av} = 0.627$, 95% CI [0.421, 0.840]), which was not the case for the CSs- ($p = .489$, $g_{av} = -0.133$, 95% CI [-0.319, 0.052]). We also observed a statistically significant interaction between time and CS category, $F(2, 212) = 28.41$, $p < .001$, partial $\eta^2 = .211$, 90% CI [.131, .283]. Happy faces were evaluated as more relevant than angry and neutral faces both before and after conditioning (all ps $< .001$, $0.57 < g_{av}$s < 1.82), and angry faces as more relevant than neutral faces (all ps $< .003$, $0.39 < g_{av}$s < 0.98). Neutral faces were additionally rated as higher in relevance after conditioning relative to before conditioning ($p < .001$, $g_{av} = 0.897$, 95% CI [0.633, 1.171]), whereas there was no statistical difference in preconditioning and postconditioning relevance ratings for angry faces ($p = .876$, $g_{av} = 0.086$, 95% CI [-0.067, 0.240]) and happy faces ($p = .786$, $g_{av} = -0.110$, 95% CI [-0.278, 0.057]).

The postconditioning ratings of CS-US contingency (see Figure S3.2.2d) revealed a statistically significant interaction between the CS categories and the CS types, $F(2, 212) = 3.35$, $p = .037$, partial $\eta^2 = .031$, 90% CI [.001, .072]. Follow-up analyses indicated that the CS+ was rated to be associated with the delivery of more electric stimulations than the CS- for angry ($p < .001$, $g_{av} = 1.876$, 95% CI [1.495, 2.278]), happy ($p < .001$, $g_{av} = 2.345$, 95% CI [1.933, 2.784]), and neutral ($p < .001$, $g_{av} = 1.817$, 95% CI [1.468, 2.188]) faces. In addition, participants evaluated the happy face CS+ as paired with more electric stimulations than the neutral face CS+ ($p = .010$, $g_{av} = 0.369$, 95% CI [0.155, 0.587]), and the angry face CS+ as marginally more likely to be associated with more electric stimulations than the neutral face

CS+ (p = .087, g_{av} = 0.312, 95% CI [0.080, 0.547]). By contrast, no difference was found between the angry and the happy face CSs+ (p = .985, g_{av} = 0.071, 95% CI [-0.136, 0.278]) or between the CSs- among the three CS categories (all ps > .16, 0.03 < g_{av}s < 0.25).

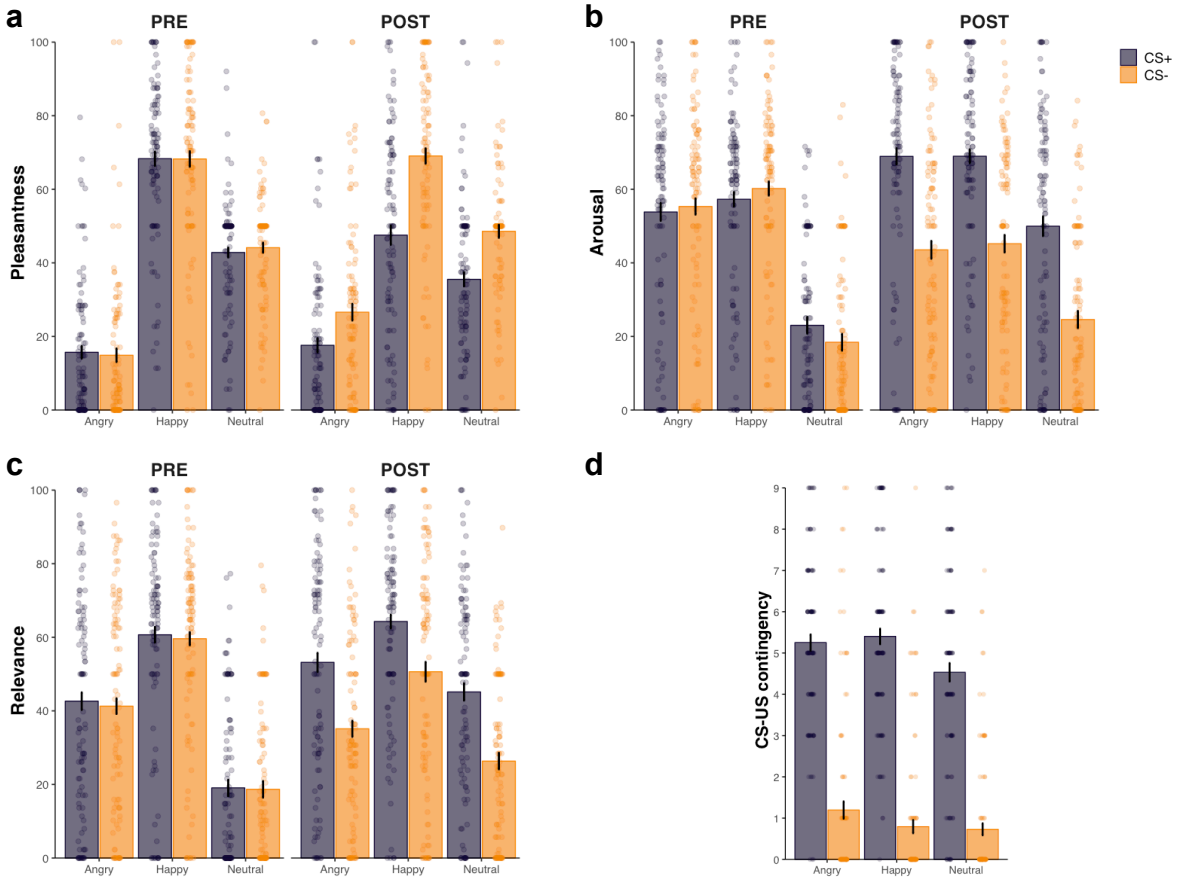

Figure S3.2.2. Subjective ratings before (pre) and after (post) the conditioning procedure as a function of conditioned stimulus type (CS+ vs. CS-) and stimulus category (angry vs. happy vs. neutral). Mean (a) pleasantness ratings, (b) arousal ratings, (c) relevance ratings, and (d) CS-US contingency ratings. The dots indicate data for individual participants. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008).

Computational modeling

To characterize and provide insights into the computations underlying the influence of angry and happy faces, relative to neutral faces, on Pavlovian aversive conditioning, we constructed simple reinforcement learning models (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see also Stussi, Pourtois, & Sander, 2018) and fitted them to the skin conductance response (SCR) data for each CS category separately in order to estimate the models' free parameters and to identify the best-fitting

model. After selection of the best-fitting model, its parameter estimates were subsequently compared across angry, happy, and neutral face CSs. We considered the following models.

Rescorla-Wagner model. According to the Rescorla-Wagner model (Rescorla & Wagner, 1972), learning occurs when events deviate from expectations and correspondingly serves to update future expectations (Niv & Schoenbaum, 2008). It formalizes the notion of prediction error by stating that associative learning is directly driven by the discrepancy between the actual and the expected outcome. In this model, the predictive value (or associative strength) V at trial $t + 1$ of a given CS j is updated on the basis of the sum of the current predictive value V_j at trial t and the prediction error between the predictive value V_j and the outcome R at trial t, weighted by a constant learning rate α:

$$V_j(t+1) = V_j(t) + \alpha \cdot (R(t) - V_j(t))$$

where the learning rate α is a free parameter within the range [0, 1]. If the unconditioned stimulus (US) was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Hybrid model. In addition to maintaining the basic assumption that learning is directly driven by prediction errors as stated in the Rescorla-Wagner model, the hybrid model proposed by Li et al. (2011) incorporates the Pearce-Hall associability mechanism (Pearce & Hall, 1980). The Pearce-Hall model specifically asserts that the CS associability determines the learning rate and is dynamically modulated on each trial as a function of unsigned past prediction errors. According to the Pearce-Hall algorithm, the CS associability decreases when the CS accurately and reliably predicts the actual outcome, whereas it increases when the CS is an unreliable predictor of the actual outcome. In the hybrid model, the predictive value V and the associability α of a given CS j are updated as follows:

$$V_j(t+1) = V_j(t) + \kappa \cdot \alpha_j(t) \cdot (R(t) - V_j(t))$$

$$\alpha_j(t+1) = \eta \cdot \left| R(t) - V_j(t) \right| + (1 - \eta) \cdot \alpha_j(t)$$

where the initial associability α_0, the learning rate κ, and the weighting factor η are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Rescorla-Wagner with dual learning rates. We additionally modified the standard version of the Rescorla-Wagner model by implementing distinct learning rates for positive (i.e., excitatory learning) and negative (i.e., inhibitory learning) prediction errors instead of a single

learning rate (see Niv, Edlund, Dayan, & O'Doherty, 2012; Stussi, Pourtois, et al., 2018). In the dual-learning-rate Rescorla-Wagner model, the predictive value V of a given CS j is updated based on the sum of the current predictive value V_j at trial t, and the prediction error between the predictive value V_j and the outcome R at trial t, weighted by different learning rates for positive and negative prediction errors as follows:

$$V_j(t+1) = \begin{cases} V_j(t) + \alpha^+ \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) > 0 \\ V_j(t) + \alpha^- \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) < 0 \end{cases}$$

where the learning rate for positive prediction errors α^+ and the learning rate for negative prediction errors α^- are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$. This model allows for parsimoniously accounting for how specific categories of stimuli can both accelerate acquisition (through the learning rate for positive prediction errors) and enhance resistance to extinction (through the learning rate for negative prediction errors) of the conditioned response (CR).

Hybrid model with dual learning rates. Similar to the modified Rescorla-Wagner model implementing different learning rates for positive and negative prediction errors, we also constructed a modified hybrid model with dual learning rates. In this modified version of the hybrid model, the predictive value V and the associability α of a given CS j are updated as follows:

$$V_j(t+1) = \begin{cases} V_j(t) + \kappa^+ \cdot \alpha_j(t) \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) > 0 \\ V_j(t) + \kappa^- \cdot \alpha_j(t) \cdot (R(t) - V_j(t)) \; if \; R(t) - V_j(t) < 0 \end{cases}$$

$$\alpha_j(t+1) = \eta \cdot \left| R(t) - V_j(t) \right| + (1 - \eta) \cdot \alpha_j(t)$$

where the initial associability α_0, the learning rate for positive prediction errors κ^+, the learning rate for negative prediction errors κ^-, and the weighting factor η are free parameters within the range [0, 1]. If the US was delivered on the current trial t, $R(t) = 1$, else $R(t) = 0$.

Model and parameter fitting. We fitted and optimized the models' free parameters using maximum a posteriori estimation, which consisted in finding the set of parameters maximizing the likelihood of each participant's trial-by-trial normalized (i.e., scaled and square-root-transformed) SCRs to the CS given the model, constrained by a regularizing prior (Gershman, 2016; Niv et al., 2012). The free parameters were constrained with a Beta (1.2,

1.2) prior distribution that favors a normal distribution of the estimated parameters. We used the trial-by-trial timeseries of CS predictive values $V(t)$ to optimize the free parameters for the Rescorla-Wagner model (RW[V]) and the modified Rescorla-Wagner model with dual learning rates (dual RW[V]). For the hybrid model and the hybrid model with dual learning rates, we optimized the free parameters separately for each possible combination based on the trial-by-trial timeseries of CS values $V(t)$ (Hybrid[V] and dual Hybrid[V]), the trial-by-trial timeseries of CS associabilities $\alpha(t)$ (Hybrid[α] and dual Hybrid[α]), or the combination of both (Hybrid[$V+\alpha$] and dual Hybrid[$V+\alpha$]; see Li et al., 2011; Zhang, Mano, Ganesh, Robbins, & Seymour, 2016). Given that participants were expecting to receive electric stimulations at the outset of the Pavlovian aversive conditioning procedure because of the work-up procedure and the instructions, we set each CS initial predictive value V_0 to 0.5. We fitted the various models using a separate set of free parameters for each participant (a) across all trials, and (b) separately for each CS category (Boll, Gamer, Gluth, Finsterbusch, & Büchel, 2013). This allowed for comparing the parameter estimates that best fitted to the SCR data between the three different CS categories. Two participants were excluded from the computational analyses because their individual parameters could not be estimated due to a lack of SCR to all the angry face CSs during the experiment. The final sample size for the computational analyses included 105 participants (83 women, 22 men; mean age = 21.79 ± 2.46 years).

Model comparison. We performed model comparison with the Bayesian information criterion (BIC; Schwarz, 1978; see also Stussi, Pourtois, et al., 2018; Zhang et al., 2016). In addition to providing a quantitative measure of the models' goodness of fit, the BIC considers and penalizes for the number of free parameters that the model includes. For each model, the mean BIC value was computed using the average of individual participant's estimated parameters. The models were additionally compared against a random model, in which the predictive value $V_j(t)$ and the prediction errors were updated at each trial by adding random noise from a uniform random distribution within the range [-0.1, 0.1] (Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013). This allowed us to confirm that the reinforcement learning models that we used outperformed a model implementing random predictions. The mean BIC values for each model are reported in Table S3.2.1.

Table S3.2.1

Goodness of fit to skin conductance responses for individual models using the mean Bayesian information criterion (N = 105)

CS category	Model								
	RW(V)	Dual RW(V)	Hybrid (V)	Hybrid (α)	Hybrid ($V + \alpha$)	Dual Hybrid (V)	Dual Hybrid (α)	Dual Hybrid ($V + \alpha$)	Random
All	-0.48	-4.79	6.39	-1.06	-1.20	2.15	1.95	1.92	13.17
Angry	-0.31	-1.11	5.47	1.14	1.08	4.38	3.14	3.64	7.31
Happy	-1.52	-2.94	4.43	0.06	0.10	2.92	1.93	2.18	5.17
Neutral	-3.83	-4.40	2.27	-2.00	-2.20	0.99	0.28	0.52	2.09

Note. RW = Rescorla-Wagner model, V = predictive values, α = associabilities, Dual = dual-learning-rate.

Relationship between modeled learning signals and participants' normalized skin conductance responses. We further assessed whether, and the extent to which, modeled predictive value and prediction error signals from the best-fitting model (i.e., the dual-learning-rate Rescorla-Wagner model; see Table S3.2.1) were predictive of the participants' trial-by-trial normalized SCRs (see Li et al., 2011; Pauli et al., 2015). To do so, we performed a multiple linear regression in which we regressed predictive value and prediction error timeseries generated with the individual parameter estimates from the dual-learning-rate Rescorla-Wagner model and averaged across participants against the averaged trial-by-trial normalized SCRs. This analysis revealed that predictive value and prediction error signals explained a statistically significant portion of variance of trial-by-trial normalized SCRs (R^2 = .580, 90% CI [.450, .677], adjusted R^2 = .570, $F(2, 87) = 60.11$, $p < .001$). Predictive value signals statistically significantly predicted trial-by-trial normalized SCRs, b = 0.385, SE = 0.035, β = .763, $t(87) = 10.96$, $p < .001$ (see Figure S3.2.3), which was not the case for prediction error signals, b = 0.013, SE = 0.021, β = .043, $t(87) = 0.62$, p = .538.

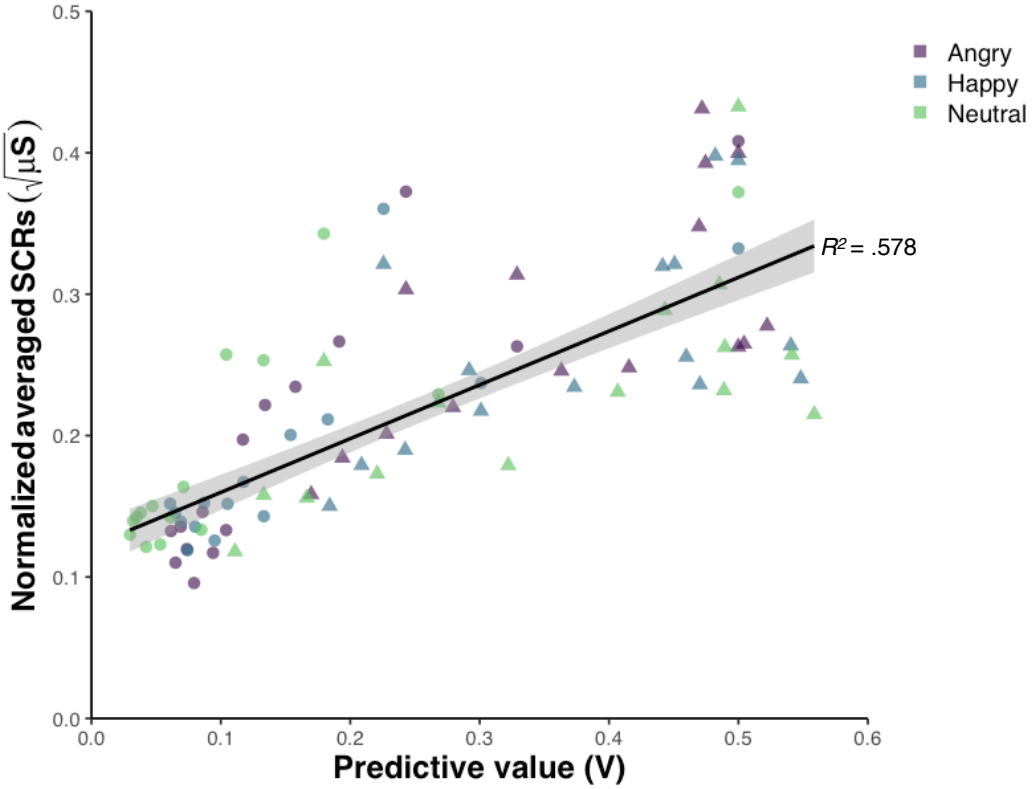

Figure S3.2.3. Relationship between modeled predictive values (*V*) and trial-by-trial normalized skin conductance responses (SCRs) averaged across participants using the individual best-fitting parameters for the Rescorla-Wagner model implementing dual learning rates. Triangles represent reinforced conditioned stimuli (CSs+) and circles represent unreinforced conditioned stimuli (CSs-). The line represents the fitted regression line using least squares estimation and 95% confidence interval.

Parameter estimates analysis. Given that model comparison revealed that the dual-learning-rate Rescorla-Wagner model provided the best fit to the SCR data (see Table S3.2.1), we accordingly compared the estimated learning-rate parameters for both positive and negative prediction errors across the three different CS categories used. A one-way repeated-measures ANOVA with CS category (angry vs. happy vs. neutral) as a within-participant factor on the learning-rate parameter estimates for positive prediction errors revealed no statistically significant difference among the three CS categories, $F(2, 208) = 1.42$, $p = .244$, partial $\eta^2 = .013$, 90% CI [.000, .044] (see Figure S3.3.4). This therefore provides no evidence that the accelerated Pavlovian aversive conditioning to angry and happy faces compared with neutral faces was driven by higher learning rates for positive prediction errors. Inversely, the CS categories differentially affected the learning-rate estimates for negative prediction errors, $F(2, 208) = 6.80$, $p = .001$, partial $\eta^2 = .061$, 90% CI [.015, .115]. Planned contrast analyses showed that these estimates were lower for angry faces (contrast weight: -1) than for neutral faces (contrast weight: +1), $t(104) = 3.83$, $p < .001$ (one-tailed), $g_{av} = 0.480$, 95% CI [0.227, 0.738], $BF_{10} = 208.192$, and happy faces (contrast weight: +1), $t(104) = 1.81$, $p = .036$ (one-tailed), $g_{av} = 0.206$, 95% CI [-0.019, 0.434], $BF_{10} = 1.331$, whereas they were marginally lower for happy faces (contrast weight: -1) compared with neutral faces (contrast weight: +1) with respect to the corrected alpha level for this contrast ($\alpha = .025$) using the Holm-Bonferroni procedure (Holm, 1979), $t(104) = 1.81$, $p = .036$ (one-tailed), $g_{av} = 0.230$, 95% CI [-0.021, 0.483], $BF_{10} = 2.030$ (see Figure S3.2.4). In line with recent findings (Stussi, Pourtois, et al., 2018), these results suggest that the enhanced resistance to extinction of the conditioned response to angry and happy faces compared with neutral faces was characterized by lower inhibitory learning rates that decreased the impact of negative prediction errors on associative learning.

In addition, we conducted multiple linear regressions to test whether the learning-rate estimates for positive and negative prediction errors to happy faces were predicted by participants' (a) extraversion level, (b) differential d' index for happy faces, and (c) differential reaction time index for happy faces. These analyses (see Table S3.2.2) indicated that participants' extraversion level, differential d' index for happy faces, and differential reaction time index for happy faces explained 5.31% ($R^2 = .053$, 90% CI [.000, .120], adjusted $R^2 = .025$, $F(3, 101) = 1.89$, $p = .136$) and 4.58% ($R^2 = .045$, 90% CI [.000, .108], adjusted $R^2 = .017$, $F(3, 101) = 1.62$, $p = .190$) of the variance of the estimated learning rates for positive and negative prediction errors, respectively. Extraversion was a marginally significant predictor of the learning-rate estimates for positive prediction errors, $b = 0.011$, $SE = 0.006$, $\beta = .179$, $t(101)$

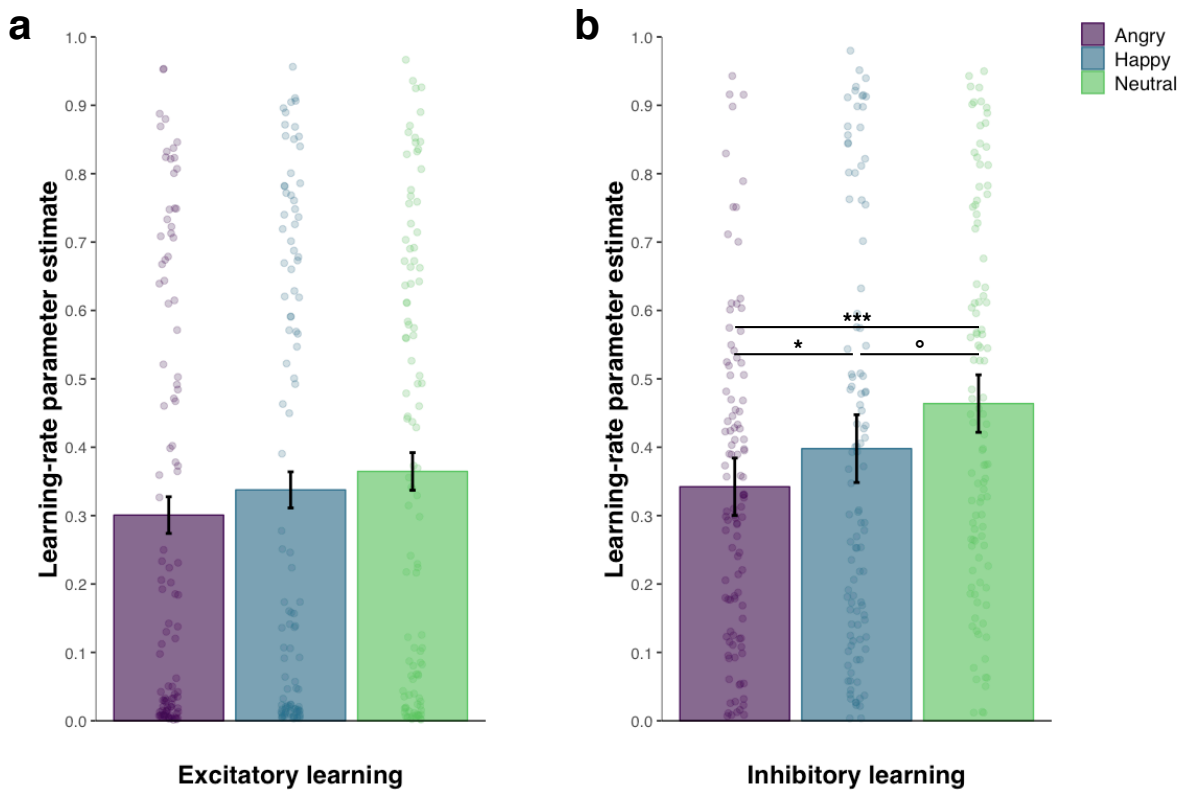

Figure S3.2.4. Learning-rate parameter estimates of the Rescorla-Wagner model implementing dual learning rates using the best-fitting parameters for positive prediction errors (excitatory learning) and negative prediction errors (inhibitory learning) as a function of the conditioned stimulus category (angry vs. happy vs. neutral). The dots indicate data for individual participants. Error bars indicate ± 1 *SEM* adjusted for within-participant designs (Morey, 2008). Asterisks indicate statistically significant differences between conditions (***$p < .001$, *$p <. 05$, °$p < .10$, one-tailed, Holm-Bonferroni corrected).

= 1.83, $p = .070$, whereas the differential reaction time index for happy faces was a marginally significant predictor of the learning-rate estimates for negative prediction errors, b = -0.001, SE = 0.001, β = -.172, $t(101)$ = -1.73, p = .087. No other relationship was observed between the predictors and the estimated learning rates for positive and negative prediction errors (all ps > .16).

Table S3.2.2

Results for the multiple linear regression analyses on the estimated learning rates for positive and negative prediction errors to happy faces (N = 105)

	Estimated learning rate for positive prediction errors to happy faces					Estimated learning rate for negative prediction errors to happy faces				
	b	*SE*	*β*	*t* (101)	*p*	*b*	*SE*	*β*	*t* (101)	*p*
Intercept	0.023	0.168		0.14	.891	0.505	0.148		3.42	<.001
Extraversion	0.011	0.006	.179	1.83	.070	-0.004	0.005	-.079	-0.81	.422
Differential d' index	0.086	0.062	.138	1.39	.167	0.075	0.055	.137	1.37	.173
Differential reaction time index	-0.000	0.001	-.051	-0.51	.609	-0.001	0.001	-.172	-1.73	.087
R^2			.053					.046		

Exploratory analyses

We carried out exploratory analyses to investigate whether the conditioned response to happy faces during early acquisition and extinction, as well as the estimated learning rates for positive and negative prediction errors to happy faces, were predicted by personality traits besides extraversion. To do so, we performed hierarchical multiple linear regressions to examine whether the addition of neuroticism (M = 24.07, SD = 9.10, range = 3-46, Cronbach's α = .89), openness (M = 29.71, SD = 6.08, range = 16-42, Cronbach's α = .70), agreeableness (M = 33.33, SD = 6.57, range = 13-46, Cronbach's α = .78), and conscientiousness (M = 31.59, SD = 7.64, range = 0-46, Cronbach's α = .83) scores improved prediction of the various dependent variables relative to a model including only extraversion, differential d' index for happy faces, and differential reaction time index for happy faces as predictors. Results of these analyses are displayed in Table S3. Altogether, the inclusion of participants' neuroticism, openness, agreeableness, and conscientiousness scores did not statistically significantly improve prediction of the conditioned response, as well as the estimated learning rates for negative prediction errors, to happy faces (all Fs < 2.25, all ps > .07). Neuroticism was negatively associated with the conditioned response to happy faces during early acquisition, b

= -0.005, *SE* = 0.002, β = -.218, *t*(103) = -2.15, *p* = .034; however, this association did not survive correction of the significance level for multiple comparisons using false discovery rate (α = 4/40*.05 = .005; see Benjamini & Hochberg, 1995). By contrast, the inclusion of these predictors improved prediction of the estimated learning rates for positive prediction errors, *F*(4, 97) = 2.69, *p* = .036, and explained an additional 9.44% of the variation thereof. Neuroticism negatively predicted the estimated learning rates for positive prediction errors to happy faces (see Figure S3.2.5), *b* = -0.011, *SE* = 0.004, β = -.305, *t*(101) = -2.98, *p* = .0037, this association remaining statistically significant after correcting for multiple testing using false discovery rate (α = 3/40*.05 = . 0038; see Benjamini & Hochberg, 1995). This exploratory result suggests that happy faces were associated with lower excitatory learning rates in participants high in neuroticism than in those lower in this trait. None of the other personality traits were found to be statistically significant predictors of the conditioned response to happy faces during early acquisition or extinction and of the learning rates for positive and negative prediction errors to happy faces, even without correcting for multiple comparisons (all *p*s > .07).

We conducted additional exploratory analyses to investigate whether personality traits and differential d' and reaction time indices derived from the GNAT were related to the conditioned response and the learning rates to angry and neutral faces. We ran multiple linear regressions with extraversion, neuroticism, openness, agreeableness, conscientiousness, differential d' index, and differential reaction time index as predictors of the conditioned response during (a) early acquisition and (b) extinction, and of the learning-rate estimates for (c) positive and (d) negative prediction errors to both angry and neutral faces. Results of the analyses for angry faces and neutral faces are shown in Table S3 and Table S4, respectively. Conscientiousness was positively associated with the conditioned response to neutral faces during extinction, *b* = 0.003, *SE* = 0.002, β = .196, *t*(103) = 2.02, *p* = .047; however, this relationship was no longer statistically significant after adjusting the significance level for multiple comparisons using false discovery rate (α = 1/28*.05 = .0018; see Benjamini & Hochberg, 1995). No other statistically significant relationship was observed between the various predictors and the dependent variables, even without correcting for multiple testing (all *p*s > .05).

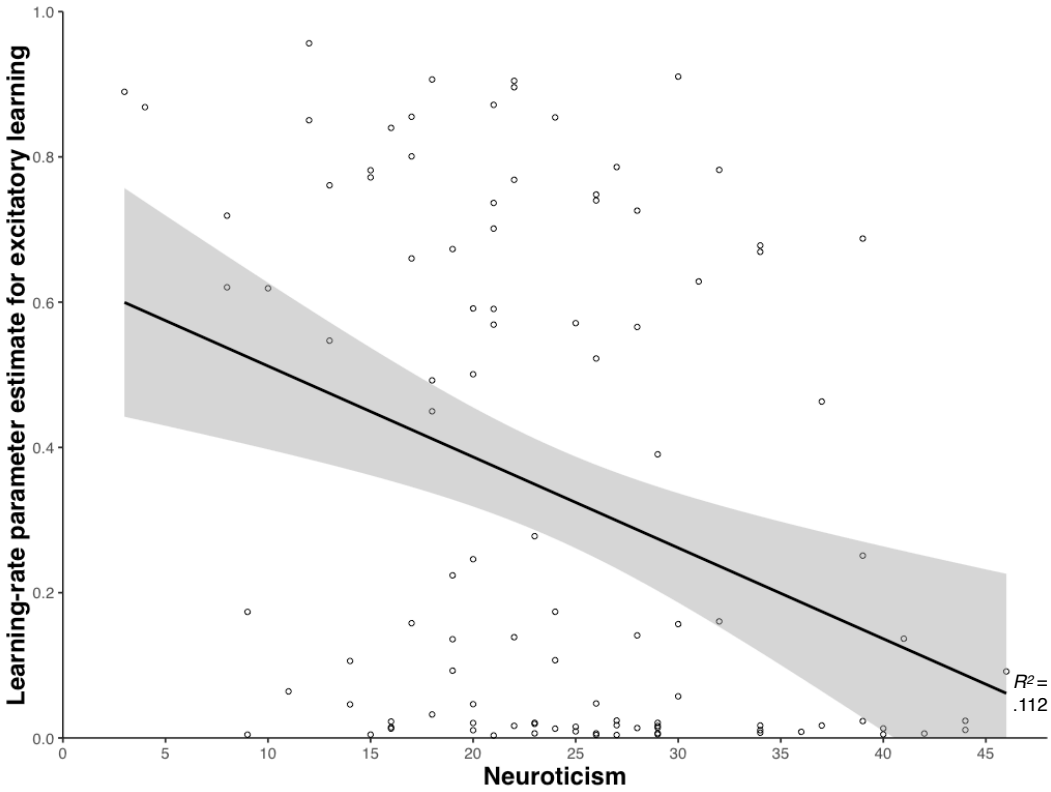

Figure S3.2.5. Relationship between neuroticism and the learning rates for positive prediction errors (excitatory learning) for happy faces. The line represents the fitted regression line using least squares estimation and 95% confidence interval.

Table S3.2.3

Results for the exploratory hierarchical multiple linear regression analyses for happy faces

	Conditioned response to happy faces during early acquisition (N = 107)					Conditioned response to happy faces during extinction (N = 107)					Estimated learning rate for positive prediction error to happy faces (N = 105)					Estimated learning rate for negative prediction error to happy faces (N = 105)				
	b	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (101)	*p*	*b*	*SE*	*β*	*t* (101)	*p*
Model 1	R^2 = .015					R^2 = .131					R^2 = .053					R^2 = .046				
Intercept	0.073	0.106		0.69	.494	0.027	0.087		0.31	.759	0.023	0.168		0.14	.891	0.505	0.148		3.42	<.001
Extraversion	0.003	0.004	.085	0.87	.388	0.002	0.003	.046	0.50	.621	0.011	0.006	.179	1.83	.070	-0.004	0.005	-.079	-0.81	.442
Differential d' index	-0.005	0.039	-.013	-0.13	.896	-0.028	0.032	-.081	-0.87	.386	0.086	0.062	.138	1.39	.167	0.075	0.055	.137	1.37	.173
Differential reaction time index	-0.001	0.001	-.096	-0.96	.341	0.002	0.0005	.360***	3.83	< .001	-0.000	0.001	-.051	-0.51	.609	-0.001	0.001	-.172	-1.73	.087
Model 2	R^2 = .097, ΔR^2 = .082, $F(4, 99) = 2.24$, p = .070					R^2 = .149, ΔR^2 = .019, $F(4, 99) = 0.55$, p = .700					R^2 = .147, ΔR^2 = .094, $F(4, 97) = 2.66$, p = .036					R^2 = .074, ΔR^2 = .028, $F(4, 97) = 0.73$, p = .573				
Intercept	0.080	0.222		0.36	.720	0.031	0.188		0.17	.868	0.179	0.346		0.52	.605	0.367	0.317		1.16	.249
Extraversion	0.002	0.004	.055	0.51	.612	-0.000	0.003	-.008	-0.08	.936	0.005	0.006	.078	0.74	.463	-0.006	0.006	-.109	-0.99	.327
Differential d' index	-0.022	0.039	-.059	-0.59	.558	-0.033	0.033	-.097	-1.00	.317	0.066	0.061	.105	1.07	.286	0.063	0.056	.116	1.13	.262

Differential reaction time index	-0.000	0.001	-.082	-0.83	.409	0.002	0.0005	.359**	3.74	< .001	-0.000	0.001	-.044	-0.45	.654	-0.001	0.001	-.180	-1.77	.080
Neuroticism	-0.005	0.002	-.224	-2.15	.034	-0.003	0.002	-.127	-1.26	.211	-0.011	0.004	-.304*	-2.98	.004	0.001	0.004	.040	0.38	.707
Openness	0.006	0.003	.165	1.70	.093	0.002	0.003	.055	0.59	.559	0.008	0.005	.136	1.42	.158	-0.004	0.005	-.085	-0.85	.397
Agreeableness	-0.003	0.003	-.106	-1.02	.312	0.001	0.003	.042	0.41	.681	0.002	0.005	.035	0.34	.736	0.003	0.005	.073	0.68	.498
Conscien-tiousness	0.003	0.003	.110	1.10	.273	0.001	0.002	.025	0.26	.797	0.000	0.004	.004	0.04	.969	0.005	0.004	.134	1.32	.192

Note. ***$p < .001$, **$p < .01$ (FDR-corrected).

Table S3.2.4

Results for the exploratory multiple linear regression analyses for angry faces

	Conditioned response to angry faces during early acquisition (N = 107)					Conditioned response to angry faces during extinction (N = 107)					Estimated learning rate for positive prediction error to angry faces (N = 105)					Estimated learning rate for negative prediction error to angry faces (N = 105)				
	b	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (101)	*p*	*b*	*SE*	*β*	*t* (101)	*p*
Intercept	-0.170	0.245		-0.69	.490	0.118	0.181		0.65	.518	-0.401	0.352		-1.14	.259	0.569	0.262		2.17	.033
Extraversion	-0.000	0.004	-.008	-0.07	.943	-0.004	0.003	-.134	-1.22	.225	0.009	0.006	.167	1.54	.127	-0.006	0.005	-.142	-1.27	.208
Differential d' index	-0.017	0.051	-.036	-0.34	.733	0.000	0.037	.000	0.00	.999	-0.062	0.073	-.089	-0.85	.396	0.003	0.054	.007	0.06	.950
Differential reaction time index	0.000	0.001	.002	0.02	.983	-0.001	0.001	-.132	-1.24	.219	0.000	0.001	.032	0.31	.761	-0.001	0.001	-.095	-0.87	.387
Neuroticism	-0.003	0.003	-.109	-1.01	.317	-0.001	0.002	-.081	-0.75	.454	-0.001	0.004	-.036	-0.33	.739	-0.001	0.003	-.056	-0.51	.609
Openness	0.003	0.004	.081	0.80	.426	-0.002	0.003	-.081	-0.79	.430	0.001	0.005	.013	0.13	.896	0.002	0.004	.042	0.40	.688
Agreeableness	0.005	0.004	.150	1.40	.166	0.002	0.003	.093	0.87	.389	0.006	0.005	.122	1.15	.254	-0.003	0.004	-.086	-0.78	.436
Conscien-tiousness	0.004	0.003	.142	1.36	.178	0.003	0.002	.146	1.40	.164	0.008	0.004	.180	1.73	.088	0.001	0.003	.022	0.21	.838
R^2			.061					.066					.098					.045		

Table S3.2.5

Results for the exploratory multiple linear regression analyses for neutral faces

	Conditioned response to neutral faces during early acquisition ($N = 107$)					Conditioned response to neutral faces during extinction ($N = 107$)					Estimated learning rate for positive prediction error to neutral faces ($N = 105$)					Estimated learning rate for negative prediction error to neutral faces ($N = 105$)				
	b	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (103)	*p*	*b*	*SE*	*β*	*t* (101)	*p*	*b*	*SE*	*β*	*t* (101)	*p*
Intercept	-0.258	0.203		-1.27	.206	-0.089	0.123		-0.72	.471	0.257	0.344		0.75	.457	0.720	0.288		2.50	.014
Extraversion	0.005	0.004	.144	1.31	.193	-0.000	0.002	.001	0.01	.993	0.005	0.006	.095	0.84	.405	-0.003	0.005	-.055	-0.48	.631
Differential d' index	-0.025	0.044	-.057	-0.58	.565	0.025	0.026	.093	0.96	.342	-0.046	0.074	-.063	-0.62	.539	0.009	0.062	.014	0.14	.891
Differential reaction time index	0.000	0.001	.010	0.10	.924	0.000	0.0003	.027	0.27	.788	0.001	0.001	.125	1.20	.232	-0.000	0.001	-.024	-0.23	.816
Neuroticism	0.000	0.002	.008	0.07	.942	-0.001	0.001	-.040	-0.39	.699	0.000	0.004	.001	0.01	.992	-0.004	0.003	-.137	-1.25	.213
Openness	0.000	0.003	.007	0.07	.944	-0.003	0.002	-.134	-1.38	.171	0.003	0.005	.057	0.55	.583	0.002	0.005	.050	0.49	.629
Agreeableness	0.005	0.003	.181	1.72	.089	0.003	0.002	.184	1.78	.078	-0.003	0.005	-.059	-0.54	.588	-0.004	0.004	-.092	-0.84	.402
Conscien-tiousness	-0.000	0.003	-.003	-0.03	.978	0.003	0.002	.196	2.02	.047	-0.001	0.004	-.023	-0.22	.827	-0.001	0.004	-.025	-0.24	.811
R^2			.075					.106					.026					.025		

3.3. STUDY 3:

ACHIEVEMENT MOTIVATION MODULATES PAVLOVIAN AVERSIVE CONDITIONING TO GOAL-RELEVANT STIMULI[12]

Abstract

Pavlovian aversive conditioning is a fundamental form of learning helping organisms survive in their environment. Previous research has suggested that organisms are prepared to preferentially learn to fear stimuli that have posed threats to survival across evolution. Here, we examined whether enhanced Pavlovian aversive conditioning can occur to stimuli that are relevant to the organism's concerns beyond biological and evolutionary considerations, and whether such preferential learning is modulated by inter-individual differences in affect and motivation. Seventy-two human participants performed a spatial cueing task where the goal-relevance of initially neutral stimuli was experimentally manipulated. They subsequently underwent a differential Pavlovian aversive conditioning paradigm, in which the goal-relevant and goal-irrelevant stimuli served as conditioned stimuli. Skin conductance response was recorded as an index of the conditioned response and participants' achievement motivation was measured to examine its impact thereon. Results show that achievement motivation modulated Pavlovian aversive learning to goal-relevant versus goal-irrelevant stimuli. Participants with high achievement motivation more readily acquired a conditioned response to goal-relevant compared with goal-irrelevant stimuli than did participants with lower achievement motivation. However, no difference was found between goal-relevant and goal-irrelevant stimuli during extinction. These findings suggest that stimuli that are detected as relevant to the organism can induce facilitated Pavlovian aversive conditioning even though they hold no inherent threat value and no biological evolutionary significance, and that the occurrence of such learning bias is critically dependent on inter-individual differences in the organism's concerns, such as achievement motivation.

[12] Reprint of: Stussi, Y., Ferrero, A., Pourtois, G., & Sander, D. (in press). Achievement motivation modulates Pavlovian aversive conditioning to goal-relevant stimuli. *npj Science of Learning*.

3.3.1. Introduction

Pavlovian aversive conditioning is a fundamental form of learning in the animal kingdom, being ubiquitous across a wide variety of species ranging from simple (e.g., fruit fly) to more complex (e.g., human) organisms (LeDoux, 1994). It consists of both the learning process and procedure whereby an environmental stimulus (the conditioned stimulus) acquires the ability to elicit a preparatory response (the conditioned response) by virtue of a single or repeated contingent pairing with a biologically significant aversive event (the unconditioned stimulus; Pavlov, 1927; Rescorla, 1988b). However, not all stimuli are equally associable in Pavlovian aversive conditioning (Garcia & Koelling, 1966). Previous research has shown that specific classes of evolutionarily threat-relevant stimuli, such as snakes, angry faces, or outgroup faces, are more rapidly (Ho & Lipp, 2014; Öhman, Eriksson, & Olofsson, 1975) and persistently (Öhman & Dimberg, 1978; Öhman, Eriksson, et al., 1975; Öhman, Fredrikson, Hugdahl, & Rimmö, 1976; Olsson, Ebert, Banaji, & Phelps, 2005) associated with an aversive outcome than nonthreatening stimuli, such as flowers, happy faces, or ingroup faces (for reviews, see Mallan, Lipp, & Cochrane, 2013; Öhman & Mineka, 2001). These preferential associations have generally been interpreted as evidence for the preparedness (Seligman, 1971) and fear module (Öhman & Mineka, 2001) theories, which posit that organisms are biologically prepared to associate stimuli that have posed threats to the species' survival across evolution with aversive events.

At variance with these evolutionary theories, an alternative framework deriving from appraisal theories of emotion (Sander, Grafman, & Zalla, 2003; Sander, Grandjean, & Scherer, 2005) asserts that preferential emotional learning is not driven by a threat-specific mechanism, but by a more general mechanism of relevance detection (Stussi, Brosch, & Sander, 2015; Stussi, Pourtois, & Sander, 2018). Relevance detection is a rapid and adaptive process that determines whether a stimulus encountered in the environment is relevant to the organism's concerns, such as their goals, needs, motives, or values (Frijda, 1986; Pool, Brosch, Delplanque, & Sander, 2016; Sander et al., 2003, 2005; Stussi, Pourtois, et al., 2018). Importantly, this proposal allows for incorporating the findings of preferential Pavlovian aversive conditioning to threat-relevant stimuli, as these stimuli are highly relevant for the organism's survival, but also generates new testable predictions: Stimuli that are detected as relevant to the organism's concerns benefit from preferential emotional learning, regardless of their valence and evolutionary status per se.

In agreement with this hypothesis, we have recently shown that, similar to threat-relevant stimuli (angry faces and snakes), positive stimuli with biological relevance (baby faces and erotic stimuli) are likewise persistently associated with an aversive outcome (electric stimulation) during Pavlovian aversive conditioning, thereby demonstrating that preferential Pavlovian aversive conditioning is not restricted to negative threat-related stimuli but extends to positive relevant stimuli (Stussi, Pourtois, et al., 2018). Nonetheless, this study did not address the question of whether the stimulus' evolutionary history might be a key ingredient in this preferential emotional learning. Existing evidence on this issue is mixed: Whereas some studies found a similar enhanced resistance to extinction of learned threat to both biological (snakes) and cultural (pointed guns) threats (Flykt, Esteves, & Öhman, 2007; Hugdahl & Johnsen, 1989), other studies observed a greater persistence of learned threat to threat-relevant stimuli from phylogenetic origin than from ontogenetic origin (E.W. Cook, Hodes, & Lang, 1986; Hugdahl & Kärker, 1981). Accordingly, whether enhanced emotional learning is confined to evolutionarily relevant stimuli or encompasses stimuli with high relevance to the organism beyond biological and evolutionary considerations remains to be better elucidated.

A key assumption of the relevance detection model is that emotional learning is largely affected by individual differences in affect and motivation. The process of relevance detection is inextricably tied to the organism's concerns, the salience and priority of which may flexibly and rapidly change based on current environmental contingencies, and which are likely to vary across individuals (Cunningham & Brosch, 2012; Frijda, 1986; Sander et al., 2005). As a result, the same stimulus may potentially produce a learning bias for a given individual, but not for another one, if these two individuals differ according to their current concerns, and hence the way in which they appraise the stimulus at stake. In line with this view, inter-individual differences are inherent and highly prevalent in Pavlovian conditioning (Lonsdorf & Merz, 2017; Pavlov, 1927), as reflected by a substantial variability across individuals in this learning process as a function of biological, experiential, or personality factors, as well as affective or cognitive biases (Byrom & Murphy, 2018; Gazendam et al., 2015; Hartley, Fischl, & Phelps, 2011; Lonsdorf & Merz, 2017; Lonsdorf et al., 2009; Sjouwerman, Scharfenort, & Lonsdorf, 2018; Zorawski, Cook, Kuhn, & LaBar, 2005). Despite these initial attempts to consider inter-individual differences for yielding a better understanding of emotional learning in humans, their contribution to Pavlovian conditioning, along with the underlying mechanisms thereof, remain yet poorly understood (Lonsdorf & Merz, 2017).

Here, we therefore aimed to investigate whether enhanced emotional learning could occur to stimuli that are relevant to the organism's concerns independently of their intrinsic evolutionary significance, as well as the modulatory role of inter-individual differences therein. To this end, we used initially neutral stimuli (i.e., geometric figures) and experimentally manipulated their relevance for task goals in a spatial cueing task (Pool, Brosch, Delplanque, & Sander, 2014; Figure 3.3.1), some stimuli being goal-relevant by predicting target location (*goal-relevant valid stimuli*) or predicting the opposite location relative to the target (*goal-relevant invalid stimuli*), and others goal-irrelevant by being nonpredictive of target location (*goal-irrelevant stimuli*). We subsequently used these stimuli as conditioned stimuli in a differential Pavlovian aversive conditioning paradigm. In this paradigm, one stimulus (CS+) from each of the three stimulus categories was systematically paired with a mild electric stimulation (US) during the acquisition phase, whereas the other stimulus (CS-) from each stimulus category was never associated with it. In the following extinction phase, the US was no longer delivered. Skin conductance response (SCR) was measured continuously throughout the entire conditioning procedure. The conditioned response (CR) was operationalised as the differential SCR to the CS+ minus CS- from the same stimulus category (Olsson et al., 2005; Stussi, Pourtois, et al., 2018) and used as an index of learning.

To assess the role of inter-individual differences, we examined the influence of participants' achievement motivation on Pavlovian aversive learning to goal-relevant versus goal-irrelevant stimuli. Achievement motivation refers to the need or concern to develop or demonstrate high ability, and to attain a standard of excellence or a success goal (Murray, 1938; Nicholls, 1984). Inter-individual differences in achievement motivation have been reported to affect how individuals appraise the relevance of objects and situations. For instance, when confronted with achievement-related situations, individuals high in achievement motivation have been shown to appraise these situations as more important than did individuals lower in this trait (Smith & Pope, 1992). Moreover, individuals with a high level of achievement motivation have been reported to be intrinsically motivated to perform a task for its own sake (French, 1955; McKeachie, 1961). In light of this evidence and given that the spatial cueing task involves an achievement component related to task performance and success, we inferred that individuals with high achievement motivation would be highly motivated to perform well in this task, thereby attaching higher relevance to the goal-relevant stimuli and lower relevance

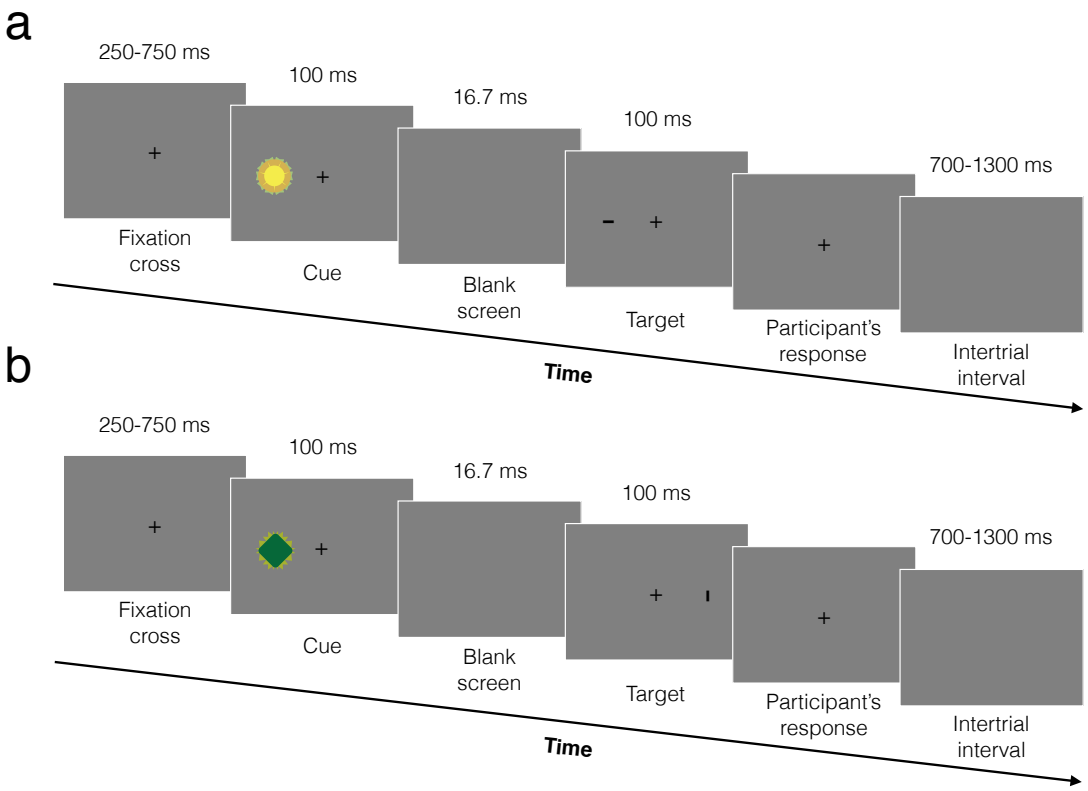

Figure 3.3.1. Illustration of the spatial cueing task used in the experiment. (a) In valid trials, the target appeared at the same location as the cue. (b) In invalid trials, the target appeared at the opposite location as the cue. The cues were geometric figures, which systematically predicted target location at the same (*goal-relevant valid*) or the opposite (*goal-relevant invalid*) location, or were nonpredictive of target location (*goal-irrelevant*). Participants were requested to detect the target orientation (horizontal vs. vertical).

to the goal-irrelevant stimuli than individuals with lower achievement motivation because of their respective informativeness and instrumentality, or lack thereof, for task accomplishment.

As preferential emotional learning is generally characterised by a faster acquisition of the conditioned response and/or an enhanced resistance to extinction of that conditioned response (Öhman & Mineka, 2001; Seligman, 1971), these two indicators being considered as equally valid (Rescorla, 1980), we hypothesised that the conditioned response to goal-relevant stimuli would be (a) acquired faster and (b) more resistant to extinction than the conditioned response to goal-irrelevant stimuli. Furthermore, we predicted that inter-individual differences in achievement motivation would modulate the acquisition readiness and the resistance to extinction of the conditioned response to goal-relevant stimuli compared with goal-irrelevant

stimuli, with higher achievement motivation leading to a greater difference in the conditioned response to goal-relevant versus goal-irrelevant stimuli during early acquisition and during extinction.

3.3.2. Results

Spatial cueing task

The reaction times in the spatial cueing task were analysed using a repeated-measures general linear model (GLM) assuming compound symmetry covariance structure with stimulus type (to-be-CS+ vs. to be-CS-) and stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) as within-participant categorical factors, and participants' standardised (z-score) achievement motivation score as a continuous predictor. This analysis revealed a marginally significant main effect of stimulus category, $F(2, 140) = 2.85$, $p = .061$, partial $\eta^2 = .039$, 90% confidence interval (CI) [.000, .095]. No other effect was observed (all Fs < 2.39, all ps > .12, all partial η^2s < .033). A polynomial contrast analysis showed a statistically significant linear trend in the reaction times as a function of stimulus category, $F(1, 70) = 5.41$, $p = .023$, partial $\eta^2 = .072$, 90% CI [.005, .181], indicating increased reaction times in detecting the target from goal-relevant valid cues (M = 496.52 ms, SD = 129.53) to goal-irrelevant cues (M = 500.92 ms, SD = 147.23) to goal-relevant invalid cues (M = 505.21 ms, SD = 140.12; Figure 3.3.2). This result reflects the occurrence of a cueing validity effect, hence suggesting that the spatial cueing task triggered attention orienting, although it is important to note that this effect was small.

Descriptive analyses revealed the presence of an outlier in the reaction time data exhibiting slow reaction times in all the conditions. Analysis excluding this outlier revealed a statistically significant main effect of stimulus category, $F(2, 138) = 3.19$, $p = .044$, partial $\eta^2 = .044$, 90% CI [.0006, .103]. No other effect was statistically significant (all Fs < 2.28, all ps > .10, all partial η^2s < .032). The linear trend remained statistically significant after the outlier exclusion, $F(1, 69) = 4.33$, $p = .041$, partial $\eta^2 = .059$, 90% CI [.001, .165].

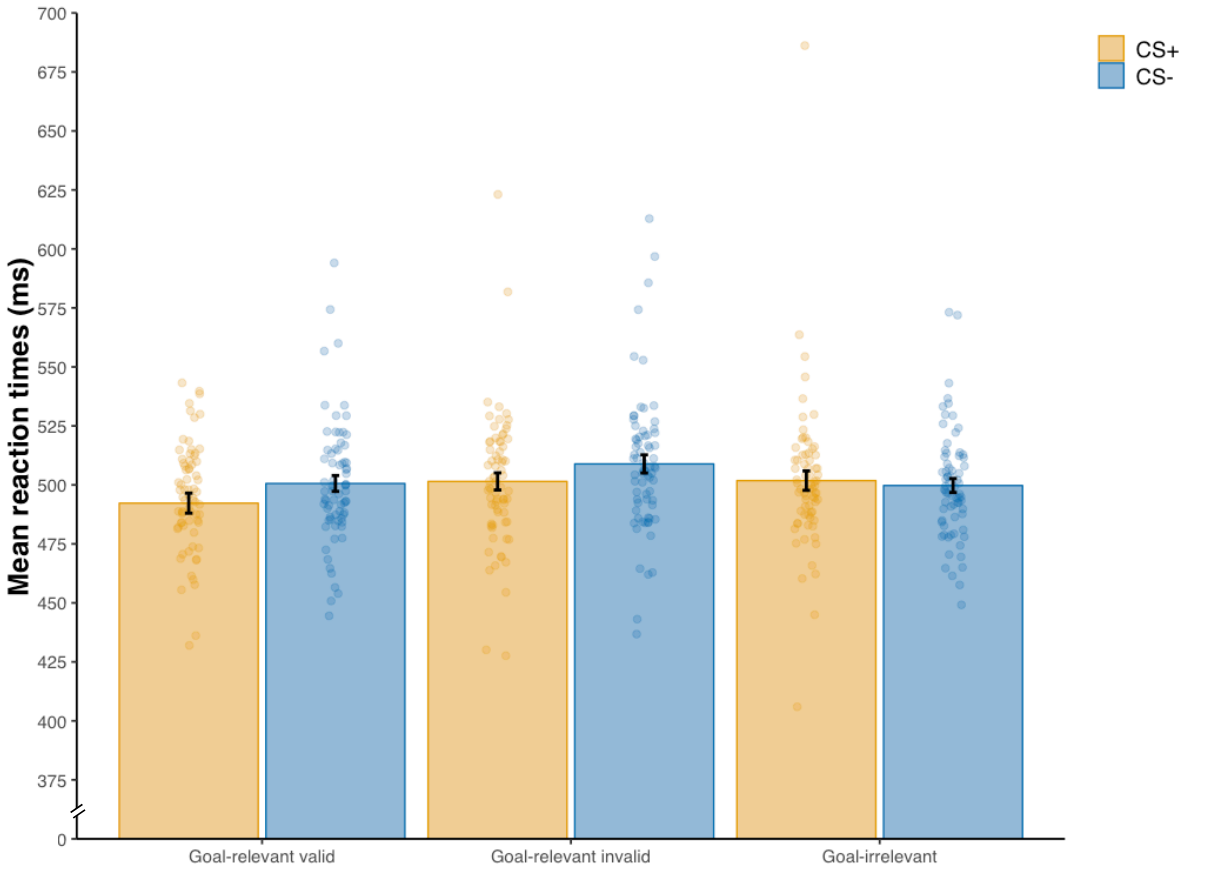

Figure 3.3.2. Mean reaction times during the spatial cueing task as a function of stimulus type (to-be-CS+ vs. to-be-CS) and stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant). The dots indicate normalised data for individual participants. Error bars indicate ± 1 standard error of the mean adjusted for within-participant designs (Morey, 2008).

Differential Pavlovian aversive conditioning

According to standard practice in the human conditioning literature (Lonsdorf et al., 2017; Olsson et al., 2005), the SCR data (Figure 3.3.3) was analysed separately for each conditioning phases. The habituation and extinction phases were each analysed with a GLM with stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) as a within-participant categorical factor and participants' standardised achievement motivation score as a continuous variable. To examine the differential CR acquisition readiness as a function of the stimulus' goal-relevance and the modulatory influence of participants' achievement motivation thereon, the acquisition phase was split into an early and a late phase, and was analysed using a repeated-measures GLM assuming compound symmetry covariance structure with time (early vs. late) and stimulus category (goal-relevant valid vs. goal-relevant

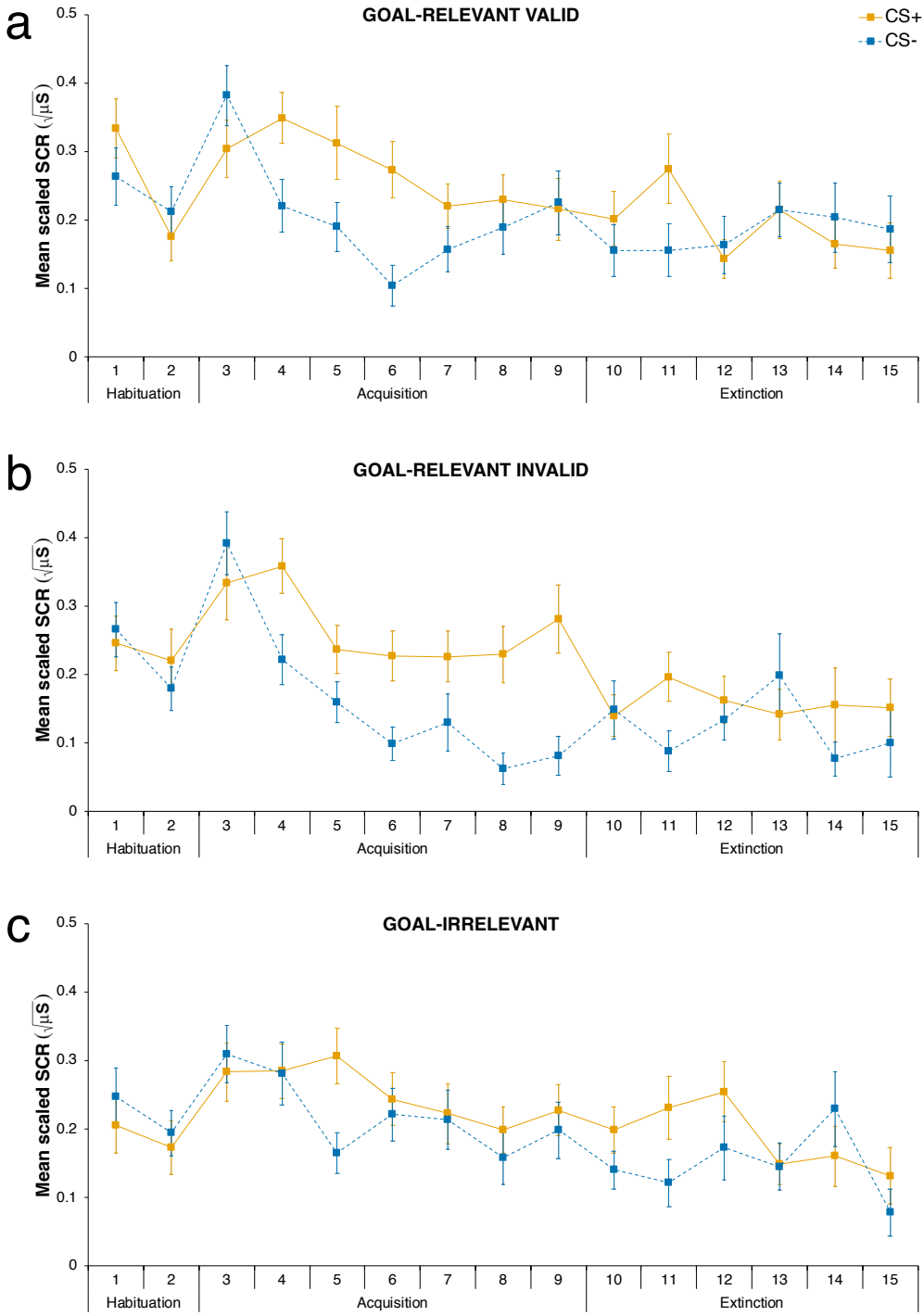

Figure 3.3.3. Mean scaled skin conductance response (SCR) to the conditioned stimuli as a function of conditioned stimulus type (CS+ vs. CS-) across trials. Mean scaled SCR to (a) goal-relevant valid stimuli, (b) goal-relevant invalid stimuli, and (c) goal-irrelevant stimuli. Error bars indicate ± 1 standard error of the mean adjusted for within-participant designs.

invalid vs. goal-irrelevant) as within-participant categorical factors, and participants' standardised achievement motivation score as a continuous predictor. During habituation, there

were no pre-existing differences in differential SCRs to the various stimulus categories, or as a function of participants' achievement motivation or the interaction between these factors (all Fs < 0.48, all ps > .62, all partial η^2s < .007).

To assess whether a CR was successfully acquired (i.e., greater SCRs to the CS+ than to the CS-) in response to the different stimulus categories (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) during acquisition as expected, we performed one-tailed one-sample t tests on the CR across the entire acquisition phase. These tests showed that the SCRs to the CS+ were larger than to the CS- for goal-relevant valid stimuli, $t(71) = 3.32$, $p < .001$ (one-tailed), $g_{av} = 0.547$, 95% CI [0.212, 0.891] and goal-relevant invalid stimuli, $t(71) = 6.09$, $p < .001$ (one-tailed), $g_{av} = 1.005$, 95% CI [0.646, 1.380], thereby indicating successful differential conditioning, whereas they were marginally larger than to the CS- for goal-irrelevant stimuli, $t(71) = 1.49$, $p = .070$ (one-tailed), $g_{av} = 0.246$, 95% CI [-0.082, 0.577]. The apparent less robust differential conditioning to goal-irrelevant stimuli was mainly driven by the existence of an outlier (-5.66 *SD* from the mean CR to goal-irrelevant stimuli), who was strongly conditioned to the goal-irrelevant CS-. The one-sample t test excluding this outlier indeed reflected a stronger differential conditioning to goal-irrelevant stimuli, $t(70) = 2.90$, $p = .002$ (one-tailed), $g_{av} = 0.482$, 95% CI [0.147, 0.824].

Moreover, the GLM revealed a statistically significant main effect of stimulus category, $F(2, 140) = 3.81$, $p = .024$, partial $\eta^2 = .052$, 90% CI [.004, .113]. A planned contrast analysis showed that goal-relevant valid (contrast weight: +1) and goal-relevant invalid (contrast weight: +1) stimuli ($M = 0.11$, $SD = 0.15$) led to the acquisition of a larger CR than goal-irrelevant stimuli (contrast weight: -2; $M = 0.04$, $SD = 0.23$), $F(1, 71) = 5.49$, $p = .022$, partial $\eta^2 = .072$, 90% CI [.006, .181]. Albeit not statistically significant, we also observed a marginal trend for the interaction between time and stimulus category, $F(2, 140) = 2.61$, $p = .077$, partial $\eta^2 = .036$, 90% CI [.000, .090], and for the three-way interaction between time, stimulus category, and achievement motivation, $F(2, 140) = 2.65$, $p = .074$, partial $\eta^2 = .036$, 90% CI [.000, .091]. No other effect reached statistical significance (all Fs < 1.14, all ps > .29, all partial η^2s < .016). To specifically test our a priori hypothesis concerning the CR acquisition readiness to goal-relevant versus goal-irrelevant stimuli and its modulation by inter-individual differences in achievement motivation, we constructed a contrast comparing the difference between the CR to goal-relevant valid (contrast weight: +1) and goal-relevant invalid (contrast weight: +1) stimuli versus goal-irrelevant stimuli (contrast weight: -2) during early acquisition, and tested whether this difference was influenced by participants' standardised achievement

motivation score by means of a repeated-measures GLM. Consistent with our prediction, this analysis indicated that the difference between the CR to goal-relevant stimuli and the CR to goal-irrelevant stimuli was modulated by participants' achievement motivation during early acquisition, $F(1, 70) = 5.15$, $p = .026$, partial $\eta^2 = .069$, 90% CI [.004, .177], with high level of achievement motivation resulting in a greater difference in CR acquisition readiness between goal-relevant and goal-irrelevant stimuli (Figure 3.3.4a). Further analyses using simple slopes congruently revealed that participants with high achievement motivation (+ 1 *SD*) more readily acquired a CR to goal-relevant stimuli ($M = 0.15$) than to goal-irrelevant stimuli ($M = -0.02$), $F(1, 70) = 8.11$, $p = .006$, partial $\eta^2 = .104$, 90% CI [.018, .222], whereas no statistically significant difference between goal-relevant stimuli ($M = 0.11$) and goal-irrelevant stimuli ($M = 0.13$) was observed for participants with lower achievement motivation (-1 *SD*), $F(1, 70) = 0.13$, $p = .717$, partial $\eta^2 = .002$, 90% CI [.000, .048] (Figure 3.3.4b). Achievement motivation conversely did not moderate the difference between the CR to goal-relevant versus goal-irrelevant stimuli in late acquisition, $F(1, 70) = 0.62$, $p = .433$, partial $\eta^2 = .009$, 90% CI [.000, .076].

Analysis of the extinction phase showed that the CR did not statistically differ across the three stimulus categories, $F(2, 140) = 0.54$, $p = .586$, partial $\eta^2 = .008$, 90% CI [.000, .037], suggesting a similar CR extinction to goal-relevant valid, goal-relevant invalid, and goal-irrelevant stimuli. The extinction of the CR was likewise not affected by participants' achievement motivation (all Fs < 0.32, all ps > .57, all partial η^2s < .005). The difference between the CR to goal-relevant stimuli and the CR to goal-irrelevant stimuli was not modulated by participants' achievement motivation either, $F(1, 70) = 0.15$, $p = .696$, partial $\eta^2 = .002$, 90% CI [.000, .050]. This result reflects that the CR persistence to goal-relevant stimuli compared with goal-irrelevant stimuli did not statistically differ as a function of participants' achievement motivation during extinction.

3.3.3. Discussion

Altogether, our results show that goal-relevant stimuli induced the acquisition of a larger conditioned response than goal-irrelevant stimuli, thus suggesting stronger Pavlovian aversive conditioning. Most importantly, this effect was notably driven by inter-individual differences in achievement motivation that modulated the acquisition readiness of the conditioned response to goal-relevant stimuli compared with goal-irrelevant stimuli, as

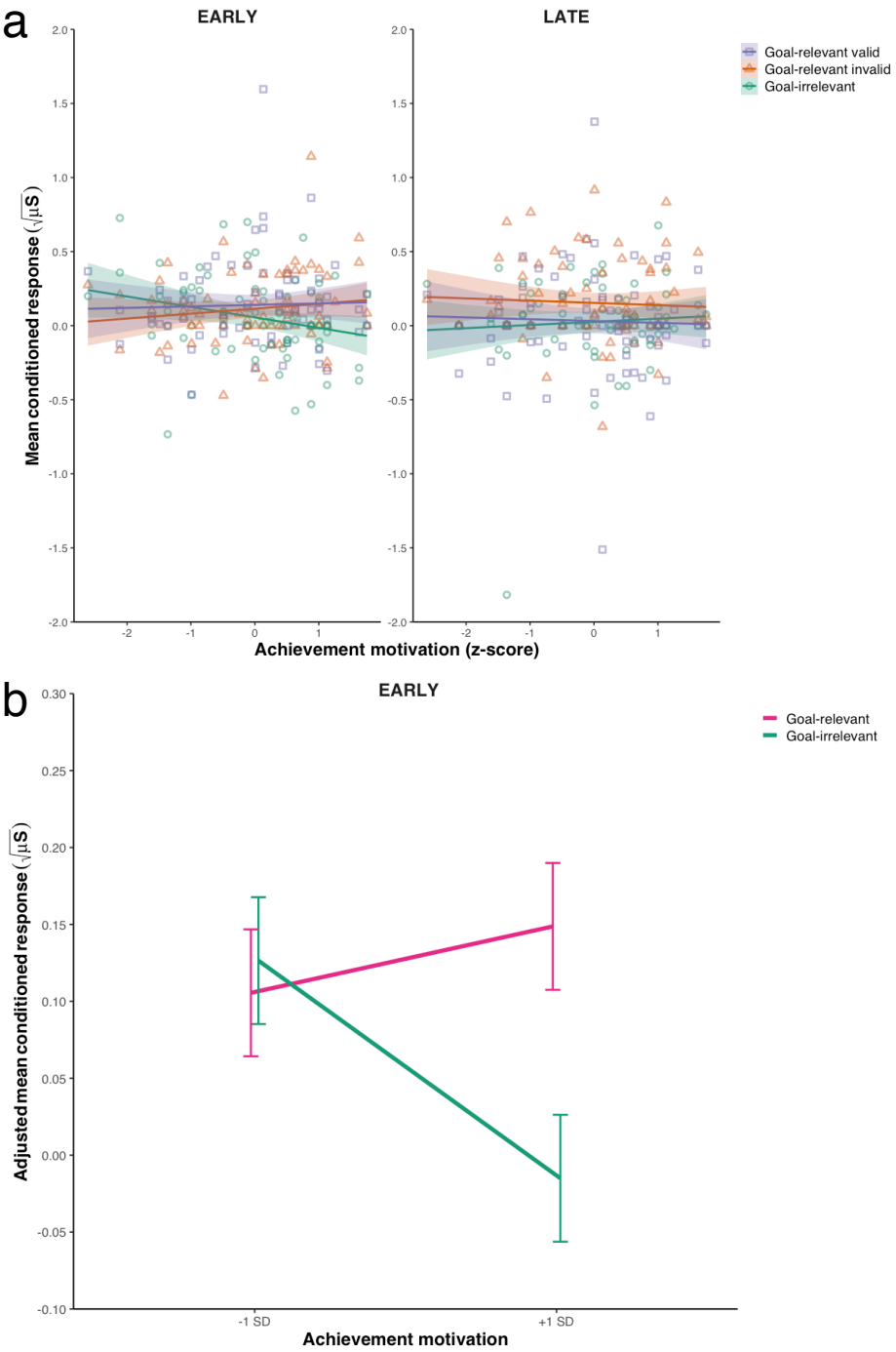

Figure 3.3.4. Influence of achievement motivation on the conditioned response to goal-relevant versus goal-irrelevant stimuli during acquisition (a) Mean conditioned response as a function of stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) and participants' standardised (z-score) achievement motivation score in the early and the late acquisition phase. The points indicate data for individual participants. The curves represent the best-fitting regression lines using least squares estimation and their 95% confidence interval. (b) Mean adjusted conditioned response to goal-relevant versus goal-irrelevant stimuli during early acquisition as a function of low (-1 *SD*) and high (+1 *SD*) achievement motivation. Error bars indicate ± 1 standard error of the mean.

revealed by an interaction of the stimulus' goal-relevance with participants' achievement motivation during early acquisition. Participants with high achievement motivation more readily acquired a conditioned response to goal-relevant stimuli than to goal-irrelevant stimuli, thus reflecting a learning bias, whereas no learning bias was observed in participants with lower achievement motivation. This indicates that inter-individual differences can produce enhanced Pavlovian aversive conditioning to the very same stimuli depending on their relevance to the individual's current concerns, such as their achievement motive. Such findings dovetail nicely with the relevance detection model (Stussi et al., 2015; Stussi, Pourtois, et al., 2018) according to which preferential emotional learning stems from the interaction between the stimulus and the organism's current concerns, thereby assigning a crucial role to inter-individual differences in enhanced emotional learning. On the other hand, we failed to observe an enhanced resistance to extinction to goal-relevant versus goal-irrelevant stimuli, and no modulatory effect of inter-individual differences in achievement motivation was reported thereon, which is at odds with our predictions.

The fact that we found faster Pavlovian aversive conditioning to goal-relevant versus goal-irrelevant stimuli in participants high in achievement motivation but not in those lower in achievement motivation may relate to the interplay between the manipulation of stimuli's relevance for task goals and participants' current concerns. The construct of goal-relevance has been suggested to cover at least three partly dissociable but related components (Severo, Walentowska, Moors, & Pourtois, 2017; Walentowska, Moors, Paul, & Pourtois, 2016; Walentowska, Paul, Severo, Moors, & Pourtois, 2018): (1) task-relevance, which pertains to the degree to which a stimulus signals the opportunity of implementing and satisfying a specific goal in a given task, (2) informativeness, which refers to the degree to which a stimulus provides reliable information about a goal's satisfaction status, and (3) the impact a stimulus has on the individual's goals. It has been further advanced that a stimulus that is task-relevant is likewise goal-relevant in terms of informativeness and impact, whereas it can be goal-relevant in terms of informativeness and/or impact without being task-relevant (Walentowska et al., 2016). Importantly, task-relevance however differs from current concerns in that it refers to task instructions, and begins and ends in task context; in contrast, current concerns involve a state of commitment about their satisfaction that extends across various contexts and situations beyond a particular task (Klinger, 1975; Pool, Brosch, et al., 2016). Accordingly, the stimuli's goal-relevance may have generalised beyond the spatial cueing task for participants with high achievement motivation because of the stimuli's informativeness and/or impact to

their achievement-related concerns, whereas it ended with the spatial cueing task for participants with lower achievement motivation, the goal-relevant stimuli no longer being task-relevant and of higher relevance to their current concerns than the goal-irrelevant stimuli.

Critically, the facilitated Pavlovian aversive conditioning to goal-relevant than to goal-irrelevant stimuli observed in participants with high achievement motivation furthermore suggests that stimuli that are detected as relevant to the organism's concerns can also be readily and preferentially conditioned to threat even though they hold no intrinsic biological evolutionary significance. This finding reflects that preferential emotional learning is not restricted to stimuli that are relevant in a phylogenetic sense. In this respect, the current study concurs with previous research on human conditioning reporting enhanced Pavlovian aversive learning to ontogenetic threat-relevant stimuli (Flykt et al., 2007; Hugdahl & Johnsen, 1989). It even adds to these earlier reports by showing that initially neutral stimuli devoid of any pre-existing threat value that have acquired goal-relevance can likewise be readily associated with a naturally aversive event in individuals high in achievement motivation. In that sense, our results thereby suggest that preferential emotional learning may be underlain by a relevance detection mechanism, as opposed to a fear- or threat-specific mechanism, allowing the organism to adaptively and flexibly produce a learning bias towards specific stimuli depending on their relevance to the organism's current concerns (Stussi et al., 2015; Stussi, Pourtois, et al., 2018).

Nonetheless, the fact that we did not find effects of stimulus' goal-relevance during extinction suggests that the preferential aversive learning to goal-relevant stimuli as a function of inter-individual differences in achievement motivation was rather modest and transient. This negative finding notably departs from the greater resistance to extinction to threat-relevant stimuli than to threat-irrelevant stimuli typically reported in the human conditioning literature (see, e.g., Mallan et al., 2013; Öhman & Mineka, 2001). Although the present experiment indicates that goal-relevant stimuli can produce facilitated Pavlovian aversive conditioning relative to goal-irrelevant stimuli even if they have no inherent threat value when considering achievement motivation, it appears that the effects of goal-relevance observed therein are likely to be smaller than those usually obtained with threat-relevant stimuli. It is worth noting, however, that such potential difference is fully consistent with our general framework supporting the relevance detection hypothesis: Whereas threat-relevant stimuli are highly relevant for the organism's survival, the goal-relevant stimuli used here were only temporarily relevant for task-related goals in labortatory settings. In other words, because survival is

arguably one of the highest prioritised concerns, survival-relevance can be conceptually considered as a high-value sub-category of goal-relevance. Accordingly, the type of goal-relevant stimuli that we used in the current study probably held a lower level of relevance to the organism than threat-relevant stimuli, thereby possibly accounting for the occurrence of seemingly weaker effects. In this context, an interesting avenue for future research would thus be to directly compare the impact of survival-relevance (e.g., using threat-relevant stimuli) to the impact of other types of goal-relevance on Pavlovian aversive conditioning, while ideally using goal-relevant stimuli of comparable relevance to that of threat-relevant stimuli (see Stussi, Pourtois, et al., 2018). Our framework would also predict that individual differences in specific survival-relevant concerns would cause various degrees of preferential aversive learning to threat-related stimuli.

Relatedly, the lack of differential resistance-to-extinction effects may tentatively be imputed to the specifics of our manipulation of goal-relevance. In particular, the use of a spatial cueing task in which the cues were presented exogenously for a brief amount of time (100 ms) may have precluded participants from forming an explicit and strong knowledge of the associations between the cues and the stimulus categories, and mainly tapped into implicit processes. Consistent with this proposition, the subjective ratings (see 3.3.6. Supplementary materials) suggested that participants did not discriminate the differential predictive value of the different stimuli used as cues during the spatial cueing task. In this context, the relevance manipulation was probably too weak to induce long-lasting effects that could as well influence the persistence of the conditioned response. Future studies are therefore needed to assess whether a stronger relevance manipulation, for instance by using an endogenous cueing task allowing participants to integrate information about the stimuli's goal-relevance at a more explicit, controlled level (Desimone & Duncan, 1995), could lead to a differential resistance-to-extinction effect for goal-relevant stimuli compared with goal-irrelevant stimuli, besides faster Pavlovian aversive learning.

Whereas our manipulation of goal-relevance by means of a spatial cueing task was probably subtle, the subjective ratings collected after extinction (see 3.3.6. Supplementary materials) clearly reflected that participants were aware of the contingencies between the conditioned stimuli and the unconditioned stimulus, and that the conditioning procedure elicited robust evaluative effects, the CSs+ being evaluated as less pleasant, more arousing, and more relevant than the CSs- (see Figure S3.3.1). Presumably, this potent conditioning procedure may have overshadowed "residual" relevance effects produced by the preceding

spatial cueing task, the salience of the CSs+ association with an electric stimulation prevailing over the stimuli's previously acquired goal-relevance, especially when considering the extinction phase. This too could potentially account for the fact that we observed faster Pavlovian aversive conditioning to goal-relevant stimuli than to goal-irrelevant stimuli in participants high in achievement motivation, but did not find differential extinction effects as a function of stimulus' goal-relevance and achievement motivation.

As goal-relevant stimuli have been shown to attract attention (Pool, Brosch, et al., 2016; Vogt, De Houwer, Moors, Van Damme, & Crombez, 2010), it is possible that the goal-relevant stimuli induced facilitated acquisition of a conditioned response in participants high in achievement motivation because more attention was allocated to them than to the goal-irrelevant stimuli. Given that the goal-relevant stimuli were also highly predictive with respect to target location in the spatial cueing task while the goal-irrelevant stimuli were associated with a high uncertainty, this suggestion aligns with the Mackintosh's (1975) attentional model of Pavlovian conditioning. According to this model, the amount of attention devoted to the conditioned stimulus is a core determinant of learning, with predictive stimuli being better attended and hence more readily conditioned. In this light, attention could provide an underlying mechanism contributing to the occurrence of learning bias to goal-relevant stimuli in participants high in achievement motivation, thus possibly mirroring the contribution of attention to the enhancement effects of emotion on memory for instance (Talmi et al., 2013).

Further consideration of the role of predictiveness and uncertainty additionally raises the question of whether these constructs may have influenced our findings. Predictiveness and uncertainty have been shown to affect associative learning and attentional processes (Beesley, Nguyen, Pearson, & Le Pelley, 2015; Le Pelley, Mitchell, Beesley, George, & Willis, 2016; Mackintosh, 1975; Pearce & Hall, 1980), in particular through their impact on stimulus' salience (Esber & Haselgrove, 2011; Mackintosh, 1975; Pearce & Hall, 1980) or informativeness (Walentowska et al., 2018), as well as are considered as an important evaluation criterion for determining the relevance of a stimulus in appraisal theories (Sander et al., 2005). Although the cues' predictiveness and/or uncertainty may have had a general influence on their appraised relevance and contributed to our findings, it seems unlikely that our results were solely driven by these factors. Indeed, it remains unclear to what extent such an account can accommodate the observed effects of inter-individual differences in achievement motivation on the acquisition readiness of the conditioned response to goal-relevant versus goal-irrelevant stimuli, without requiring the involvement of additional

explanatory mechanisms directly tied to the organism's achievement-related concerns. Accordingly, it appears that goal-relevance offers a more parsimonious and plausible key mechanistic explanation of our findings. Further research would nevertheless be necessary to disentangle the specific contributions of predictiveness and/or uncertainty and of goal-relevance to faster Pavlovian aversive conditioning, for instance by implementing a paradigm enabling the orthogonalisation of these factors (Walentowska et al., 2018).

Considering that our sample mostly consisted of women participants, we cannot be sure that our results can generalise to men, which represents a limitation of our study. As women and men can differ in conditioned threat acquisition (e.g., Milad et al., 2006), it could be possible that the modulation of Pavlovian aversive conditioning to goal-relevant versus goal-irrelevant stimuli may have been affected by sex differences in achievement motivation. However, women (M = 4.04, SD = 0.82) and men (M = 4.32, SD = 0.75) participants in our sample did not statistically differ in achievement motivation scores, as reflected by a Welch's t test for unequal sample sizes, $t(18.79) = -1.16$, $p = .262$, $g_s = -0.332$, 95% CI [-0.956, 0.250]. This result thereby provides no evidence that sex differences in achievement motivation influenced our results. Another caveat relates to the fact that we did not consider the role of the hormonal cycle stage of our women participants, which has been shown to affect skin conductance response, notably during extinction learning (Milad et al., 2006). Although we cannot exclude the possibility that this factor may have had an effect on our results, we are not aware of any empirical evidence suggesting that the hormonal cycle stage specifically facilitates the acquisition of a conditioned response to certain categories of stimuli, such as goal-relevant stimuli in the present case, relative to other stimulus categories, such as goal-irrelevant stimuli.

In sum, our study suggests that stimuli without any inherent biological evolutionary significance but temporarily associated with a higher goal-relevance can also induce facilitated Pavlovian aversive learning provided that specific individual motivation dispositions are met concurrently, thus reflecting that the occurrence of a learning bias is crucially dependent on inter-individual differences in the organism's current concerns. In the present case, the learning bias towards goal-relevant stimuli in comparison with goal-irrelevant stimuli was expressed as a greater conditioned response acquisition, and, importantly, as a facilitated conditioned response acquisition in participants scoring high on achievement motivation, whereas no effect on the persistence of the conditioned response was observed. Although the impact of goal-relevance was modest and transient, these findings lean towards the view that Pavlovian

aversive conditioning may be driven by a general mechanism of relevance detection that is not necessarily selective for stimuli holding a pre-existing threat value (Stussi, Pourtois, et al., 2018). This mechanism yields flexibility in the way specific stimuli encountered in the environment are eventually learned preferentially, depending primarily on the complex interplay between the stimulus at hand and the organism's current concerns. Hence, relevance detection provides a flexible theoretical framework that can not only incorporate the extant evidence on preferential Pavlovian aversive learning in the human conditioning literature but also account for the large inter-individual differences typically observed in human emotional learning. In this perspective, the relevance detection approach holds promise for contributing to an improved mechanistic understanding of emotional learning in humans. Ultimately, this alternative framework could also contribute to unravelling emotional learning impairments preceding or following the onset and maintenance of specific affective disorders, such as anxiety disorders and phobias, thus hopefully aiding in developing and validating new individualised and targeted interventions for these conditions.

3.3.4. Methods

Participants

Eighty-eight participants took part in the experiment, which was approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva. All ethical regulations were complied with. Sixteen participants were excluded from the analyses based on predetermined criteria (Olsson et al., 2005; Stussi et al., 2015; Stussi, Pourtois, et al., 2018): seven because of technical problems, three for displaying virtually no SCR, and six for failing to acquire a conditioned response to at least one of the conditioned stimuli predictive of the unconditioned stimulus. The final sample size consisted of 72 participants (59 women), aged between 18 and 70 years old (mean age = 22.67 ± 7.58 years).

We established the sample size prior to data collection by means of a power analysis conducted with G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), which indicated that a total sample of 71 participants would be required to obtain a power of 80% to detect a relatively small effect (d = 0.3) as reported in a previous study (Flykt et al., 2007). For counterbalancing purposes, we sought to recruit a sample of 72 participants that were conditioned to at least one of the three stimulus categories, and stopped data collection when the required number of participants had been reached.

Stimuli and apparatus

Six neutral complex geometric figures commonly used in human conditioning paradigms (Gottfried, O'Doherty, & Dolan, 2003; Pool, Brosch, et al., 2014) served as cues in the spatial cueing task and subsequently as conditioned stimuli (CSs) in the Pavlovian differential aversive conditioning paradigm. The geometric figures were divided into three stimulus categories as a function of their goal-relevance and predictive power of target location in the spatial cueing task: (a) the *goal-relevant valid stimuli*, which consistently predicted target location, (b) the *goal-relevant invalid stimuli*, which consistently predicted the opposite location relative to the target, and (c) the *goal-irrelevant stimuli*, which were nonpredictive of target location by predicting target location and the opposite location with an equal probability (50%). The goal-relevant valid and the goal-relevant invalid geometric figures allowed participants to anticipate target location, and were therefore relevant for the spatial cueing task goals. By contrast, the goal-irrelevant geometric figures were uninformative about upcoming target location, thus being irrelevant for the spatial cueing task. We used two types of goal-relevant stimuli in order to be able to dissociate a general effect of goal-relevance from a mere cue (in)validity effect. The attribution of the stimulus categories to the six geometric figures were counterbalanced across participants. In the differential Pavlovian aversive conditioning procedure, one geometric figure from each of the three stimulus categories served as a CS+, whereas the other one served as a CS-; this assignment being counterbalanced across participants. The unconditioned stimulus (US) was a mild electric stimulation (200-ms duration, 50 pulses/s) delivered to the participants' nondominant wrist through a Grass SD9 stimulator (Grass Medical Instruments, West Warwick, RI) charged by a stabilized current.

The conditioned response (CR) was assessed through SCR measured with two Ag-AgCl electrodes (6-mm contact diameter) filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the index and middle fingers of the participants' nondominant hand. The SCR data was continuously recorded at 1000 Hz with a BIOPAC MP150 system (Santa Barbara, CA) and analysed offline with AcqKnowledge software (Version 4.2; BIOPAC Systems Inc., Goleta, CA).

Procedure

Upon arrival at the laboratory, participants were informed about the general procedure of the experiment and provided written informed consent. They next performed the spatial cueing task. Participants were then asked to rate the geometric figures on several dimensions

(see 3.3.6. Supplementary materials) before undertaking the differential Pavlovian aversive conditioning procedure. After the end of the conditioning procedure, they were again asked to provide subjective ratings of the geometric figures (see 3.3.6. Supplementary materials). Finally, participants completed the Unified Motive Scales (UMS; Schönbrodt & Gerstenberg, 2012) to measure their achievement motivation.

Spatial cueing task. In this task (Pool, Brosch, et al., 2014; Figure 3.3.1), each trial started with a fixation cross presented for a duration randomly varying between 250 and 750 ms. A cue was subsequently presented either on the left or the right side of the fixation cross for 100 ms. The cues consisted of the six geometric figures, divided into the three stimulus categories (i.e., two goal-relevant valid cues, two goal-relevant invalid cues, and two goal-irrelevant cues). Following a brief interval after the cue was removed (blank screen; 16.7 ms), a target consisting of a black bar was presented onscreen for 100 ms. Participants were requested to press as quickly and accurately as possible with the second digit of their dominant hand the "B" key when the target was displayed horizontally and the "N" key when it was displayed vertically, and their reaction times and accuracy were measured. The target appeared either at the same location as the cue (valid trial) or at the opposite location (invalid trial; Figure 3.3.1). After participants' response, each trial ended with an intertrial interval randomly varying between 700 and 1300 ms. Participants were asked to look at the fixation cross during the entire task.

Participants first undertook a training session of 24 trials. Each of the six cues was presented four times. The training session was repeated until participants reached an accuracy of 75%, after which the experimental task started. It was composed of 144 trials, divided into 48 trials for each stimulus category, each cue being presented 24 times. During both the training session and the experimental task, the valid and invalid trials were equally presented, and the left or right position of the cue and the target, as well as the horizontal and vertical orientation of the target, were counterbalanced, and the order of the trials pseudorandomised. All responses that were incorrect (4.06% of the trials), faster than 200 ms (0.09% of the trials), or more than three standard deviations from the participant's mean (1.63% of the trials) were removed prior to analysis (Pool, Brosch, et al., 2014).

Differential Pavlovian aversive conditioning. Before conditioning, the electrodes for measuring SCR were placed on participants and a work-up procedure was conducted to individually set the electric stimulation intensity (M = 33.73 V, SD = 9.48, range = 10-50 V) to a level reported as "uncomfortable, but not painful". During the initial habituation phase,

each of the six geometric figures serving as CSs was presented twice without being reinforced. In the following acquisition phase, each CS was presented seven times. This phase always began with a CS+ trial. Five of the seven presentations of each CS+ coterminated with an electric stimulation delivery, whereas the CSs- were never paired with the US. In the extinction phase, each CS was presented six times and the US was no longer delivered. During all the conditioning phases, the CSs were presented for 6 s with a variable intertrial interval ranging from 12 to 15 s. The CSs' order of presentation was pseudorandomised into 12 different orders.

Unified Motive Scales (UMS). At the end of the experiment, participants filled out the UMS (Schönbrodt & Gerstenberg, 2012). This questionnaire offers an explicit measure of individuals' motives. It is composed of 54 items measured on a 6-point Likert scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). These items assess various types of motivation, including achievement motive, power motive, affiliation motive, intimacy motive, fear of losing control, fear of failure, fear of rejection, and fear of losing emotional contact (Schönbrodt & Gerstenberg, 2012). Given our a priori hypotheses, we exclusively focused on the achievement motive subscale, which comprised 10 items (standardised Cronbach's α = .85). Each participant's responses to these items were averaged to compute their achievement motivation score (M = 4.09, SD = 0.81, range = 2.0-5.5; see Figure S3.3.2).

Response definition

SCR was scored for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window following CS onset. The minimal response criterion was 0.02 μS. Responses below this criterion were scored as zero and remained in the analysis. Before analysis, the SCR data was low-pass filtered (Blackman -92 dB, 1 Hz). SCRs were detected automatically with AcqKnowledge software and manually checked for artefacts and response (mis)detection. Trials containing artefacts influencing the coding of event-related SCRs (< 0.001%) were omitted from the analyses. The raw SCRs were square-root-transformed to normalise the distributions and scaled according to each participant's mean square-root-transformed unconditioned response (UR). The UR was scored as the peak-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5-4.5 s temporal window after the US delivery. The habituation means were composed of the first two presentations of each CS. To investigate the CR acquisition readiness, the acquisition means were split into an early (i.e., the first three presentations of each CS subsequent to the first pairing between the CS+ from the stimulus category and the US) and a late (i.e., the following three presentations of each CS) phase (Lonsdorf et al., 2017;

Stussi et al., 2015; Stussi, Pourtois, et al., 2018). Because the CSs+ became predictive of the US solely after their first association with the electric stimulation, the first acquisition trial for each CS was removed from the analyses. The extinction means included the last six presentations of each CS. The conditioning data analyses were performed on the CR, which was computed as the SCR to the CS+ minus the SCR to the CS- from the same stimulus category (Olsson et al., 2005; Stussi et al., 2015; Stussi, Pourtois, et al., 2018).

Statistical analyses

Statistical analyses were performed with R (R Core Team, 2018) and the afex package (Singmann, Bolker, Westfall, & Aust, 2018). An alpha level of .05 was adopted for all the analyses performed. When descriptive analyses revealed the presence of outliers (value smaller than the lower quartile minus three times the interquartile range, or value larger than the upper quartile plus three times the interquartile range; Tukey, 1977), we conducted the analyses including and excluding the outliers, and report the outcome of both analyses when the outlier removal altered statistical significance. Otherwise, we only report the results of the analyses including the outliers. We report partial η^2 or Hedges' g_{av} and their 90% or 95% confidence interval, respectively, as estimates of effect sizes.

Code availability

The code used for data analysis can be found in the Open Science Framework repository at https://doi.org/10.17605/OSF.IO/EQA6S.

3.3.5. Data availability

The datasets generated and analysed during the current study, as well as the materials used therein, are available in the Open Science Framework repository at https://doi.org/10.17605/OSF.IO/EQA6S.

3.3.6. Supplementary materials

Supplementary methods and results

Subjective ratings

Subsequent to the spatial cueing task, participants completed subjective ratings of the geometric figures' (a) predictive power, (b) liking, (c) arousal, and (d) relevance. In this procedure, the geometric figures were presented to participants, accompanied by a visual analogue scale (VAS). For the prediction power ratings, participants were asked to rate to what extent the geometric figure was predictive of target location on the same side from 0 (*never*) to 100 (*always*). For the liking ratings, they were asked to rate to what extent the geometric figure was unpleasant or pleasant from 0 (*very unpleasant*) to 100 (*very pleasant*). For the arousal and the relevance ratings, participants were asked to rate to what extent the geometric figure was arousing from 0 (*not at all arousing*) to 100 (*very arousing*), and to what extent it was important to them from 0 (*not at all important*) to 100 (*very important*), respectively.

After the extinction phase of the conditioning procedure, participants completed (a) CS-US contingency ratings, along with (b) liking, (c) arousal, and (d) relevance ratings of the geometric figures. For the CS-US contingency ratings, they were asked to rate to what extent the geometric figure was predictive of the delivery of an electric stimulation on a VAS going from 0 (*never*) to 100 (*always*). The procedure for the liking, arousal, and relevance ratings was identical to the one used in the preconditioning ratings. The order of the geometric figure presentations and the questions was randomised across participants for both the preconditioning and the postconditioning subjective ratings. Finally, participants rated to what extent it was important to them to perform well in the spatial cueing task on a VAS ranging from 0 (*not at all important*) to 100 (*very important*; $M = 74.87$, $SD = 15.88$, range = 17.05-100).

The preconditioning ratings of the stimuli's predictive power were analysed with a repeated-measures general linear model (GLM) assuming compound symmetry covariance structure including stimulus type (CS+ vs. CS-) and stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) as within-participant categorical factors, and participants' standardised (z-score) achievement motivation score as a continuous predictor. This analysis showed that there was no statistical difference across the three stimulus categories, or as a function of stimulus type, achievement motivation, or the interaction between any of these factors in the predictive power ratings of the stimuli (all Fs < 1.07, all ps

> .34, all η^2_ps < .016; Figure S3.3.1a). These results tentatively suggest that participants did not seem to be able to explicitly distinguish the predictive power of the different stimuli used as cues during the spatial cueing task.

The liking, the arousal, and the relevance ratings were each analysed using a repeated-measures GLM assuming compound symmetry covariance structure including stimulus type (CS+ vs. CS-), stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant), and time (pre vs. post) as within-participant categorical factors, and participants' standardised achievement motivation score as a continuous predictor. Analysis of the liking ratings revealed statistically significant main effects of stimulus type, $F(1, 70) = 10.18$, $p = .002$, $\eta^2_p = .127$, 90% CI [.029, .249], and of time, $F(1, 70) = 12.27$, $p < .001$, partial $\eta^2 = .149$, 90% CI [.042, .274]. These main effects were however qualified by the interaction between stimulus type and time, $F(1, 70) = 20.09$, $p < .001$, partial $\eta^2 = .223$, 90% CI [.093, .350]. Follow-up simple effects analyses indicated that there was no statistical difference in the preconditioning liking ratings of the CSs+ and the CSs-, $F(1, 70) = 0.15$, $p = .696$, partial $\eta^2 = .002$, 90% CI [.000, .050], whereas the CSs+ were rated as less pleasant than the CSs- after conditioning, $F(1, 70) = 17.00$, $p < .001$, partial $\eta^2 = .195$, 90% CI [.072, .322] (Figure S3.3.1b). No other effect reached statistical significance (all Fs < 3.13, all ps > .08, all partial η^2s < .043).

For the arousal ratings, the analysis showed a statistically significant main effect of stimulus type, $F(1, 70) = 53.43$, $p < .001$, partial $\eta^2 = .433$, 90% CI [.285, .541], which was qualified by the higher-order interaction between stimulus type and time, $F(1, 70) = 65.29$, $p < .001$, partial $\eta^2 = .483$, 90% CI [.338, .584]. Simple effects analysis showed that participants did not statistically differ in their subjective ratings of the CSs+ and the CSs- arousal value before conditioning, $F(1, 70) = 0.95$, $p = .334$, partial $\eta^2 = .013$, 90% CI [.000, .087], but deemed the CSs+ more arousing than the CSs- after it, $F(1, 70) = 75.92$, $p < .001$, partial $\eta^2 = .520$, 90% CI [.380, .615] (Figure S3.3.1c). No other effect was found for the arousal ratings (all Fs < 3.15, all ps > .08, all partial η^2s < .044).

The relevance ratings analysis revealed main effects of stimulus type, $F(1, 70) = 15.87$, $p < .001$, partial $\eta^2 = .185$, 90% CI [.065, .311], and of time, $F(1, 70) = 14.36$, $p < .001$, partial $\eta^2 = .170$, 90% CI [.055, .296]. These main effects were however qualified by their interaction, $F(1, 70) = 18.42$, $p < .001$, partial $\eta^2 = .208$, 90% CI [.082, .335]. Further simple effects analyses showed no statistical difference in relevance ratings between the CSs+ and the CSs- prior to conditioning, $F(1, 70) = 0.96$, $p = .330$, partial $\eta^2 = .014$, 90% CI [.000, .087], whereas the CSs+ were evaluated as more relevant than the CSs- after it, $F(1, 70) = 19.67$, $p < .001$,

partial η^2 = .219, 90% CI [.090, .346] (Figure S3.3.1d). A statistically significant three-way interaction between stimulus category, time, and participants' achievement motivation was additionally found, $F(2, 140) = 3.41$, $p = .036$, partial $\eta^2 = .046$, 90% CI [.002, .106]. Follow-up simple slopes analyses revealed a marginal trend for the simple interaction effect of stimulus category and time in participants high in achievement motivation (+1 *SD*), $F(2, 140) = 2.64$, $p = .075$, partial $\eta^2 = .036$, 90% CI [.000, .091], reflecting that these participants evaluated the goal-relevant invalid ($p = .028$) and the goal-irrelevant ($p < .001$), but not the goal-relevant valid ($p = .186$), stimuli as more relevant after than before conditioning, whereas no simple interaction effect was observed in participants lower in achievement motivation (-1 *SD*), $F(2, 140) = 1.18$, $p = .311$, partial $\eta^2 = .017$, 90% CI [.000, .057]. No other effect yielded statistical significance (all Fs < 3.28, all ps > .06, all partial η^2s < .045).

A repeated-measures GLM assuming compound symmetry covariance structure including stimulus type (CS+ vs. CS-) and stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) as within-participant categorical factors and participants' standardised achievement motivation score as a continuous predictor was used to analyse the postconditioning CS-US contingency ratings. This analysis showed that the CSs+ were deemed more likely to be predictive of the US than the CSs-, $F(1, 70) = 164.38$, $p < .001$, partial $\eta^2 = .701$, 90% CI [.599, .763] (Figure S3.3.1e). They conversely did not statistically differ as a function of stimulus category or participants' achievement motivation, and no interaction effect was observed (all Fs < 2.21, all ps > .11, all partial η^2s < .031).

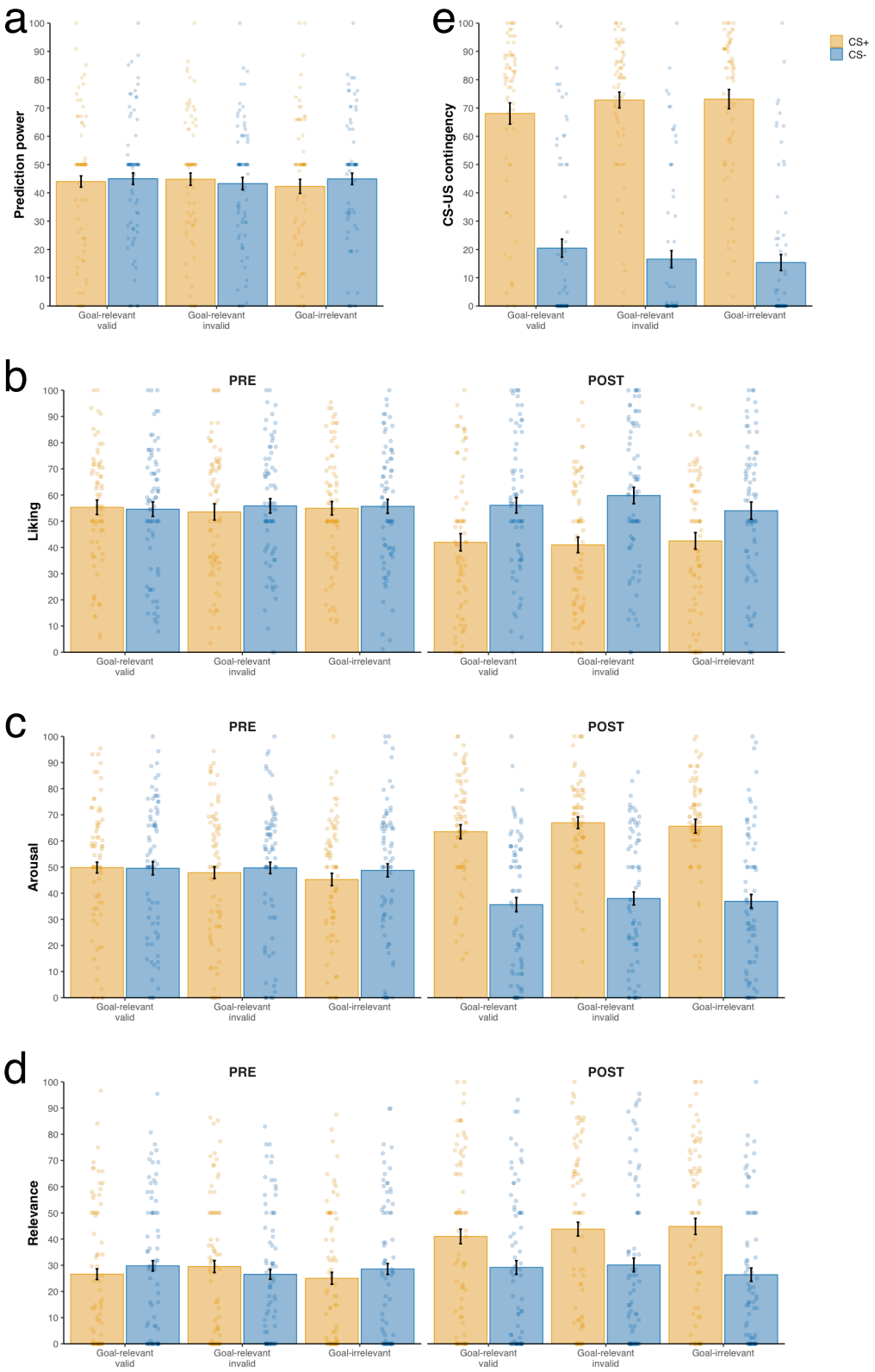

Figure S3.3.1. Mean subjective ratings before (pre) and after (post) the conditioning procedure as a function of conditioned stimulus type (CS+ vs. CS-) and stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant). Mean (a) prediction power ratings, (b) liking ratings, (c) arousal ratings, (d) relevance ratings, and (e) CS-US contingency ratings. The dots indicate data for individual participants. Error bars indicate ± 1 standard error of the mean adjusted for within-participant designs.

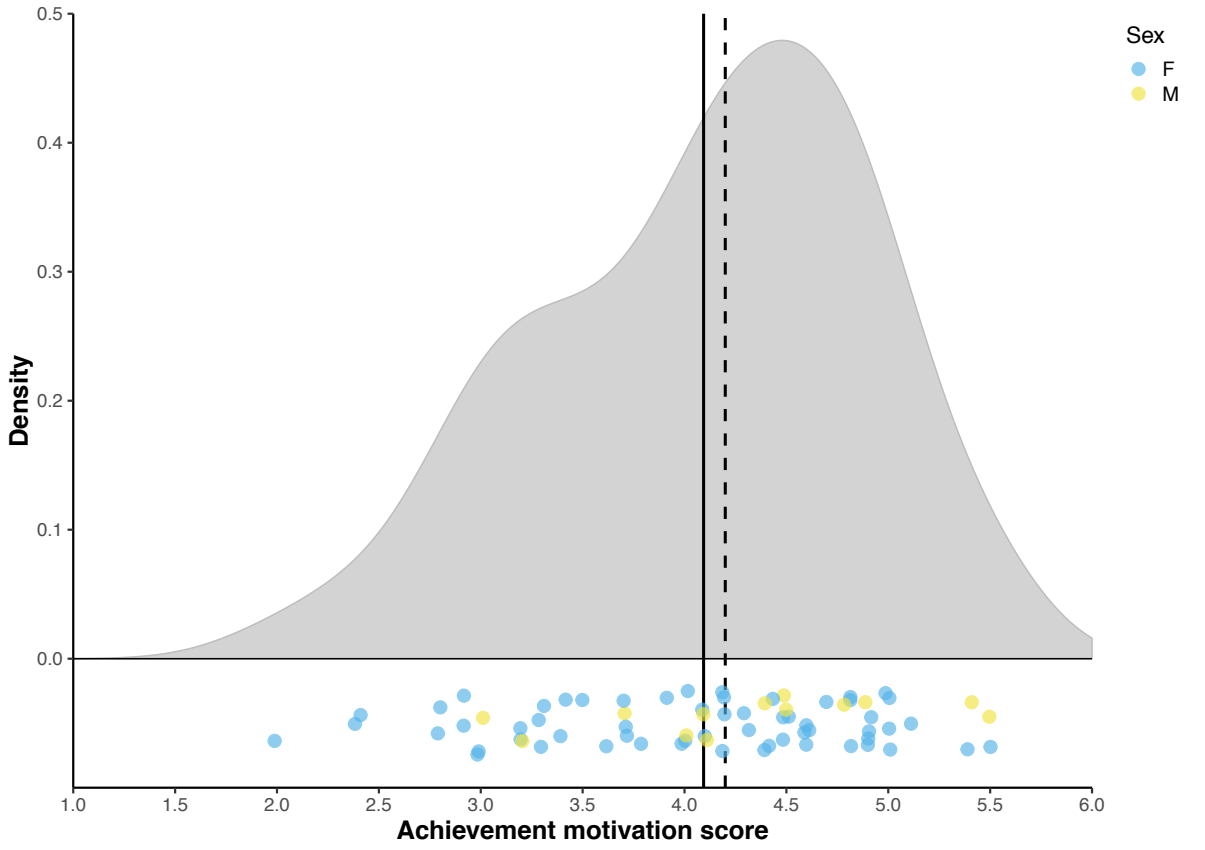

Figure S3.3.2. Distribution of achievement motivation scores as measured with the Unified Motive Scales (Schönbrodt & Gerstenberg, 2012). The dots indicate data for individual participants. The solid line indicates the mean achievement motivation score, and the dashed line indicates the median achievement motivation score.

3.4. STUDY 4:

MEASURING PAVLOVIAN APPETITIVE CONDITIONING IN HUMANS WITH THE POSTAURICULAR REFLEX[13]

Abstract

Despite its evolutionary and clinical significance, appetitive conditioning has been rarely investigated in humans. It has been proposed that this discrepancy might stem from the difficulty in finding suitable appetitive stimuli that elicit strong physiological responses. However, this might also be due to a possible lack of sensitivity of the psychophysiological measures commonly used to index human appetitive conditioning. Here, we investigated whether the postauricular reflex – a vestigial muscle microreflex that is potentiated by pleasant stimuli relative to neutral and unpleasant stimuli – may provide a valid psychophysiological indicator of appetitive conditioning in humans. To this end, we used a delay differential appetitive conditioning procedure, in which a neutral stimulus was contingently paired with a pleasant odor (CS+), while another neutral stimulus was not associated with any odor (CS-). We measured the postauricular reflex, the startle eyeblink reflex, and skin conductance response (SCR) as learning indices. Taken together, our results indicate that the postauricular reflex was potentiated in response to the CS+ compared with the CS-, whereas this potentiation extinguished when the pleasant odor was no longer delivered. In contrast, we found no evidence for startle eyeblink reflex attenuation in response to the CS+ relative to the CS-, and no effect of appetitive conditioning was observed on SCR. These findings suggest that the postauricular reflex is a sensitive measure of human appetitive conditioning and constitutes a valuable tool for further shedding light on the basic mechanisms underlying emotional learning in humans.

[13] Reprint of: Stussi, Y., Delplanque, S., Coraj, S., Pourtois, G., & Sander, D. (2018). Measuring Pavlovian appetitive conditioning in humans with the postauricular reflex. *Psychophysiology, 55*, e13073. https://doi.org/10.1111/psyp13073

3.4.1. Introduction

Learning to predict the presence of potentially harmful or beneficial events in the environment is a critical adaptive function that enables organisms to shape appropriate behaviors fostering survival and reproduction. This kind of learning principally occurs through Pavlovian aversive and appetitive conditioning processes. In Pavlovian conditioning, the organism learns to associate an environmental stimulus (the conditioned stimulus, CS) with a motivationally salient aversive or appetitive stimulus (the unconditioned stimulus, US) through one or several contingent pairings (Pavlov, 1927; Rescorla, 1988b).

While aversive conditioning has been extensively studied both in animals and humans (e.g., Delgado, Olsson, & Phelps, 2006; LaBar & Cabeza, 2006; Phelps & LeDoux, 2005), appetitive conditioning has been rarely investigated systematically in humans (Andreatta & Pauli, 2015; Hermann, Ziegler, Birmbauer, & Flor, 2000; Martin-Soelch, Linthicum, & Ernst, 2007). This paucity and asymmetry is rather surprising given that Pavlovian appetitive processes are considered to play a central role in reward processing (Berridge & Robinson, 2003; Pool, Sennwald, Delplanque, Brosch, & Sander, 2016) and to represent a crucial mechanism in the etiology, maintenance, and treatment of several major psychiatric conditions, including depression, addiction, and eating disorders (Martin-Soelch et al., 2007). It has been proposed that this discrepancy might be explained by the difficulty in finding appropriate appetitive stimuli that are able to elicit physiological responses that are similarly intense to the ones elicited by the aversive USs (e.g., electric stimulations) used in aversive conditioning (Hermann et al., 2000; Martin-Soelch et al., 2007), thereby resulting in potentially subtler effects (see Rescorla & Wagner, 1972). However, this discrepancy might also stem from a possible lack of sensitivity of the psychophysiological measures commonly used to systematically detect physiological changes induced by appetitive conditioning.

In line with this suggestion, human appetitive conditioning has generally been successfully evidenced using subjective measures (e.g., US expectancy and CS valence ratings; Van Gucht, Baeyens, Vansteenwegen, Hermans, & Beckers, 2010; Van Gucht, Vansteenwegen, Van den Bergh, & Beckers, 2008), behavioral measures (e.g., reaction times; Pool, Brosch, Delplanque, & Sander, 2014; Pool, Delplanque, et al., 2014; Van Gucht et al., 2008), or brain activity (e.g., Delgado, 2007; Franken, Huijding, Nijs, & van Strien, 2011; Gottfried, O'Doherty, & Dolan, 2002, 2003; Klucken et al., 2009; Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013), whereas the use of peripheral physiology measures (e.g., skin conductance response, SCR) has mainly yielded mixed or inconclusive results (see, e.g.,

Hermann et al., 2000). Developing psychophysiological indicators of appetitive conditioning thus constitutes an important purpose to eventually remedy the scarcity of knowledge about key mechanisms involved in emotional learning in humans.

In this vein, Andreatta and Pauli's (2015) study recently suggested that the startle reflex – an automatic defensive response to a sudden, intense, and unexpected stimulus – might be a putative index of human appetitive conditioning. In this study, the authors implemented a concurrent differential aversive and appetitive conditioning paradigm, in which three types of CS were used: One stimulus (aversive CS+) was associated with an electric stimulation (i.e., aversive US), one stimulus (appetitive CS+) was paired with sweet or salty food (i.e., appetitive US), and another stimulus (CS-) was not associated with any US. Overall, the aversive CS+ was rated as more negative and more arousing than the CS-, and elicited enhanced SCRs, while the appetitive CS+ was rated as more positive and also induced larger SCRs than the CS-, but was not rated as more arousing. Of particular interest, the startle eyeblink reflex was potentiated in response to the aversive CS+ compared with the CS-, whereas it was attenuated in response to the appetitive CS+, thereby replicating key findings obtained in rodents (e.g., Koch, Schmid, & Schnitzler, 1996). These results concurred with prior research in the human startle literature indicating that the startle eyeblink reflex is specifically potentiated in response to unpleasant stimuli and attenuated in response to pleasant stimuli (Lang, Bradley, & Cuthbert, 1990). It has been, however, argued that the startle eyeblink response is primarily an index of the defensive motivational system, being hence optimal for studying aversive processes, but is not ideally suited for indexing appetitive processing (Dichter, Benning, Holtzclaw, & Bodfish, 2010). Although it is widely accepted that the startle eyeblink reflex does index defensive responding, mixed findings have been indeed reported regarding its role as an indicator of appetitive responding (Dillon & LaBar, 2005; D. C. Jackson, Malmstadt, Larson, & Davidson, 2000; for a review, see Grillon & Baas, 2003). Therefore, it remains unclear to what extent the startle eyeblink reflex is the most appropriate measure of appetitive conditioning in humans: The attenuation of this reflex may reflect an inhibition of defensive responding rather than appetitive responding per se.

In contrast, the postauricular reflex (PAR) has previously been suggested to provide a reliable index of appetitive processing (Benning, Patrick, & Lang, 2004; Sandt, Sloan, & Johnson, 2009). The PAR is a vestigial muscle microreflex in humans that serves to pull the ear backward and upward (Bérzin & Fortinguerra, 1993; H. Gray, 1901/1995). As for the eyeblink reflex, the PAR can be elicited with an acoustic startle probe. However, the PAR

latency is faster than the eyeblink reflex latency (9-11 ms vs. 45-50 ms, respectively; Hackley, Woldorff, & Hillyard, 1987), suggesting that these two reflexes do not share the same underlying neural circuitry (Hackley, 2015). Importantly, a key aspect of the PAR lies in its sensitivity to affective modulation. Accumulating evidence has demonstrated that the PAR magnitude is potentiated during presentation of pleasant stimuli relative to neutral or unpleasant stimuli (Aaron & Benning, 2016; Benning, 2011; Benning et al., 2004; Dichter et al., 2010; Gable & Harmon-Jones, 2009; Hackley, Muñoz, Hebert, Valle-Inclán, & Vila, 2009; Hebert, Valle-Inclán, & Hackley, 2015; Hess, Sabourin, & Kleck, 2007; Johnson, Valle-Inclán, Geary, & Hackley, 2012; Sandt et al., 2009) and in particular during viewing of appetitive images, such as food or erotic scenes (Sandt et al., 2009). These observations support the view that the PAR is an index of appetitive processing and accordingly suggest that the PAR may constitute a suitable psychophysiological measure for indexing human appetitive conditioning.

The current study therefore aimed to test whether appetitive conditioning may be measured with the PAR in humans. To this end, we applied a differential appetitive conditioning procedure, in which two initially neutral stimuli were presented. During the initial habituation phase, the two stimuli were presented without being reinforced. In the subsequent acquisition phase, one stimulus (CS+) was systematically paired with a pleasant odor (US), while the other stimulus (CS-) was not associated with any odor. We used a pleasant odor as US because pleasant odors have been shown to be an efficient primary reinforcer to trigger appetitive conditioning in humans (Gottfried et al., 2002, 2003; Pool, Brosch, et al., 2014; Pool, Brosch, Delplanque, & Sander, 2015). During the final extinction phase, the US was no longer delivered. The PAR, the startle eyeblink reflex, and SCRs were measured concurrently during all the conditioning phases as putative psychophysiological indices of appetitive conditioning, thus enabling a systematic comparison thereof. Subjective ratings were additionally collected after the conditioning procedure to assess learning at the subjective level. Our main hypothesis was that the PAR magnitude would be potentiated in response to the CS+ compared with the CS- during acquisition. Based on previous findings (Andreatta & Pauli, 2015), we also expected the CS+, in comparison with the CS-, to elicit larger SCRs, and a startle eyeblink reflex attenuation during acquisition.

3.4.2. Method

Participants

Sixty-three volunteers participated in the study, which was approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva. They received either partial course credit or monetary compensation for their participation. The sample size was determined prior to data collection with the aim of recruiting approximately 60 participants and based on previous research investigating the PAR in humans (Gable & Harmon-Jones, 2009; Hebert et al., 2015; Sandt et al., 2009). Eight participants were excluded from the analyses due to technical problems. The final sample consisted of 55 participants (34 women, 21 men), aged between 18 and 40 years old (mean age = 25.27 ± 5.56 years). From this sample, four participants (3 women, 1 man) were further excluded from the SCR analysis because of technical problems with the SCR recordings.

Stimuli and apparatus

Conditioned stimuli. The CSs were two neutral geometric figures commonly used in human conditioning paradigms (Gottfried et al., 2002, 2003; Pool, Brosch, et al., 2014; Pool et al., 2015; see Figure 3.4.1A). Each geometric figure served either as the CS+ or as the CS-, this assignment being counterbalanced across participants.

Unconditioned stimulus. The US consisted of a pleasant odor selected among a set of 17 different odors (Firmenich SA, Geneva, Switzerland; see Table 3.4.1). The odor that the participant rated as the most pleasant and intense was selected as the US for the appetitive conditioning procedure. More precisely, the most pleasant odor was chosen if its intensity was evaluated above or equal to a predefined threshold (i.e., 50 on a scale from 0 to 100). In case the intensity of the most pleasant odor was rated below this threshold, the second most pleasant odor was selected if (a) its intensity was rated as higher than the most pleasant odor and (b) the pleasantness difference score between the most pleasant and second most pleasant odor was below or equal to 10. Otherwise, the most pleasant odor was chosen. Given the high and inherent variability of affective responses to odors across individuals (e.g., Ferdenzi et al., 2013), this procedure was warranted to ensure that the selected odor was pleasant, sufficiently intense, and had rewarding properties for the participant, thus constituting an appropriate appetitive US. During both the US selection and appetitive conditioning procedures, the odors were released through a custom-made, computer-controlled olfactometer with an airflow fixed at 1 L/min delivering the olfactory stimulation rapidly, without thermal and tactile confounds,

via a nasal cannula (see Ischer et al., 2014; Pool, Brosch, et al., 2014; Pool et al., 2015; Pool, Delplanque, et al., 2014).

Table 3.4.1

Odors used in the unconditioned stimulus (US) selection procedure.

Odorant name	Odor family	Concentration (% in di-propylene glycol)	Mean liking (SD)	Mean intensity (SD)	Number of times selected as the US
Aladinate	Floral	50	32.95 (19.92)	63.49 (22.45)	0
Ariana	Detergent	20	64.69 (22.26)	66.96 (14.58)	10
Caramel	Sweet food	20	39.94 (25.01)	60.43 (19.27)	3
Chocolate	Sweet food	20	39.65 (26.38)	69.36 (20.88)	3
Galbex	Floral	50	57.23 (21.69)	52.69 (22.04)	3
Geraniol	Floral	50	39.32 (22.17)	59.32 (22.81)	2
Green tea	Floral green	50	50.72 (15.16)	33.43 (24.65)	1
Lavender	Floral	20	46.14 (23.78)	61.74 (20.14)	1
Linalol	Floral	50	50.85 (20.89)	49.55 (24.40)	2
Magnolia grandiflora	Floral	50	53.29 (23.91)	60.91 (20.18)	4
Peach	Fruity	50	56.05 (21.35)	45.39 (21.40)	1
Pine	Woody	33	48.88 (19.88)	48.64 (24.09)	1
Pipol	Herbal	20	29.63 (20.79)	65.19 (24.76)	0
Speculaas	Sweet food	20	39.42 (22.85)	61.74 (19.24)	1
Strawberry	Fruity	20	58.88 (19.30)	60.27 (21.30)	4
Tiare	Floral	50	48.97 (22.02)	51.76 (24.26)	3
Tutti frutti	Fruity	20	64.69 (25.24)	62.48 (23.42)	16

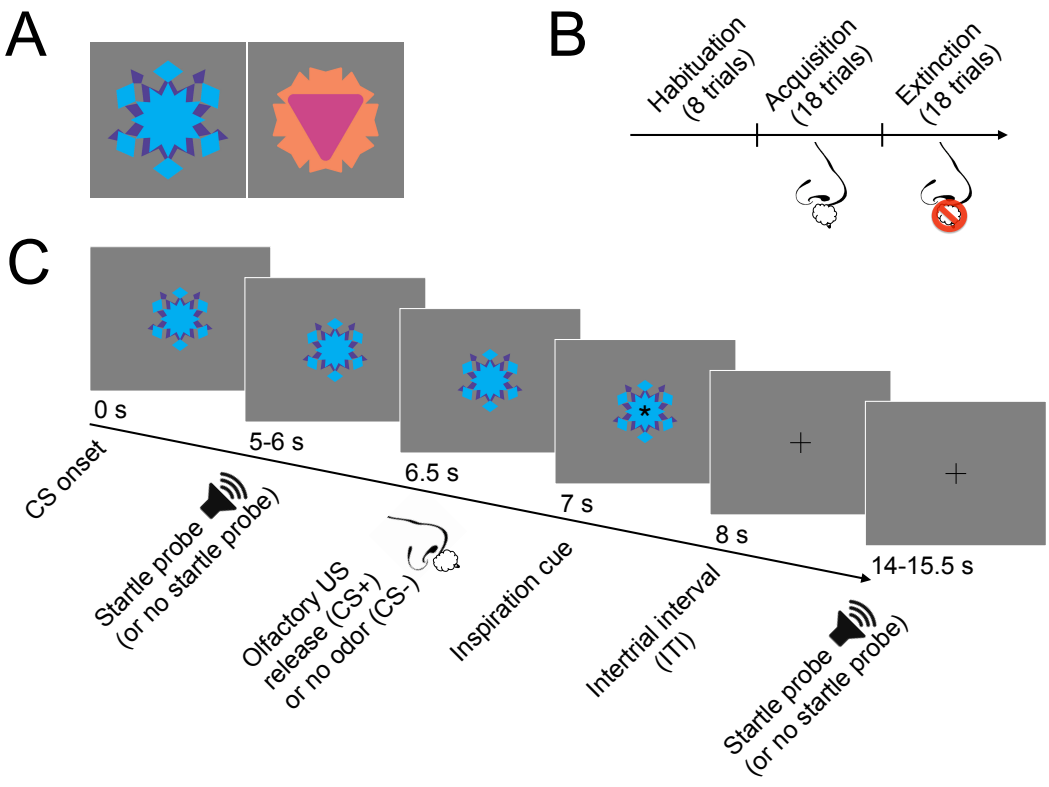

Figure 3.4.1. Experimental design. (A) Geometric figures used as conditioned stimuli. (B) Conditioning phases. (C) Trial structure during the acquisition phase.

Acoustic startle probe. The acoustic startle probe was a 50-ms white noise burst (105 dB) with a nearly instantaneous rise time (< 1 ms). The startle probe was presented binaurally through loudspeakers and delivered between 5 and 6 s after CS onset, or between 6 and 7.5 s after CS offset during intertrial intervals (ITIs).

Procedure

Prior to coming to the laboratory, participants were requested to refrain from eating before the experiment, which took place between 8.30 am and 12.30 pm. This procedure aimed to increase the likelihood that participants were in a hunger state, thereby optimizing the chances of the olfactory US to be rewarding, as is typically done in animal (e.g., Koch et al., 1996) and human (Andreatta & Pauli, 2015) appetitive conditioning studies.

Upon arrival at the laboratory, participants read and signed an informed consent form. They were then invited to provide background information, such as their age and gender, and to indicate their hunger level on a Likert scale from 1 (*not hungry at all*) to 10 (*very hungry*). Participants reported a mean hunger level of 5.75 (*SD* = 2.44). Next, the skin conductance

electrodes and the nasal cannula were attached to them. Subsequently, participants performed the US selection procedure, in which the various odors (see Table 3.4.1), along with odorless air, were delivered to them in a randomized order. Each trial started with a 3-s countdown followed by an inspiration cue that indicated to participants to breathe in evenly. The odors were released 0.5 s before the inspiration cue for a duration of 1.5 s. Participants were then asked to rate each odor according to its subjective pleasantness and intensity on visual analog scales (VASs) going from 0 (*extremely unpleasant* on the pleasantness VAS or *not perceived* on the intensity VAS) to 100 (*extremely pleasant* on the pleasantness VAS or *extremely strong* on the intensity VAS). Each trial ended with an ITI whose duration was adapted as a function of participants' rating pace (i.e., the ITI duration lasted for 15 s minus the time the participant took to rate the odor, with a minimal duration of 0.5 s).

Once the US selection procedure was completed, the electrodes for measuring the PAR and the startle eyeblink reflex were placed on participants. The room light was also turned dim to facilitate the acoustic startle reflex (Grillon, Pellowski, Merikangas, & Davis, 1997). Before the start of conditioning, 10 acoustic startle probes were delivered with an interstimulus interval randomly varying between 10 and 20 s to reduce the initial startle reactivity. The differential appetitive conditioning paradigm used a delay conditioning procedure and was composed of three contiguous phases (see Figure 3.4.1B). The habituation phase comprised four unreinforced presentations of each one of the two CSs. During the acquisition phase, each CS was presented nine times. Each CS+ trial co-terminated with the pleasant olfactory US, which was released 6.5 s after CS+ onset for a duration of 1.5 s (see Figure 3.4.1C), while the CS- trials were paired with odorless air. The extinction phase consisted of nine presentations of each CS, and no olfactory US was delivered during this phase. During all the conditioning phases, the CSs were presented for 8 s with an ITI ranging from 12 to 15 s, during which a fixation cross was presented onscreen (see Figure 3.4.1C). An inspiration cue indicating to participants to breathe in evenly was presented on each trial 7 s after CS onset (see Figure 3.4.1C). Startle probes were delivered on an equal number of trials for each CS (2 out of 4 during habituation, 6 out of 9 during acquisition, and 6 out of 9 during extinction). Additional startle probes were presented during ITIs (2 during habituation, 6 during acquisition, and 6 during extinction) between 6 and 7.5 s post-CS offset in order to decrease their predictability (see Figure 3.4.1C).

After the extinction phase, participants completed CS-US contingency and CS liking ratings to assess their awareness of the reinforcement contingencies and the evaluative effects

of appetitive conditioning, respectively. In this procedure, the CSs were presented again to participants and were accompanied by a VAS. For CS-US contingency, participants were asked to rate to what extent the stimulus was predictive of the pleasant odor delivery on a VAS going from 0 (*never*) to 100 (*always*). For CS liking, participants were asked to rate to what extent the stimulus was unpleasant or pleasant on a VAS going from 0 (*very unpleasant*) to 100 (*very pleasant*). The order of the CS presentations and the questions was randomized across participants.

Physiological recordings and response definition

Postauricular reflex and startle eyeblink reflex. The PAR was measured through electromyography (EMG) by pulling the left pinna forward and placing two 4-mm contact diameter Ag-AgCl electrodes filled with electrolyte gel on each side of the tendon of insertion for the PAR. One electrode was placed directly posterior to the tendon on the pinna surface, while the other electrode was placed over the postauricular muscle (Sollers & Hackley, 1997). The eyeblink reflex was measured through EMG recordings of the left orbicularis oculi muscle with two 4-mm contact diameter Ag-AgCl electrodes filled with electrolyte gel. Consistent with recent guidelines (Blumenthal et al., 2005), one electrode was placed below the lower left eyelid in line with the pupil in forward gaze and the second one 1-2 cm laterally. Two additional electrodes positioned on the top of the forehead were used as recording reference and ground electrodes (see http://www.biosemi.com/faq/cms&drl.htm for further information).

The EMG data were continuously recorded at 2048 Hz through a BIOSEMI Active-Two amplifier system (BioSemi Biomedical Instrumentation, Amsterdam, the Netherlands). The EMG analyses were carried out offline using Brain Vision Analyzer software (version 2.1; Brain Products GmbH, Gilching, Germany). Conventional bipolar montages were calculated from electrode pairs for the PAR and eyeblink reflex by subtracting the recorded activity of one electrode from the activity of the neighboring electrode. Prior to analysis, the PAR signal was band-pass (10-400 Hz) and notch filtered (50 Hz) before being rectified. The eyeblink reflex signal was bandpass (20-400 Hz) and notch filtered (50 Hz), rectified, and then low-pass filtered (40 Hz; see Blumenthal et al., 2005). The filtered EMG signals were segmented into epochs from 100 ms prior to startle probe onset to 250 ms after probe onset. The 50 ms prior to startle probe onset were used as a baseline. Each segment was visually inspected, and segments identified as containing excessive baseline shifts or blinks in progress were removed by hand from the analyses (4.16% of the trials for the PAR, and 4.16% of the trials for the eyeblink reflex).

Given its low signal-to-noise ratio as a microreflex, the PAR was scored after signal averaging of the rectified waveforms across trials within conditions (Aaron & Benning, 2016; Benning, 2011; Benning et al., 2004; Hackley et al., 1987, 2009; Hebert et al., 2015; Hess et al., 2007; Sollers & Hackley, 1997). The PAR magnitude was scored from the aggregate waveform as the baseline-to-peak amplitude for each condition. The peak was calculated as the maximum EMG activity occurring within a 5-35 ms time window after startle probe onset (Gable & Harmon-Jones, 2009; Sandt et al., 2009).

The startle eyeblink reflex was analyzed by means of a single-trial analysis, which corresponds to the most common method of analyzing eyeblink reflex data (Blumenthal et al., 2005). Accordingly, the eyeblink reflex was scored for each trial as the baseline-to-peak amplitude of the maximum EMG activity occurring within 21-120 ms after startle probe onset (Blumenthal et al., 2005). The raw eyeblink scores were standardized within participants using T scores. The eyeblink reflex magnitudes were calculated by averaging the T scores for each condition.

Skin conductance response. SCR was measured with two 6-mm contact diameter Ag-AgCl electrodes filled with 0.5% NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants' nondominant hand. The SCR data were recorded at 2000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). The SCR analysis was performed offline with AcqKnowledge software (version 4.2; BIOPAC Systems Inc., Goleta, CA). Before analysis, the SCR data were downsampled to 1000 Hz and low-pass filtered (1 Hz). SCR was scored for each trial as the peak-to-peak amplitude difference in skin conductance of the largest response occurring in the 0.5-4.5 s temporal window after CS onset. The minimal response criterion was 0.02 µS. Responses below this criterion were scored as zero and remained in the analysis. SCRs were detected automatically with an AcqKnowledge routine and manually screened for artifacts and misdetections. The raw SCRs were square-root-transformed to reduce the distributions' positive skew. The square-root-transformed SCRs were then scaled according to each participant's maximal square-root-transformed SCR in order to take into account individual differences (Lykken & Venables, 1971). The habituation means included the first four presentations of each CS. The acquisition means comprised the nine presentations of each CS following the first pairing between the CS+ and the US. The extinction means were composed of the last eight presentations of each CS following the first US omission.

Statistical analyses

Paired *t* tests were performed on the pleasantness and intensity ratings collected during the US selection procedure in order to ensure that the odor selected as the US was more pleasant and intense than odorless air. To assess whether there were differences in stimulus conditions in the conditioning phases, the PAR and the startle eyeblink reflex data were each analyzed with a one-way multivariate analysis of variance (MANOVA) with stimulus type (CS+ vs. CS- vs. ITI) as a within-participant factor and treating the habituation, acquisition, and extinction phases as multiple dependent variables. Separate one-way repeated measures ANOVAs with stimulus type (CS+ vs. CS- vs. ITI) as a within-participant factor were next conducted to investigate differences in stimulus conditions within each conditioning phase. Significant main effects were followed up with pairwise comparisons. To specifically test our a priori hypothesis, we performed a planned contrast comparing the PAR magnitude to the CS+ with the PAR magnitude to the CS- during acquisition. Likewise, we performed a planned contrast comparing the startle eyeblink reflex magnitude to the CS+ with the startle eyeblink magnitude to the CS- during the acquisition phase. Within each repeated measures ANOVA conducted, a stringent Bonferroni correction was applied on the pairwise comparisons' *p* value to correct for multiple testing (i.e., $3 \times p$). SCR was analyzed separately for habituation, acquisition, and extinction with paired *t* tests comparing the CS+ versus the CS-. We additionally conducted an exploratory correlational analysis using Pearson's correlation coefficients to investigate whether (a) the PAR potentiation to the CS+ during acquisition and/or (b) the CS+/CS- differentiation as measured by the PAR were associated with participants' subjective hunger level. Finally, the CS-US contingency and the CS liking ratings were each analyzed with a paired *t*-test comparing the CS+ versus the CS-.

An alpha level of .05 was adopted for all the statistical analyses performed. We provide the Huyhn-Feldt correction value (ε_{HF}) and the corrected *p* value for the one-way repeated measures ANOVAs. We moreover report either partial η^2 or Hedges' g_{av} as estimates of effect size (see Lakens, 2013) and their 90% or 95% confidence interval (CI), respectively.

3.4.3. Results

Olfactory US evaluation

The odor selected as the US was evaluated as more pleasant ($M = 83.84$, $SD = 13.53$) than odorless air ($M = 47.56$, $SD = 14.99$), $t(54) = 14.76$, $p < .001$, $g_{av} = 2.506$, 95% CI =

[1.952, 3.122]. Likewise, the odor selected as the US was rated as more intense (M = 70.19, SD = 16.59) than odorless air (M = 24.46, SD = 22.18), $t(54) = 12.82$, $p < .001$, $g_{av} = 2.302$, 95% CI = [1.764, 2.896].

Postauricular reflex

The multivariate omnibus test revealed a statistically significant difference between the stimulus types in the conditioning phases, $F(6, 49) = 3.44$, $p = .006$, Wilks's $\Lambda = .703$, partial $\eta^2 = .297$, 90% CI = [.056, .380][14]. The one-way repeated measures ANOVA for the habituation phase revealed a statistically significant main effect of stimulus type, $F(2, 108) = 5.31$, $p = .007$, $\varepsilon_{HF} = 0.98$, partial $\eta^2 = .090$, 90% CI = [.016, .173]. Follow-up comparisons showed that the PAR magnitude was greater during the ITI than to both the CS+, $t(54) = 3.01$, $p = .012$ (Bonferroni corrected), $g_{av} = 0.239$, 95% CI = [0.077, 0.406], and the CS-, $t(54) = 2.48$, $p = .048$ (Bonferroni corrected), $g_{av} = 0.224$, 95% CI = [0.042, 0.411] (see Figure 3.4.2A). These results replicate previous findings showing smaller PAR magnitudes during stimulus presentation than during ITIs (Benning, 2011; Benning et al., 2004), the PAR being generally inhibited by perceptual engagement with a stimulus (Benning, 2011; Hackley et al., 1987). Conversely, there was no statistical difference in PAR magnitude in response to the CS+ relative to the CS-, $t(54) = -0.11$, $p > .99$ (Bonferroni corrected), $g_{av} = -0.010$, 95% CI = [-0.184, 0.164] (see Figure 3.4.2A).

In the acquisition phase, a main effect of stimulus type was found, $F(2, 108) = 6.87$, $p = .003$, $\varepsilon_{HF} = 0.80$, partial $\eta^2 = .113$, 90% CI = [.029, .201]. Congruent with our a priori hypothesis, the PAR magnitude was potentiated to the CS+ compared with the CS-, $t(54) = 2.97$, $p = .013$ (Bonferroni corrected), $g_{av} = 0.095$, 95% CI = [0.030, 0.161] (see Figure 3.4.2B). Further comparisons revealed that the PAR magnitude was greater during the ITI than to the CS-, $t(54) = 3.33$, $p = .005$ (Bonferroni corrected), $g_{av} = 0.166$, 95% CI = [0.063, 0.271], whereas there was no statistical difference in PAR magnitude during the ITI relative to the CS+, $t(54) = 1.47$, $p = .444$ (Bonferroni corrected), $g_{av} = 0.074$, 95% CI = [-0.027, 0.177] (see Figure 3.4.2B).

[14] Although the aim of the present study was not to specifically assess changes between the stimulus types across the different conditioning phases, we nonetheless performed a two-way repeated measures ANOVA on the postauricular reflex data for the sake of completeness. This analysis revealed a statistically significant main effect of stimulus type, $F(2, 108) = 11.67$, $p < .001$, $\varepsilon_{HF} = 0.90$, partial $\eta^2 = .178$, 90% CI = [.069, .279], and a marginal main effect of phase, $F(2, 108) = 2.97$, $p = .063$, $\varepsilon_{HF} = 0.87$, partial $\eta^2 = .052$, 90% CI = [.000, .130], whereas the Stimulus Type × Phase interaction did not reach statistical significance, $F(4, 216) = 1.33$, $p = .266$, $\varepsilon_{HF} = 0.81$, partial $\eta^2 = .024$, 90% CI = [.000, .057] (but see 3.4.5. Supplementary materials, for the outcome of more powerful planned contrasts testing specific patterns of results for the postauricular reflex).

The one-way repeated measures ANOVA for extinction showed a statistically significant main effect of stimulus type, $F(2, 108) = 6.34$, $p = .004$, $\varepsilon_{HF} = 0.89$, partial $\eta^2 = .105$, 90% CI = [.024, .192]. Follow-up comparisons revealed that the PAR magnitude was larger during the ITI than to the CS-, $t(54) = 3.35$, $p = .004$ (Bonferroni corrected), $g_{av} = 0.184$, 95% CI = [0.071, 0.301], and marginally larger than to the CS+, $t(54) = 2.28$, $p = .080$ (Bonferroni corrected), $g_{av} = 0.135$, 95% CI = [0.016, 0.257] (see Figure 3.4.2C). Importantly, the PAR magnitude was no longer potentiated in response to the CS+ compared with the CS-, $t(54) = 0.95$, $p > .99$ (Bonferroni corrected), $g_{av} = 0.043$, 95% CI = [-0.047, 0.134] (see Figure 3.4.2C).

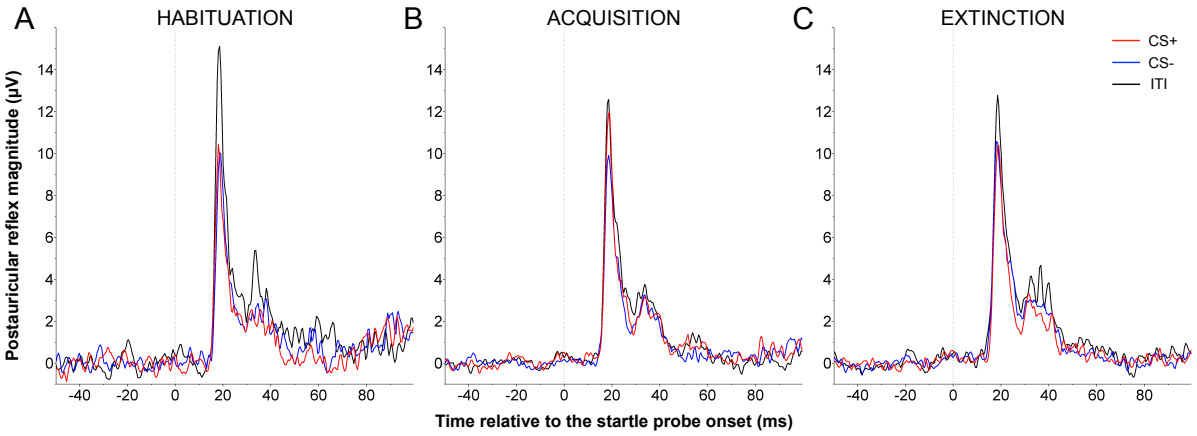

Figure 3.4.2. Grand-averaged postauricular reflex waveforms as a function of stimulus type (CS+ vs. CS- vs. ITI) across the (A) habituation, (B) acquisition, and (C) extinction phases.

Startle eyeblink reflex

The one-way MANOVA yielded a statistically significant effect of stimulus type on the startle eyeblink reflex, $F(6, 49) = 5.91$, $p < .001$, Wilks's $\Lambda = .580$, partial $\eta^2 = .420$, 90% CI = [.174, .500][15]. During habituation, a statistically significant main effect of stimulus type was observed, $F(2, 108) = 6.33$, $p = .003$, $\varepsilon_{HF} = 0.99$, partial $\eta^2 = .105$, 90% CI = [.024, .192]. Pairwise comparisons indicated that the startle eyeblink reflex magnitude was higher in response to both the CS+, $t(54) = 2.81$, $p = .021$ (Bonferroni corrected), $g_{av} = 0.452$, 95% CI =

[15] As for the postauricular reflex, we ran a two-way repeated measures ANOVA on the startle eyeblink reflex data for the sake of completeness. This analysis yielded statistically significant main effects of stimulus type, $F(2, 108) = 15.63$, $p < .001$, $\varepsilon_{HF} = 1$, partial $\eta^2 = .225$, 90% CI = [.110, .322], and of phase, $F(2, 108) = 63.65$, $p < .001$, $\varepsilon_{HF} = 0.83$, partial $\eta^2 = .541$, 90% CI = [.418, .621]. In contrast, the Stimulus Type × Phase interaction was not statistically significant, $F(4, 216) = 1.41$, $p = .239$, $\varepsilon_{HF} = 0.83$, partial $\eta^2 = .025$, 90% CI = [.000, .059].

[0.125, 0.788], and the CS-, $t(54) = 3.37$, $p = .004$ (Bonferroni corrected), $g_{av} = 0.633$, 95% CI = [0.247, 1.033], than during the ITI, reflecting that it was potentiated by the CSs (see Figure 3.4.3). However, there was no statistical difference in eyeblink reflex magnitude in response to the CS+ relative to the CS-, $t(54) = 0.86$, $p > .99$ (Bonferroni corrected), $g_{av} = 0.162$, 95% CI = [-0.213, 0.540] (see Figure 3.4.3).

Analysis of the acquisition phase showed a statistically significant main effect of stimulus type, $F(2, 108) = 8.94$, $p < .001$, $\varepsilon_{HF} = 1$, partial $\eta^2 = .142$, 90% CI = [.047, .234]. The eyeblink reflex magnitude was, however, not attenuated in response to the CS+ compared with the CS-, $t(54) = 1.79$, $p = .237$ (Bonferroni corrected), $g_{av} = 0.304$, 95% CI = [-0.036, 0.650] (see Figure 3.4.3). Further comparisons revealed that the eyeblink reflex magnitude was greater to both the CS+, $t(54) = 2.47$, $p = .050$ (Bonferroni corrected), $g_{av} = 0.526$, 95% CI = [0.097, 0.966], and the CS-, $t(54) = 4.02$, $p < .001$ (Bonferroni corrected), $g_{av} = 0.842$, 95% CI = [0.404, 1.297] than during the ITI (see Figure 3.4.3).

In the extinction phase, a main effect of stimulus type was found, $F(2, 108) = 4.05$, $p = .020$, $\varepsilon_{HF} = 1$, partial $\eta^2 = .070$, 90% CI = [.006, .147]. Follow-up comparisons showed that the CS- elicited a higher eyeblink reflex magnitude compared with the ITI, $t(54) = 2.64$, $p = .033$ (Bonferroni corrected), $g_{av} = 0.467$, 95% CI = [0.109, 0.834], whereas the eyeblink reflex magnitude to the CS+ was only marginally higher than during the ITI, $t(54) = 2.34$, $p = .068$ (Bonferroni corrected), $g_{av} = 0.442$, 95% CI = [0.062, 0.830] (see Figure 3.4.3). In addition, the eyeblink reflex magnitudes to the CS+ and to the CS- did not statistically differ, $t(54) = 0.04$, $p > .99$ (Bonferroni corrected), $g_{av} = 0.007$, 95% CI = [-0.342, 0.357] (see Figure 3.4.3).

Skin conductance response

No preexistent difference was found in SCRs to the CS+ ($M = 0.07$, $SD = 0.11$) relative to the CS- ($M = 0.06$, $SD = 0.09$) during habituation, $t(50) = 0.71$, $p = .479$, $g_{av} = 0.097$, 95% CI = [-0.173, 0.369]. Similarly, SCRs to the CS+ ($M = 0.03$, $SD = 0.05$) were not larger than to the CS- ($M = 0.02$, $SD = 0.04$) during the acquisition phase, $t(50) = 0.88$, $p = .381$, $g_{av} = 0.113$, 95% CI = [-0.141, 0.369]. Analysis of the extinction phase likewise showed no statistical difference in SCRs to the CS+ ($M = 0.03$, $SD = 0.05$) compared with the CS- ($M = 0.03$, $SD = 0.05$), $t(50) = -0.52$, $p = .606$, $g_{av} = -0.073$, 95% CI = [-0.352, 0.206][16].

[16] A two-way repeated measures ANOVA on the SCR data revealed a statistically significant main effect of phase, $F(2, 100) = 8.81$, $p = .002$, $\varepsilon_{HF} = 0.70$, partial $\eta^2 = .150$, 90% CI = [.038, .270], reflecting a decrease in SCR magnitude from the habituation phase to the other conditioning phases. By contrast, the main effect of stimulus type was not statistically significant, $F(1, 50) = 0.41$, $p = .525$, $\varepsilon_{HF} = 1$, partial $\eta^2 = .008$, 90% CI = [.000, .090],

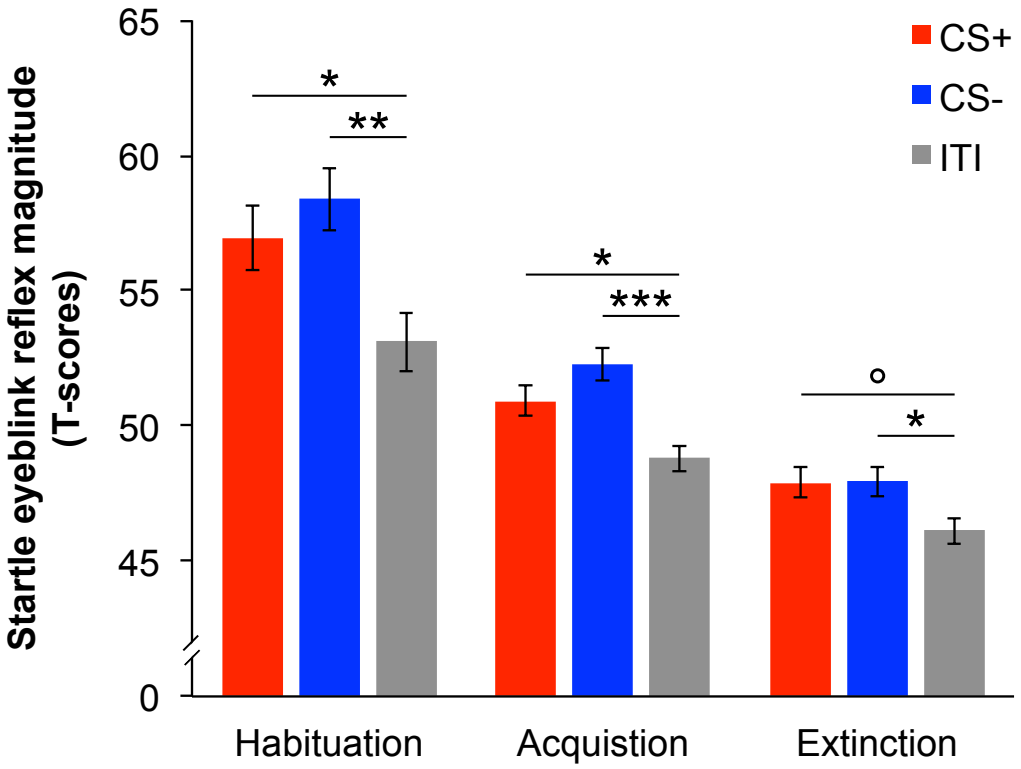

Figure 3.4.3. Mean startle eyeblink reflex magnitudes as a function of stimulus type (CS+ vs. CS- vs. ITI) across the habituation, acquisition, and extinction phases. Error bars represent ± 1 standard error of the mean. ***p < .001, **p < .01, *p < .05, °p < .10 (Bonferroni corrected).

Correlational analysis

The exploratory correlational analysis did not show that participants' subjective hunger level was associated either with the PAR magnitude to the CS+ during acquisition, $r(53)$ = .190, p = .165, 95% CI [-.079, .433] or with the CS+/CS- discrimination as measured by the PAR (i.e., PAR magnitude to the CS+ minus PAR magnitude to the CS-), $r(53)$ = .113, p = .412, 95% CI [-.157, .367].

and no Stimulus Type × Phase interaction effect was observed, $F(2, 100) = 0.56$, $p = .511$, $\varepsilon_{HF} = 0.70$, partial η^2 = .011, 90% CI = [.000, .073].

Subjective ratings

Ratings of CS-US contingency revealed that the CS+ was rated as being more predictive of the olfactory US than the CS-, $t(54) = 4.78$, $p < .001$, $g_{av} = 0.944$, 95% CI = [0.522, 1.386] (see Figure 3.4.4A). In addition, ratings of CS liking showed that the CS+ was evaluated as more pleasant than the CS- after the extinction phase, $t(54) = 2.77$, $p = .008$, $g_{av} = 0.584$, 95% CI = [0.155, 1.024] (see Figure 3.4.4B).

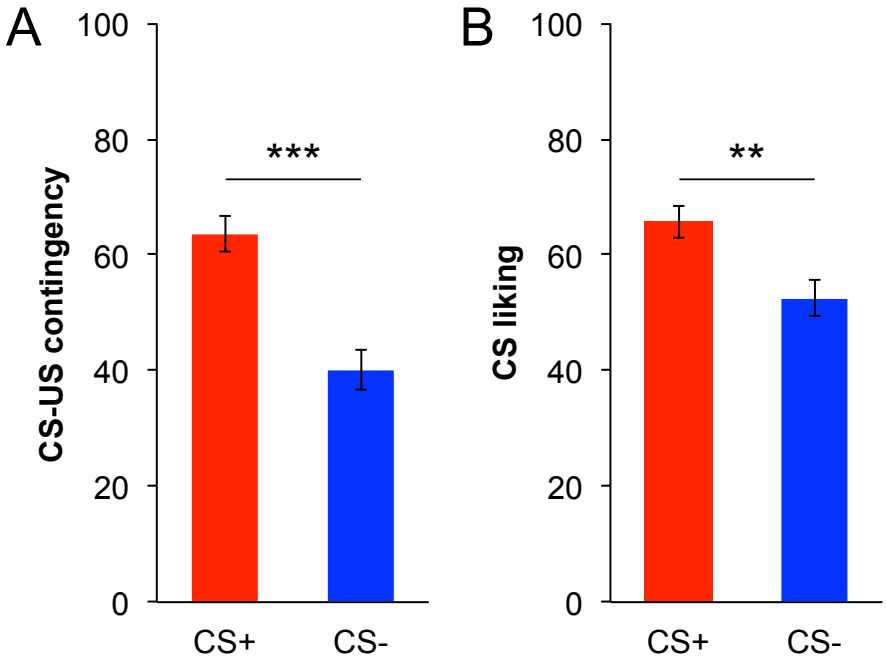

Figure 3.4.4. Mean (A) CS-US contingency ratings and (B) CS liking ratings after the appetitive conditioning procedure as a function of stimulus type (CS+ vs. CS-). Error bars represent ± 1 standard error of the mean. ***$p < .001$, **$p < .01$.

3.4.4. Discussion

In this study, we aimed to assess whether appetitive conditioning may be measured with the postauricular reflex in humans. We used a delay differential appetitive conditioning paradigm, in which a neutral stimulus (CS+) was systematically paired with a pleasant odor, while another neutral stimulus (CS-) was not paired with any odor. We predicted that the postauricular reflex magnitude would be greater to the CS+ compared with the CS- during the acquisition phase. Taken together, our study provides initial evidence that the postauricular reflex can be used to index appetitive conditioning in humans.

First, subjective ratings show that we successfully induced appetitive conditioning in our participants. Overall, the CS+ was deemed more likely to be associated with the pleasant olfactory US than the CS-, indicating that participants were well aware of the contingencies between the CSs and the US. Moreover, the CS+ was evaluated as being more pleasant than the CS- after extinction. These evaluative effects highlight that appetitive conditioning had an impact on the CSs' subjective valence, and therefore demonstrate that the paradigm that we used was efficient in triggering appetitive conditioning.

Most importantly, our results indicate that the postauricular reflex constitutes a sensitive indicator of human appetitive conditioning. The postauricular reflex was indeed specifically potentiated in response to the CS+ compared with the CS- during acquisition, thereby reflecting appetitive learning at the psychophysiological level. This effect is consistent with prior findings that showed a greater postauricular reflex magnitude during presentation of pleasant/appetitive stimuli relative to neutral or unpleasant/aversive stimuli (Aaron & Benning, 2016; Benning, 2011; Benning et al., 2004; Dichter et al., 2010; Gable & Harmon-Jones, 2009; Hackley et al., 2009; Hess et al., 2007; Johnson et al., 2012; Sandt et al., 2009), and does not seem to have been related to participants' subjective hunger level. During the extinction phase, the postauricular reflex magnitude was no longer potentiated to the CS+ in comparison with the CS-, which suggests that its potentiation to the CS+ was conditioned to the pleasant odor delivery.

It is important to note that we were, however, not able to assess whether acquisition and extinction of the postauricular reflex potentiation to the CS+ occurred straight at the outset of the acquisition and extinction phase, respectively, or more gradually. Because we analyzed the postauricular reflex data using signal averaging due to its low signal-to-noise ratio and did not probe every trial, a trial-by-trial analysis of the postauricular reflex modulation was neither possible nor warranted. Nonetheless, these results jointly suggest that (a) the postauricular reflex was sensitive to the contingency between the CS+ and the olfactory US, and (b) the postauricular reflex magnitude modulation and the evaluative effects of appetitive conditioning potentially dissociated. This latter interpretation should nevertheless be considered with caution. As we did not measure ratings trial-by-trial, it is indeed possible that participants rated the conditioned stimuli according to their memories related to the acquisition phase, which might thus not reflect the actual pleasantness of the conditioned stimuli during or after extinction. However, since the CS+ was evaluated as more pleasant than the CS- after extinction, whereas the postauricular reflex potentiation to the CS+ extinguished when the

pleasant odor was no longer delivered, our findings therefore do not provide evidence for the view that affective postauricular reflex modulation merely reflects the stimulus' subjective pleasantness per se (Gable & Harmon-Jones, 2009; Hebert et al., 2015). On the other hand, rather they suggest that the postauricular reflex indexes the predictive or current reward value of the stimulus at stake, which is likely to reflect the interplay of several components, without being limited to positive valence (see, e.g., Berridge & Robinson, 2003). In this respect, our study aligns with previous research suggesting that the postauricular reflex provides a valid psychophysiological indicator of motivational appetitive processes (Aaron & Benning, 2016; Benning, 2011; Benning et al., 2004; Hackley et al., 2009; Sandt et al., 2009).

As rewarding stimuli are typically arousing, it could be alternatively argued that the specific postauricular reflex potentiation to the CS+ relative to the CS- resulted from the CS+ being more arousing than the CS- during acquisition, and that the CS+ arousal value was conversely no longer higher than the CS- during extinction. Although we cannot completely rule out this possibility, we do not think that the postauricular reflex was sensitive to the arousal dimension of the reward-related stimulus. Such an account of our data would indeed be inconsistent with previous findings in the postauricular reflex literature. Specifically, it has been reported that the stimulus arousal level does not appear to modulate the postauricular reflex in response to pleasant or unpleasant stimuli (Gable & Harmon-Jones, 2009). Appetitive-related stimuli have also been shown to evoke a greater postauricular reflex potentiation than nonappetitive pleasant stimuli, although both were reported as similarly arousing (Sandt et al., 2009). In addition, the fact that we observed no modulation of SCR, a prototypical measure of physiological arousal (e.g., Critchley, Elliott, Mathias, & Dolan, 2000), during the acquisition phase likewise does not align with the assumption that the postauricular reflex was modulated by arousal effects.

It should be noted that the greater postauricular reflex magnitude in response to the CS+ relative to the CS- could be conceptualized as a disinhibition of the postauricular reflex rather than a potentiation per se. This conceptualization seems to be consistent with the fact that the postauricular reflex magnitude was smaller in response to the conditioned stimuli than during the ITI in the habituation phase, whereas the postauricular reflex magnitudes to the CS+ and during the ITI were both greater than to the CS-, but did not statistically differ, in the acquisition phase. Putative neurophysiological processes responsible for this modulation pattern might involve a disinhibitory influence of appetitive stimuli within the postauricular reflex neural pathway that counteracts the reduced excitability of the neurons induced by perceptual

engagement with a visual stimulus (see Hackley et al., 1987; Hackley, Ren, Underwood, & Valle-Inclán, 2017). The postauricular reflex neural circuitry is thought to comprise a disynaptic pathway from the cochlear root nucleus to the medial subdivision of the facial motor nucleus that, in turn, activates the postauricular muscle (Hackley, 2015). Based on animal work on the pinna reflex (Li & Frost, 1996), the analog of the human postauricular reflex, it could be speculated that this disinhibitory influence is underlain by inputs from midbrain dopaminergic structures associated with reward processing (e.g., retrorubral nucleus; Waraczynski & Perkins, 2000) to the motoneurons of the facial nerve innervating the pinna (see Benning et al., 2004). However, further research is definitely needed to better understand the neurophysiological mechanisms of the postauricular reflex and elucidate whether its modulation to appetitive stimuli is best conceptualized as a potentiation or as a disinhibition.

With regard to the other psychophysiological measures collected, we found no evidence for startle attenuation in response to the CS+ relative to the CS- during the acquisition phase, and no effect of appetitive conditioning was observed on SCR. These results fail to replicate Andreatta and Pauli's (2015) study, which evidenced both startle attenuation and enhanced SCRs to the CS+ associated with the appetitive US relative to the CS-. However, this inconsistency might arise from several methodological disparities between this study and ours, including in particular the paradigm used (concurrent differential aversive and appetitive conditioning vs. differential appetitive conditioning only), as well as the conditioning procedure used during acquisition (compound conditioning vs. single-element conditioning). Another potential explanation relates to the use of a pleasant odor as appetitive US instead of food. Although both odors and food are primary rewards (Gottfried, 2011), odors constitute a generally less potent class of stimuli than food in humans. Consequently, appetitive olfactory conditioning might lead to smaller effects than appetitive food conditioning (see Rescorla & Wagner, 1972). In line with this proposition, Hermann et al. (2000) were unsuccessful in showing differential appetitive conditioning effects on startle eyeblink magnitude and SCR using a pleasant vanilla odor as US, which contrasts with Andreatta and Pauli's results using an appetitive food US.

Furthermore, other aspects can be equally advanced to account for the lack of statistically significant appetitive conditioning effects on the startle eyeblink reflex and SCR in our study: The startle response, as an aversive and defensive reflex (Lang et al., 1990), has been reported to be an unreliable indicator of appetitive processing in humans (Dichter et al., 2010; Dillon & LaBar, 2005; D. C. Jackson et al., 2000; for a review, see Grillon & Baas,

2003), while SCR, as a biomarker of autonomic arousal (Critchley et al., 2000), may be particularly sensitive to the US intensity, thereby possibly failing to consistently detect subtle changes caused by appetitive conditioning. Of note, the postauricular reflex has also been shown to be resistant to habituation (Hackley et al., 2017), which contrasts with the startle eyeblink reflex (e.g., Bradley, Lang, & Cuthbert, 1993; Grillon & Baas, 2003; Hackley et al., 2017; Rimpel, Geyer, & Hopf, 1982) and SCR (e.g., Bradley et al., 1993; Hare, Wood, Britain, & Shadman, 1970) that are both sensitive to habituation, and is thus less affected by repetitive stimulus presentations, as is the case in human conditioning paradigms. In sum, the fact that we observed differential appetitive conditioning at the psychophysiological level with the postauricular reflex suggests that it provides a sensitive psychophysiological measure of human appetitive conditioning, probably even more sensitive than both the startle eyeblink reflex and SCR.

Interestingly, whereas the postauricular reflex was inhibited by the presentation of the conditioned stimuli relative to the ITI (see also Benning, 2011; Benning et al., 2004; Hackley et al., 1987), the opposite pattern of results was obtained for the startle eyeblink reflex, which was generally potentiated in response to the conditioned stimuli compared with the intertrial interval. This modulation pattern seems to align with previous reports in the human conditioning literature showing an overall greater startle eyeblink reflex magnitude to the CS- than during the ITI (e.g., Andreatta & Pauli, 2015; Hamm, Greenwald, Bradley, & Lang, 1993). Given that startle modulation is affected by multiple processes (Bradley, Codispoti, & Lang, 2006), it might possibly reflect the influence of attentional processes facilitating the enhancement of the acoustic eyeblink reflex during long lead intervals (e.g., when the interval between the stimulus onset and the startle probe is longer than 3 s), typically resulting in larger eyeblink reflex magnitude than during the ITI (e.g., Lipp, Blumenthal, & Adam, 2001), or, alternatively, the impact of specific stimulus characteristics, such as perceptual complexity (see Stanley & Knight, 2004). However, such eyeblink reflex modulation pattern has not been consistently reported across human conditioning studies, some of which observe no enhanced startle eyeblink magnitude to the CS- relative to that during the ITI for instance (see, e.g., Hamm & Vaitl, 1996; Lipp, Sheridan, & Siddle, 1994). This stresses that further investigation is required to better outline the determinants and the robustness of the eyeblink reflex modulation in response to (visual) conditioned stimuli versus during the intertrial interval.

More generally, a caveat pertains to the number of trials included in each conditioning phase. In line with the current standards in the human conditioning literature (see, e.g.,

Lonsdorf et al., 2017), our study was specifically designed to assess changes between the different stimulus types used within each conditioning phase rather than between these phases. Therefore, we implemented a standard differential conditioning paradigm comprising fewer trials for each stimulus type in the habituation phase than in the acquisition and extinction phases, as is typically done in human conditioning paradigms (see, e.g., Andreatta & Pauli, 2015; Olsson, Ebert, Banaji, & Phelps, 2005). However, such differences in trial counts (and hence signal-to-noise ratios) may turn out to be somewhat problematic if one is interested in specifically testing whether the differences between the stimulus types are statistically different between the different conditioning phases (i.e., testing the interaction term). This issue especially holds for the postauricular reflex due to its relatively low signal-to-noise ratio. The postauricular reflex magnitude is likely to be considerably affected by the number of aggregated trials when only few of them are eventually included per condition. In fact, the minimal amount of trials required for obtaining a reliable, stable measure of the postauricular reflex remains to be determined (but see Tooley, Carmel, Chapman, & Grimshaw, 2017, for a recent study suggesting that including at least 12 trials per condition seems to produce a robust estimate of the postauricular reflex magnitude). Those differences in trial numbers between phases (or conditions) may thus complicate the interpretation of the interaction effect, and even potentially produce statistically significant but spurious postauricular reflex magnitude differences. Consequently, future research aiming to specifically assess changes in psychophysiological responses to various stimulus types (e.g., CS+ vs. CS-) between the different conditioning phases should test and explicitly report such interaction term (or, alternatively, a planned contrast analysis, see Rosenthal & Rosnow, 1985), while ideally keeping the number of trials equal within each phase.

In conclusion, the present study suggests that the postauricular reflex arguably represents one of the most suitable psychophysiological indices for measuring appetitive conditioning in humans. In particular, the postauricular reflex sensitivity to appetitive contingencies indicates that this reflex is modulated by the stimulus' reward value, which supports its suitability as a measure of Pavlovian appetitive conditioning. These findings highlight that the postauricular reflex represents a promising psychophysiological indicator for studying Pavlovian reward learning, and more generally reward processing, in humans. Accordingly, future research should notably tackle in more detail whether the postauricular reflex provides a specific index for assessing – and potentially dissociating under particular circumstances – the distinct reward components of wanting, liking, and reward learning (see

Berridge & Robinson, 2003; Pool, Sennwald, et al., 2016). Importantly, this research should however employ an appropriate concept operationalization of the reward components, and ideally take into account potential confounds (e.g., expected pleasantness; see Pool, Sennwald, et al., 2016), along with the stimulus' affective relevance for the organism's concerns (see Pool, Brosch, Delplanque, & Sander, 2016; Pool, Sennwald, et al., 2016). In this perspective, the postauricular reflex constitutes a valuable tool for further shedding light on the basic mechanisms underlying appetitive conditioning and reward processing in humans, as well as their dysfunctions in specific disorders, such as depression, addiction, and food-related disorders.

3.4.5. Supplementary materials

Supplementary results

For the sake of completeness and in order to further assess changes between the different stimulus types used in the experiment (CS+ vs. CS- vs. ITI) across the different conditioning phases (Habituation vs. Acquisition vs. Extinction) with more powerful statistical analyses than the omnibus interaction, we performed planned contrast analyses (see Rosenthal & Rosnow, 1985) specifically testing the two alternative patterns of results that could be expected for the postauricular reflex modulation. The first one predicts the extinction of the difference between the CS+ and the CS- during the extinction phase (see Figure S3.4.1A), and the second predicts the resistance to extinction of the difference between the CS+ and the CS- (see Figure S3.4.1B). These contrasts were both statistically significant, $F(1, 54) = 13.02$, $p < .001$, partial $\eta^2 = .194$, 90% CI = [.058, .337] (see Figure S3.4.1A), and $F(1, 54) = 8.78$, $p = .005$, partial $\eta^2 = .140$, 90% CI = [.027, .279] (see Figure S3.4.1B), respectively, thus supporting the hypothesis that the postauricular reflex was modulated by appetitive conditioning.

Further, we computed a likelihood ratio to examine the relative evidence provided by the postauricular reflex data in favor of one contrast over the other. More specifically, we compared the unexplained variation of the contrast predicting the extinction of the difference between the CS+ and the CS- during the extinction phase (contrast 1; see Figure S3.4.1A) with the unexplained variation of the contrast predicting the resistance to extinction of the difference between the CS+ and the CS- (contrast 2; see Figure S3.4.1B) (see Bortolussi & Dixon, 2003; Glover & Dixon, 2004, for a detailed description of the procedure used to compute likelihood ratios for evaluating competing hypotheses). Given that we used a pure repeated measures design, we calculated the unexplained sum of squares uniquely for the within-participant effects (see Bortolussi & Dixon, 2003). To this end, we subtracted the explained sum of squares associated with contrast 1 from the total within-participant sum of squares, and did the same separately for the explained sum of squares associated with contrast 2 (see Table S3.4.1). Finally, we computed the likelihood ratio using the following formula (see Appendix in Bortolussi & Dixon, 2003):

$$\lambda = \left(\frac{SS_{w,2}}{SS_{w,1}} \right)^{\frac{n(c-1)}{2}}$$

where $SS_{w,1}$ is the unexplained within-participant sum of squares for contrast 1, $SS_{w,2}$ is the unexplained within-participant sum of squares for contrast 2, n is the sample size (here, n = 55), and c is the number of repeated measures (here, c = 9).

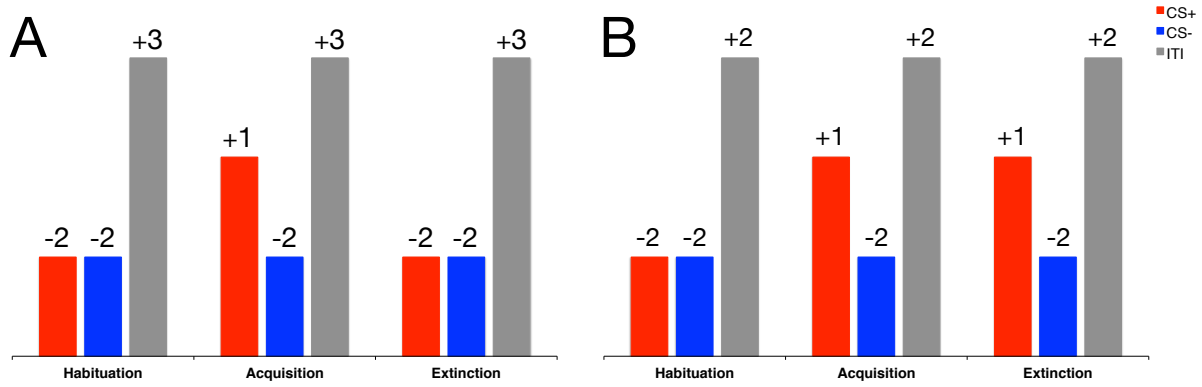

Figure S3.4.1. Alternative expected patterns of results for the postauricular reflex across stimulus types (CS+ vs. CS- vs. ITI) and conditioning phases (Habituation vs. Acquisition vs. Extinction) as tested with planned contrast analyses. (A) Expected pattern of results predicting the extinction of the CS+ vs. CS- difference during extinction. (B) Expected pattern of results predicting the resistance to extinction of the CS+ vs. CS- difference during extinction. Numbers represent the contrast weights for each condition, respectively.

This analysis indicated that the postauricular reflex data observed in the present experiment are 9.61 times more likely given contrast 1 (see Figure S3.4.1A) than given contrast 2 (see Figure S3.4.1B), thereby providing moderate evidence (see, e.g., Royall, 1997) in favor of the view that the postauricular reflex potentiation to the CS+ compared with the CS- extinguished during the extinction phase rather than resisted to extinction.

Table S3.4.1

Results for the 3 (Stimulus Type: CS+ vs. CS- vs. ITI) × 3 (Phase: Habituation vs. Acquisition vs. Extinction) repeated-measures ANOVA and the two alternative planned contrasts on the postauricular reflex data (N = 55).

Source	Degrees of freedom	Sums of squares	Mean Square	*F*
Between-participant				
Error(Participants)	54	230077.91	4260.70	
Within-participant				
Stimulus Type	2	2072.67	1036.33	11.67
Error(Stimulus Type)	108	9587.74	88.78	
Phase	2	957.42	478.71	2.97
Error(Phase)	108	17391.65	161.03	
Stimulus Type × Phase	4	317.20	79.30	1.33
Error(Stimulus Type × Phase)	216	12904.55	59.74	
Total	440	43231.22		
Contrasts				
Contrast 1	1	1707.71	1707.71	13.02
Error(Contrast 1)	54	7080.08	131.11	
Contrast 2	1	1278.42	1278.42	8.78
Error(Contrast 2)	54	7866.98	145.68	

4. GENERAL DISCUSSION & CONCLUSION

4.1. SYNTHESIS AND INTEGRATION OF THE MAIN FINDINGS

In the present thesis, we sought to establish whether relevance detection is a key determinant of preferential emotional learning in humans. More specifically, we tested the theoretical prediction deriving from appraisal theories of emotion (e.g., Sander et al., 2003, 2005, 2018) asserting that stimuli detected as relevant to the organism's concerns are preferentially learned during Pavlovian conditioning, independently of their valence and evolutionary status per se (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018). To this end, we performed a series of empirical studies that aimed to (a) examine whether, like negative threat-relevant stimuli, positive stimuli that are affectively relevant to the organism are preferentially associated with an aversive outcome during Pavlovian aversive conditioning (Studies 1 and 2), (b) characterize at the computational level the influence of the stimulus' affective relevance on Pavlovian aversive conditioning (Studies 1 and 2), (c) assess whether stimuli detected as relevant to the organism's concerns beyond biological and evolutionary considerations can also produce preferential Pavlovian aversive conditioning (Study 3), (d) investigate the impact of inter-individual differences in the organism's concerns on preferential Pavlovian aversive conditioning (Studies 2 and 3), and (e) test and validate the postauricular reflex as a valid psychophysiological indicator of Pavlovian appetitive conditioning in humans in order to provide a sensitive psychophysiological measure that could be further used to probe the generality of a relevance detection mechanism in appetitive learning beyond aversive learning (Study 4).

Across three experiments, Study 1 demonstrated that, similar to threat-relevant stimuli (angry faces and snakes), positive stimuli with biological relevance (baby faces and erotic stimuli) are likewise preferentially associated with an aversive event (electric stimulation) during Pavlovian aversive conditioning. These preferential associations (or learning biases) were characterized by an enhanced resistance to extinction of the conditioned response to both threat-relevant and positive biologically relevant stimuli compared with the conditioned response to neutral stimuli.

Study 2 replicated and extended these findings by showing that both threat-related (angry faces) and positive (happy faces) social stimuli can produce learning biases during Pavlovian aversive conditioning. Similar to Study 1, the conditioned response to angry and happy faces was more resistant to extinction than the conditioned response to neutral faces. Angry and happy faces furthermore induced a faster acquisition of a conditioned response than

neutral faces, thereby reflecting that positive affectively relevant stimuli can also be readily associated with an aversive outcome, as is the case for threat-relevant stimuli. The observation of facilitated Pavlovian aversive conditioning to both negative threat-related and positive stimuli with affective relevance in Study 2 but not in Study 1 might relate to the use of a larger sample (N = 107 in Study 2 vs. N = 40-60 in Study 1), which likely entailed a higher power to detect this effect. Results from Studies 1 and 2 in fact suggest that the effects of faster Pavlovian aversive conditioning in response to negative and positive affectively relevant stimuli appear to be of relatively smaller magnitude than those of enhanced resistance to extinction, being hence probably harder to detect. In line with this suggestion, the human conditioning literature has generally shown a lack of experimental support for the occurrence of facilitated Pavlovian aversive conditioning to specific stimulus categories, such as threat-relevant stimuli (see McNally, 1987; Öhman & Mineka, 2001, for reviews). For instance, although the effects of enhanced resistance to extinction to threat-relevant stimuli have been frequently reported in response to threat-relevant stimuli in the past (e.g., Mallan et al., 2013; Öhman & Mineka, 2001; but see Åhs et al., 2018), evidence for faster Pavlovian aversive conditioning to threat-relevant stimuli remains scarce by comparison (Atlas & Phelps, 2018; Ho & Lipp, 2014; Öhman et al., 1975). This asymmetry has been argued to arise from methodological factors and, in particular, from the use of high reinforcement rates whereby the CSs+ reliably predict the unconditioned stimulus (Ho & Lipp, 2014). High reinforcement rates can notably induce a rapid acquisition of a conditioned response to all the stimulus categories used within a few pairings, which might mitigate the emergence of differences in the acquisition readiness of the conditioned response across the different stimulus categories (Ho & Lipp, 2014; Lissek et al., 2006), and consequently lead to relatively modest effects overall.

Importantly, while the effects of faster conditioning to angry and happy faces were both of moderate size in Study 2, the effect of enhanced resistance to extinction to happy faces was smaller than that to angry faces. This effect was further modulated by inter-individual differences in happy faces' affective evaluation, as indicated by a greater persistence of the conditioned response to happy faces in individuals who were faster in associating them with the attribute of importance versus unimportance in a separate Go/No-go Association Task. Conversely, we did not find evidence that the persistence of the conditioned response to angry and neutral faces was influenced by inter-individual differences in their affective evaluation. These results are consistent with the notion that angry faces are likely to be consistently appraised as highly relevant across individuals due to their high relevance for the organism's

survival and well-being, whereas the appraised relevance of happy faces is more likely to vary among individuals as a function of inter-individual differences, thereby resulting in happy faces holding a lower level of relevance to the organism than angry faces at the group level (Brosch et al., 2008, 2010; Pool, Brosch, et al., 2016; see also Sander et al., 2005). In this respect, Study 2 suggests that the way stimuli are eventually evaluated by the individual can modulate the occurrence of learning biases in response to them, thus delineating the central role of inter-individual differences in preferential Pavlovian aversive learning (see also Lonsdorf & Merz, 2017).

In agreement with the relevance detection model of emotional learning, Studies 1 and 2 therefore critically show that both negative and positive affectively relevant stimuli are more readily and more persistently associated with an aversive event during Pavlovian aversive conditioning than neutral stimuli with less relevance. These results reflect that preferential Pavlovian aversive learning is not restricted to threat-relevant stimuli, but extends to positive stimuli that are affectively relevant to the organism. Albeit somewhat counterintuitive, our findings thereby lend support to the hypothesis that stimuli that are affectively relevant to the organism are preferentially learned during Pavlovian aversive conditioning irrespective of their valence. Whereas it could still be argued that these findings were (partly) mediated by other factors, such as arousal or salience, we contend that relevance detection provides a more parsimonious and appropriate mechanistic explanation thereof, as we will discuss in more detail later.

Additional computational analyses performed in Studies 1 and 2 using simple reinforcement learning models (Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) suggested that the influence of both negative and positive affectively relevant stimuli, as opposed to neutral, less relevant stimuli, could be characterized by a lower learning rate for negative prediction errors. In Study 1, both angry and baby faces were associated with a lower estimated learning rate for negative prediction errors than neutral faces[17]. In Study 2, the estimated learning rate for negative prediction errors to angry faces was lower than that to happy and neutral faces, and the learning rate for negative prediction errors to happy faces was lower than that to neutral faces, although the latter difference only reached marginal

[17] Note that we did not observe lower estimated learning rates for negative prediction errors to snake and erotic images relative to neutral colored squares in Experiment 3 of Study 1, possibly due to noisier skin conductance response data, as suggested by a reduced fit to the data in this experiment (for a more detailed discussion, see chapter 3.1.5, section “Pavlovian learning models”).

significance after correcting for multiple comparisons. Given that learning rates ultimately affect the impact of prediction errors on Pavlovian learning, these findings suggest that affectively relevant stimuli may bias inhibitory learning underpinning extinction (Dunsmoor, Niv, et al., 2015) through a diminished impact of negative prediction errors, thereby producing an enhanced resistance to extinction of the conditioned response to these stimuli. Of note, we however found no evidence in Study 2 that faster Pavlovian aversive conditioning to angry and happy faces was underlain by higher learning rates for positive prediction errors. Altogether, these results provide initial insights into the computational mechanisms whereby the influence of stimulus' affective relevance on Pavlovian aversive conditioning operates, hence contributing to characterizing the role of relevance detection in emotional learning. As these findings mostly constitute a first attempt at elucidating the impact of stimulus' affective relevance at the computational level, it is nonetheless worth noting that these computational mechanisms remain yet to be better pinpointed and characterized in further research.

In Study 3, we further showed that initially neutral stimuli that acquired goal-relevance for participants were more readily learned during Pavlovian aversive conditioning than goal-irrelevant stimuli in participants high in achievement motivation, but not in participants lower in this trait. These results indicate that stimuli temporarily associated with a higher goal-relevance can produce facilitated Pavlovian aversive conditioning even though they hold no pre-existing threat value, provided that specific individual motivation dispositions are met concurrently. Therefore, results of Study 3 suggest that stimuli that are relevant to the organism beyond biological or evolutionary considerations can induce accelerated Pavlovian aversive learning, thus concurring with the hypothesis that stimuli detected as relevant to the organism's concerns can produce a learning bias independently of their evolutionary status. Moreover, they suggest that the occurrence of such learning bias crucially depends on inter-individual differences in affect and motivation. Correspondingly, Study 3 aligns with Study 2 in highlighting the key importance of inter-individual differences in the organism's concerns in preferential Pavlovian aversive learning, as suggested by the relevance detection model (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018; see also Sander et al., 2005).

On the other hand, we did not find any effect of enhanced resistance to extinction to goal-relevant stimuli compared with goal-irrelevant stimuli in Study 3. This null finding notably deviates from the observations of enhanced resistance to extinction to both negative and positive affectively relevant stimuli reported in Studies 1 and 2, and suggests that the preferential aversive learning to goal-relevant stimuli in individuals with high achievement

motivation was rather modest and transient. It thereby appears that the effects of stimulus' goal-relevance reported in Study 3 were probably smaller and somewhat less robust than the effects of stimulus' affective relevance found in Studies 1 and 2. Given that the goal-relevant stimuli used in Study 3 were temporarily made relevant for task-related goals through an experimental manipulation in laboratory settings, whereas the affectively relevant stimuli used in Studies 1 and 2 were of general relevance for the organism's survival and/or well-being, it is very likely that the former had an overall lower level of relevance to the organism than the latter. This might notably explain the occurrence of seemingly weaker effects of resistance to extinction in Study 3 relative to Studies 1 and 2. As survival and well-being are among the highest prioritized concerns and can thus be conceptually considered as high-value sub-categories of goal-relevance, stimuli that are relevant at this level are likely to be appraised as more relevant than stimuli that are relevant to other types of concerns holding lower priority. Interestingly, this conjecture is consistent with the findings of Study 2 showing that the effects of resistance to extinction to happy faces were smaller than those found for angry faces; happy faces typically having a lower level of relevance than angry faces across individuals (Brosch et al., 2008, 2010; Pool, Brosch, et al., 2016). Together, these considerations align with the view that the modulation of preferential Pavlovian aversive learning is directly related to the stimulus' affective relevance, with stimuli being detected as more relevant inducing stronger learning biases than stimuli associated with a lower level of relevance. In that regard, they additionally suggest that learning biases to affectively relevant stimuli during Pavlovian aversive conditioning occur flexibly and dynamically in a dimensional rather than a dichotomous fashion as a function of the interplay between specific stimulus characteristics and the organism's current concerns.

Last, Study 4 consisted of a methodologically-oriented experimental attempt, in which we implemented a differential Pavlovian appetitive conditioning paradigm to investigate whether the postauricular reflex may provide a valid psychophysiological measure of human Pavlovian appetitive conditioning, while comparing this measure with the startle eyeblink reflex and skin conductance response. Results indicated that the postauricular reflex was potentiated in response to an initially neutral conditioned stimulus (CS+) that was contingently paired with a pleasant odor (appetitive US) in comparison with another neutral conditioned stimulus (CS-) that was never associated with the pleasant odor. This potentiation furthermore extinguished when the pleasant odor was no longer delivered. By contrast, we found no attenuation of the startle eyeblink reflex in response to the CS+ relative to the CS-, and no

effect of Pavlovian appetitive conditioning was observed on skin conductance response. These findings suggest that the postauricular reflex provides a valid measure of Pavlovian appetitive conditioning in humans, which is likely more sensitive than both the startle eyeblink reflex and skin conductance response. In this perspective, the postauricular reflex emerges as a valuable tool for further investigating whether stimuli that are affectively relevant benefit from preferential Pavlovian appetitive conditioning in humans, beyond their valence and evolutionary status per se. As such, the postauricular reflex stands as a promising psychophysiological measure for systematically testing in future studies the generality of relevance detection in appetitive contingencies. In particular, this tool could contribute to assessing at the empirical level the key hypothesis of the relevance detection model predicting that both negative and positive stimuli with enhanced affective relevance would be readily and persistently associated with an appetitive outcome during Pavlovian appetitive conditioning.

Altogether, the empirical studies reported here demonstrate that preferential Pavlovian aversive conditioning is not limited or confined to threat-related stimuli, but more generally extends to positive biologically relevant stimuli and even to initially neutral stimuli that are made relevant to the organism's concerns through experimental manipulation, independently of their intrinsic valence or evolutionary history per se. They further highlight that the occurrence of such learning biases flexibly hinges upon inter-individual differences in affect and motivation. In this respect, our results are congruent with the relevance detection model deriving from appraisal theories of emotion (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018), according to which the emergence of learning biases arises from the appraisal of the stimulus' affective relevance consisting of the interaction between the stimulus at stake and the individual's current concerns. Our data are notably congruent with neurobiological and psychological evidence suggesting that the encoding and processing of negatively and positively valenced stimuli rely on at least partially overlapping and shared brain circuits (e.g., Brosch et al., 2008; Canli et al., 2002; Janak & Tye, 2015; Jin et al., 2015; Namburi et al., 2015; Paton et al., 2006; Sander et al., 2003; Seymour et al., 2007; Shabel & Janak, 2009), neurotransmitter systems (e.g., M. Matsumoto & Hikosaka, 2009), and psychological processes (e.g., Atias et al., 2018; Aviezer, Trope, & Todorov, 2012; Solomon & Corbit, 1974; see also Brosch et al., 2008; Pool, Brosch, et al., 2016).

In contrast, our demonstration of learning biases to affectively relevant stimuli irrespective of their valence and evolutionary status starkly departs from previous research suggesting that enhanced Pavlovian aversive conditioning is selectively restricted to specific

classes of stimuli that have threatened the survival of the species across evolution (e.g., Öhman & Dimberg, 1978; Öhman, Eriksson, et al., 1975; Öhman et al., 1976), as posited by the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories. Whereas the fear module theory acknowledges that this module can also be activated by threat-related ontogenetic stimuli under certain circumstances (e.g., extensive training) in addition to its preferential activation by threat-related phylogenetic stimuli within aversive contexts (Öhman & Mineka, 2001), it is conceptualized as a fear-specific mechanism that is not thought to be differentially activated in response to stimuli that do not differ in their inherent threat value. Accordingly, this theoretical account cannot easily accommodate the observation of learning biases to affectively relevant stimuli occurring irrespective of their valence, as reported here in Studies 1, 2, and 3. In a similar vein, our results are also somewhat inconsistent with the expectancy bias model (Davey, 1992, 1995). Indeed, this model specifically asserts that preferential aversive associations stem from heightened expectancies of aversive outcomes in response to specific (threat-relevant) stimuli that primarily result from the stimulus' appraised dangerousness (Davey, 1995), thus incorporating the assumption that enhanced Pavlovian aversive conditioning is selective to threat-related stimuli. By showing that both negative and positive stimuli with affective relevance to the organism can be preferentially associated with an aversive event during Pavlovian aversive conditioning, the present studies suggest instead that such a priori expectancy biases do not seem necessary to induce preferential Pavlovian aversive learning. Because we found no support for the occurrence of selective sensitization across our experiments, it also appears unlikely that selective sensitization constitutes a necessary mechanism for the emergence of learning biases during Pavlovian aversive conditioning in humans. Rather, the learning biases that we observed in our studies most likely resulted from an associative learning process rather than pre-existing response tendencies (e.g., Lovibond et al., 1993).

As affectively relevant stimuli are typically highly arousing, it could be reasonably argued that this specific emotional dimension could have mediated our findings in addition to – or rather than – their affective relevance. In general, it remains challenging to disentangle the specific contribution of relevance detection from that of arousal to emotional effects on various cognitive processes (see Montagrin & Sander, 2016; Pool, Brosch, et al., 2016; Sander, 2013). According to appraisal theories of emotion (e.g., Sander et al., 2003, 2005, 2018), stimuli that are detected as relevant to the individual's current concerns very often trigger a motivational state, which in turn elicits a physiological state of arousal that might even be felt and

experienced consciously (Pool, Brosch, et al., 2016). At variance with the arousal account positing that preferential emotional learning depends on a specific component of the emotional response, the relevance detection hypothesis however states that the key determinant of preferential emotional learning corresponds to the relevance detection process involved in emotion elicitation, which thereby highlights that these two accounts substantially differ in terms of the hypothesized psychological mechanisms thought to be responsible for the emergence of these learning biases.

Nonetheless, it should be noted that the construct of "arousal" is often ill-defined, loose, and conceptually unclear when considering its use in the affective sciences literature (see, e.g., Mather, Clewett, Sakaki, & Harley, 2016; Montagrin & Sander, 2016; Sander, 2013). For instance, the notion of arousal traditionally refers to either (a) the activation state of the experienced affect that is felt consciously, as typically used in core affect theories of emotion (i.e., felt arousal; e.g., Russell, 2003), or (b) the bodily reaction or physiological state during an emotional episode (i.e., the changes in the sympathetic nervous system referred to as physiological arousal; e.g., Bradley, Miccoli, Escrig, & Lang, 2008; see also Montagrin & Sander, 2016). Frijda (1986, p. 168) furthermore advanced that the concept of arousal or activation can relate to three response systems: autonomic arousal, electrocortical arousal, and behavioral activation. This proposition contrasts with the notion of a general, unique arousal system, and rather suggests that arousal is probably not a unitary process (see also Robbins, 1997). Indeed, although subjective ratings of arousal are often considered as mirroring the activation of the sympathetic nervous system as measured with psychophysiological indicators (e.g., Bradley et al., 2008; Lang et al., 1993), it remains unclear to what extent felt arousal represents a direct, unaltered reflection of physiological arousal, or whether these two types of arousal are usually correlated, but distinct phenomena.

At the empirical level, several lines of evidence argue against felt and/or physiological arousal having a sufficient role to enhance emotional learning. First, a meta-analysis on attentional bias for positive emotional stimuli (Pool, Brosch, et al., 2016) has shown that both arousal and affective relevance influenced the attentional bias magnitude, but that it was only significantly predicted by affective relevance when the contribution of arousal was statistically controlled. Similarly, it has been shown that an induced state of physiological arousal can occur without affecting or enhancing memory processes (e.g., Adolphs, Tranel, & Buchanan, 2005; Christianson & Mjörndal, 1985; Libukman, Nichols-Whitehead, Griffith, & Thomas, 1999). By drawing a parallel between emotional attention and emotional memory with emotional

learning, these findings tentatively and indirectly suggest that affective relevance appears more likely to drive learning biases than arousal. Second, previous studies in the human conditioning literature (Hamm et al., 1993; Hamm & Stark, 1993; Hamm & Vaitl, 1996) reported that negative and positive stimuli evaluated as highly arousing, without considering their relevance to the organism's current concerns, did not induce preferential Pavlovian aversive conditioning in comparison with stimuli with a lower level of arousal. Third, analyses of the skin conductance response during the habituation phase across Studies 1 to 3 provided no evidence that affectively relevant stimuli triggered enhanced physiological arousal relative to stimuli with less relevance before conditioning[18], this factor being thus unlikely to have driven the subsequent occurrence of learning biases to these stimuli. Last, the effects of Pavlovian appetitive conditioning reported in Study 4 were specifically captured by the postauricular reflex, which does not appear to be modulated by arousal (Gable & Harmon-Jones, 2009; Sandt et al., 2009), but not by skin conductance response, which is usually seen as a relatively "pure" measure of autonomic arousal (e.g., Cuthbert, Schupp, Bradley, Birmbauer, & Lang, 2000; Lang et al., 1993). Speculatively, this observation suggests that a state of heightened arousal is probably not a strictly necessary condition for Pavlovian conditioning to occur. Combined together, these considerations denote that arousal alone seems insufficient to produce enhanced Pavlovian (aversive) conditioning, and accordingly suggest that relevance detection provides a more plausible and parsimonious mechanism to account for our new findings.

Another possible factor that could have contributed to the occurrence of learning biases to affectively relevant stimuli compared with neutral, less relevant stimuli across our empirical studies is stimulus salience. Indeed, stimuli that are affectively relevant are typically highly salient (e.g., Cunningham & Brosch, 2012). Stimulus salience is also incorporated as a key parameter determining the learning rate to a given conditioned stimulus in a number of formal models of Pavlovian conditioning (e.g., Li et al., 2011; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see chapter 2.2.4). In analogy with arousal, the construct of "salience" is sometimes used as an umbrella term; however, it is not a unitary process and may refer to different types of salience (see chapter 2.2.2, section "Stimulus novelty and intensity",

[18] In Study 1, we observed no statistically significant difference across angry faces, baby faces, and neutral faces, or across snake images, erotic images, and neutral colored squares during habituation. In Study 2, angry faces elicited greater skin conductance responses than happy faces during the habituation phase, but no difference was found between angry and neutral faces, or between happy and neutral faces. In Study 3, no main effect of the stimulus category (goal-relevant valid vs. goal-relevant invalid vs. goal-irrelevant) or interaction effect with participants' standardized (z-score) achievement motivation score were observed at the level of the skin conductance response during habituation (all Fs < 1.19, all ps > .30, all partial η^2s < .017).

and chapter 2.3.3, section "Conditioned stimulus salience"; see also chapter 3.1.4). In the Pavlovian conditioning literature, the notion of salience generally refers to the stimulus' intrinsic physical characteristics (see, e.g., Öhman & Mineka, 2001; Pearce & Hall, 1980; but see Rescorla, 1988a). In this context, stimuli that are more salient or intense in terms of physical or perceptual salience have been shown to be more easily conditioned than less salient stimuli (e.g., Kamin & Schaub, 1963; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla, 1988a; Rescorla & Wagner, 1972); by contrast, it has been reported that neutral stimuli that are highly perceptually salient do not induce a greater resistance to extinction of the conditioned response during Pavlovian aversive conditioning than neutral stimuli with lower perceptual salience (Öhman et al., 1976). Moreover, according to classical models of Pavlovian conditioning (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; see also chapter 2.3), the conditioned response to more salient or intense stimuli should, all else considered equal, extinguish faster than the conditioned response to less salient stimuli (Siddle & Bond, 1988; see also Kamin & Gaioni, 1974; Kremer, 1978; Taylor & Boakes, 2002, for experiments in rats providing either direct or indirect evidence for this prediction). Accordingly, physical or perceptual salience appears as an insufficient and unlikely factor to explain the effects of enhanced resistance to extinction to both threat-relevant and positive relevant stimuli observed in Studies 1 and 2 (see also McNally, 1987; Öhman & Mineka, 2001).

Nevertheless, a broader conceptualization of salience as not confined to the mere properties of the stimulus but also including its relative importance to motivational contingencies pertaining to the organism's needs and goals (see Cunningham & Brosch, 2012; Öhman & Mineka, 2001; Rescorla, 1988a) seems more appropriate to account for the results observed across the experiments reported in this thesis. In fact, stimuli that are relevant to the organism's current concerns could be considered as holding a high motivational or incentive salience. In particular, the process of incentive salience (see, e.g., Berridge, 2007; Berridge & Robinson, 1998, 2016; Schultz, 2015) has been suggested to be conceptually very closely related to the process of relevance detection as implemented in appraisal theories of emotion (see Pool, Sennwald, et al., 2016; Sennwald, Pool, & Sander, 2017). In that sense, a post-hoc explanation of our results according to a motivational salience hypothesis would very closely mirror the a priori predictions of the relevance detection model, these two accounts being virtually equivalent. Indeed, despite the fact that the constructs of "motivational salience" and "relevance detection" have different conceptual historical roots and originate from different research traditions, they both share the fundamental assumption that the key determinant of

preferential emotional learning in humans relies on the interaction between the stimulus at hand and its motivational relevance for the organism's current concerns.

In sum, the set of experiments conducted in this thesis challenges the view that threat-relevant stimuli benefit from enhanced Pavlovian aversive conditioning because they have been associated with threat across evolution. These experiments alternatively suggest that the key determinant underlying preferential Pavlovian aversive conditioning in humans rather corresponds to a process of relevance detection, as opposed to a threat-specific mechanism as previously thought. Our findings therefore provide support for the existence of a general mechanism that is shared across stimuli that are detected as relevant to the organism's concerns enhancing emotional learning. Such mechanism appears particularly functional as it prioritizes the learning of stimuli that are pertinent according to specific individual motivations through accelerated and more persistent learning, thereby helping the organism flexibly shape appropriate responses to these stimuli and ultimately interact with, and adapt to, their environment.

4.2. THEORETICAL IMPLICATIONS

4.2.1. Theoretical models of emotional learning in humans

The empirical findings reported in the present thesis have important implications for the theoretical modeling of the determinants of preferential emotional learning in humans. More particularly, they critically advocate a change of perspective in the conceptualization of the basic mechanisms underlying enhanced Pavlovian aversive learning. In fact, it has long been posited that only stimuli that have threated survival across evolution are preferentially learned during Pavlovian aversive conditioning, thus conferring a privileged status to negative stimuli carrying threat-related information from phylogenetic origin to foster emotional learning (e.g., Öhman & Mineka, 2001; Seligman, 1971). In line with this view, influential theoretical models of emotional learning have proposed that enhanced Pavlovian aversive conditioning is underlain by a biological preparedness process (Seligman, 1970, 1971) and/or an evolved fear module that is preferentially activated by evolutionarily prepared threat stimuli (Öhman & Mineka, 2001). In contradiction with the major tenets of these evolutionary theories, we showed that preferential Pavlovian aversive conditioning is not selective to threat-relevant stimuli from evolutionary origin as previously thought, but instead extends to stimuli that are relevant to the organism's concerns beyond their valence and evolutionary status, and this even though they are devoid of any pre-existing threat value. Our findings thus indicate that the emergence of aversive learning biases is likely driven by a more flexible mechanism than a fear- or a threat-specific mechanism. They further suggest that this mechanism is likely to be shared across negative and positive relevant stimuli, as well as previously neutral stimuli having acquired enhanced affective relevance through experimental manipulation, thereby aligning with the predictions of the relevance detection model of emotional learning deriving from appraisal theories of emotion (see Figure 4.1; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018; see also Sander et al., 2003, 2005, 2018).

In addition, the fact that initially neutral stimuli temporarily associated with higher (goal-)relevance can also benefit from facilitated Pavlovian aversive conditioning when considering specific individual motivation dispositions suggests that learning biases are determined by the interaction between the stimulus at play and the individual's current concerns, rather than by the stimulus' inherent properties, as postulated by appraisal theories (e.g., Frijda, 1986, 1988; Moors et al., 2013; Sander et al., 2005, 2018). These results hence highlight the central influence of the individual's motivational state on enhanced emotional

learning. Specifically, the salience and priority of the organism's concerns may flexibly and rapidly change according to current environmental contingencies and task demands (e.g., Cunningham & Brosch, 2012), which enables the individual to learn preferentially specific stimuli that are detected as relevant to their current concerns at a certain moment in time, thereby yielding high flexibility in the production of learning biases, and this way fostering a high degree of adaptation. Nonetheless, it should be noted that specific stimuli that are relevant to concerns that are shared among individuals, relatively stable, and commonly recognized (i.e., "source concerns"; e.g., survival- and reproduction-related concerns; Frijda, 2009) are more likely to consistently produce learning biases across individuals than other stimuli that are relevant to concerns that are specific to a given individual at a given time and situation (i.e., "surface concerns"; e.g., attachment or attraction to a specific person; Frijda, 2009)[19]. For instance, this might notably apply to threat-relevant stimuli or babies and sexual stimuli because of their enhanced biological relevance to the common source concerns of survival and/or reproduction shared across practically all individuals, which may account for the relatively robust learning biases found in response to these stimuli. In this light, the present thesis emphasizes the key role of the organism's current concerns in the enhancement of Pavlovian aversive conditioning (see Figure 4.1), thus stressing the importance of considering the organism's motivational state and dispositions for promoting a better understanding of the functioning of emotional learning in humans.

Relatedly, the evidence reported in this thesis furthermore delineates the crucial impact of inter-individual differences on the occurrence of learning biases during Pavlovian aversive conditioning. Although it has been well known that inter-individual differences are inherent and highly prevalent in Pavlovian conditioning (see, e.g., Beckers, Krypotos, Boddez, Effting, & Kindt, 2013; Lonsdorf & Merz, 2017; Pavlov, 1927) – which varies considerably according to biological, experiential, or personality factors, along with affective and cognitive biases (e.g., Arnaudova et al., 2013; Byrom & Murphy, 2018; Gazendam et al., 2015; Hartley et al., 2011; Lonsdorf et al., 2009; Sjouwerman et al., 2018; Zorawski et al., 2005; for a review, see Lonsdorf & Merz, 2017) – they have surprisingly mostly been regarded as "noise" or epiphenomenal rather than carrying meaningful information in many previous studies.

[19] Please note, however, that source and surface concerns are not opposite or mutually exclusive categories, as a given stimulus can be relevant to both source and surface concerns simultaneously (e.g., a food stimulus can be relevant to survival-related concerns such as nourishment and to the specific concern of hunger; Rodriguez Mosquera, Fischer, & Manstead, 2004; see also Pool, Brosch, et al., 2016).

Contrasting with this view, our results suggest that inter-individual differences in affect and motivation can dynamically modulate preferential Pavlovian aversive conditioning to specific stimuli depending on their relevance to the individual's current concerns. These findings thereby advocate a careful consideration and modeling of inter-individual differences in emotional learning as a valuable source of variability that can inform us about the processes underpinning emotional learning in humans. Despite some initial attempts made to take into account inter-individual differences in Pavlovian conditioning, their contribution to emotional learning, as well as the basic determinants thereof, remain however poorly understood currently (see Lonsdorf & Merz, 2017).

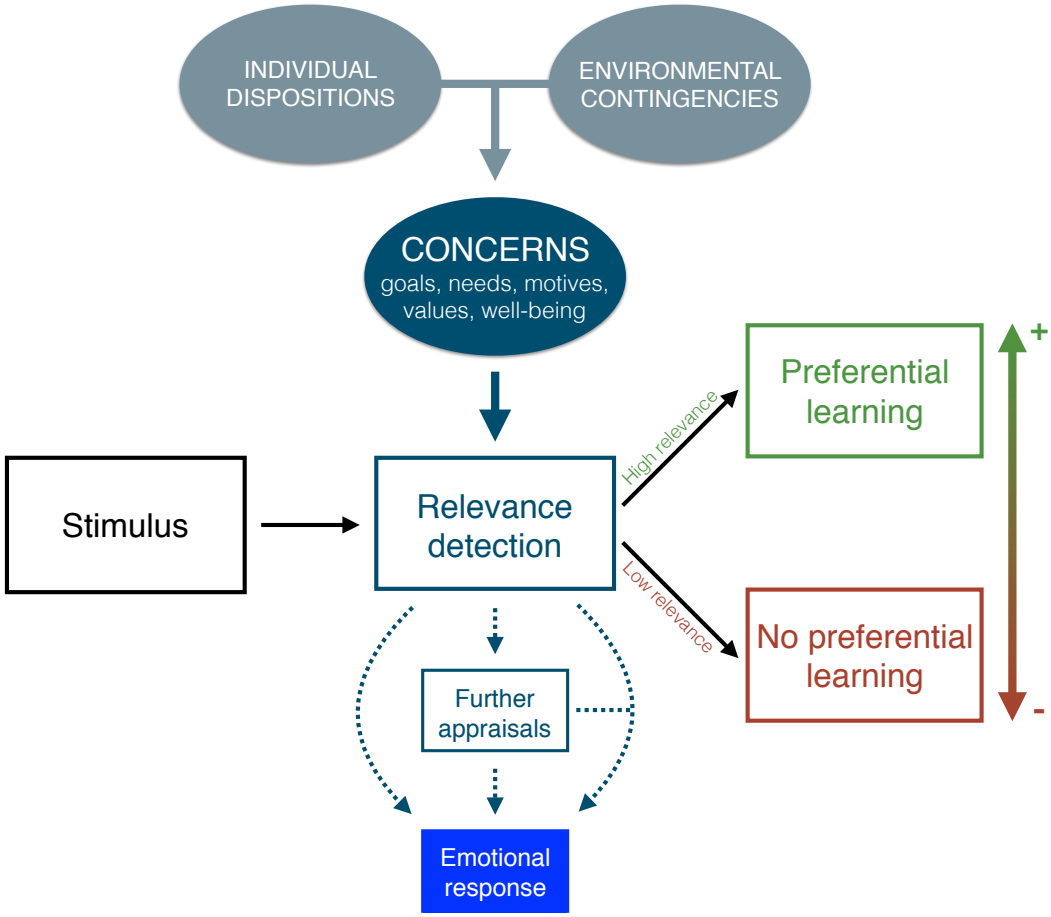

Figure 4.1. Illustration of the relevance detection model of emotional learning deriving from appraisal theories of emotion proposed in this thesis. According to this model, preferential emotional learning is driven by a mechanism of relevance detection relying on the organism's concerns, which depend on individual dispositions and current environmental contingencies. More specifically, the relevance detection model predicts that a stimulus that is detected as relevant to the organism's concerns benefits from preferential learning independently of its valence and evolutionary status per se, and that such preferential learning occurs in a dimensional fashion as a function of the stimulus' level of relevance to the organism.

In this perspective, the relevance detection model put forward in this thesis (see Figure 4.1) might provide a valid theoretical framework to account for the large (inter-)individual differences commonly observed in Pavlovian conditioning across varying situations and contexts. Albeit speculative, it could moreover help model and better understand other forms of emotional learning, besides Pavlovian aversive conditioning. According to this model, the process of relevance detection differs between individuals, or even between different moments in time for the same individual, depending on their current concerns, which are very likely to vary substantially across them (e.g., Cunningham & Brosch, 2012; Frijda, 1986; Sander et al., 2005). For instance, this flexibility notably allows the same stimulus to induce a learning bias for a given individual, but not for another one, as a function of their respective current concerns' hierarchy and the way in which they eventually appraise the stimulus at hand. More generally, appraisal theories suggest that appraisal mechanisms are largely influenced by individual differences (e.g., Sander et al., 2005), which in turn underlies individual variations in preferential emotional learning. The relevance detection framework therefore offers a putative mechanistic account for both inter-and intra-individual differences in emotional learning, which may contribute to shedding new light on the role and determinants thereof in this learning process.

4.2.2. Computational models of Pavlovian conditioning

In Studies 1 and 2, we provided initial insights into the computational mechanisms by which stimulus' affective relevance might operate to modulate Pavlovian aversive conditioning. Specifically, negative and positive stimuli with affective relevance were associated with a lower learning rate for negative prediction errors compared with neutral, less relevant stimuli, which consequently diminished the influence of negative prediction errors on the updating of the conditioned stimulus predictive value. This reduced impact of negative prediction errors likely contributed to weakening inhibitory learning that underlies extinction (e.g., Dunsmoor, Niv, et al., 2015), thereby characterizing the enhanced persistence of the conditioned response to affectively relevant stimuli relative to neutral stimuli with lesser relevance. Accordingly, these results suggest that preferential (i.e., facilitated or faster and more persistent) Pavlovian aversive conditioning in humans might be best captured and characterized by modeling separate learning rates for positive (i.e., excitatory learning) and negative (i.e., inhibitory learning) prediction errors.

Of particular importance, these findings are somewhat inconsistent with the Rescorla-Wagner (Rescorla & Wagner, 1972), the Mackintosh (1975), the Pearce-Hall (Pearce & Hall, 1980; Pearce et al., 1982), and the hybrid (Li et al., 2011) models of Pavlovian conditioning (see chapter 2.2.4). For instance, the Rescorla-Wagner model allows various conditioned stimuli to differentially affect the acquisition and extinction of a conditioned response through different learning rates, which are essentially determined by the conditioned stimulus salience and which ultimately alter the influence of prediction errors on Pavlovian conditioning. However, this model also assumes that such learning rates are identical for both excitatory and inhibitory learning, thus implying that conditioned stimuli associated with a high learning rate (i.e., highly salient or intense conditioned stimuli) are supposed to induce not only a faster conditioned response acquisition, but also an accelerated extinction of the corresponding conditioned response.

In a similar vein, whereas the Mackintosh model accounts for variations in the occurrence of facilitated Pavlovian conditioning by means of varying initial levels of associability affecting the influence of prediction errors, with conditioned stimuli with higher initial associability increasing the weight attributed to prediction errors and consequently accelerating the acquisition of a conditioned response, it does not tease the impact of positive versus negative prediction errors apart. Accordingly, the Mackintosh model postulates that conditioned stimuli that have higher associability (i.e., that reliably predict the unconditioned stimulus) should likewise produce a faster extinction of the conditioned response.

Alternatively, the Pearce-Hall model posits that conditioned stimuli may achieve preferential Pavlovian conditioning through heightened learning rates, which relate to their intensity or intrinsic salience, and/or heightened initial levels of associability. Conditioned stimuli that are more intense and/or initially more associable (i.e., that are unreliable predictors of the unconditioned stimulus) are correspondingly thought to lead to facilitated Pavlovian conditioning as opposed to conditioned stimuli with lower intensity and/or initial associability. Although excitatory and inhibitory learning are clearly distinct in this model, the learning rates for both positive and negative prediction errors are assumed to be determined by the conditioned stimulus intrinsic salience, which entails that conditioned response to more salient conditioned stimuli should also extinguish more rapidly than that to less salient stimuli. Similarly to the Mackintosh model, conditioned stimuli that have higher associability are also supposed to induce faster extinction than less associable conditioned stimuli.

As for the hybrid model, it likewise predicts that conditioned stimuli that are more salient or associable should provoke both faster acquisition and faster extinction of the conditioned response compared with less salient or associable conditioned stimuli, in a manner analogous to the Rescorla-Wagner model and the Pearce-Hall model, respectively.

In this context, the current work may suggest potential avenues for the development of a revised computational model of Pavlovian learning through the incorporation of distinct learning rates for positive and negative prediction errors, which could allow for a more accurate modeling of the occurrence of learning biases to specific stimuli observed during Pavlovian conditioning. Importantly, our data further suggest that these learning rates should not exclusively depend on a salience parameter as implemented in the Rescorla-Wagner, the Pearce-Hall, or the hybrid models, for instance, but also a distinct computational parameter tracking the conditioned stimulus' affective relevance to the organism. Such relevance parameter could notably contribute to increasing the influence of positive prediction errors, hence resulting in facilitated excitatory learning, while diminishing the impact of negative prediction errors, thus entailing weakened inhibitory learning.

As the affective relevance of a given stimulus is established based on the interplay between the stimulus and the organism's current concerns, we suggest that, in accordance with appraisal theories (e.g., Sander et al., 2005, 2018), a relevance parameter would also be dynamically modulated by the current motivational state of the individual, rather than constituting a fixed parameter determined solely by the conditioned stimulus' inherent properties. This suggestion is consistent with the computational model of incentive salience proposed by Zhang and colleagues (Zhang, Berridge, Tindell, Smith, & Aldridge, 2009), according to which the incentive salience of a Pavlovian conditioned stimulus fluctuates in a dynamic fashion as a function of the interaction between the organism's direct experience with the outcome – or the lack thereof – through learning processes and the organism's current physiological state. For instance, this model predicts that the same food cue would be associated with a higher incentive salience value when the organism is in a hunger state than in a satiety state, even though the contingency between the conditioned stimulus and the outcome remains identical in both states. However, the Zhang et al.'s (2009) model does not provide a formal computational account of how transitions between different physiological states of the organism occur to subsequently modulate incentive salience. It has been proposed that the accomplishment of such state transformations essentially require *model-based* computations rather than *model-free* ones (Dayan & Berridge, 2014). In contrast to *model-free*

learning by which the organism learns and updates the stimulus' value based on past reinforcement (i.e., via direct, retrospective experience and prediction errors) without constructing representations of the environment, *model-based learning* relies on an "internal model" of the environment that is used to build and update expectations, predictions, and transformations of the stimulus value as a function of the current or future state of the organism, as well as to represent the probabilities governing state transitions (e.g., Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005; Gläscher, Daw, Dayan, & O'Doherty, 2010). Although Pavlovian conditioning has been considered to mostly depend on model-free computations (e.g., Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Sutton & Barto, 1998), it has been suggested that model-based computations also play a key role in this form of learning (Courville, Daw, & Touretzky, 2006; O'Doherty et al., 2017; Pauli, Gentile, Collette, Tyszka, & O'Doherty, 2019; Prévost et al., 2013; see also Gershman et al., 2010; Gershman & Niv, 2012). Based on these considerations, we therefore propose that computational modeling of relevance detection may involve model-based computations relying on the organism's current (or future) motivational state, which would thereby allow the conditioned stimulus' affective relevance to flexibly vary as a function of these motivational states. It is however important to note that an algorithmic implementation of such Pavlovian model-based computations remains yet to be developed and assessed systematically (see, e.g., Dayan & Berridge, 2014).

4.2.3. Theories of emotion and emotional modulation of cognitive processes

Main families of emotion theories (see Figure 4.2) – namely basic emotion theories (e.g., Ekman, 1972, 1992, 1999; Izard, 1992; Panksepp, 1998), dimensional or core affect theories (e.g., Barrett, 2006; Lindquist, Wager, Kober, & Barrett, 2012; Russell, 2003; Russell & Barrett, 1999), and appraisal theories (e.g., Moors et al., 2013; Sander et al., 2003, 2005, 2018; Scherer & Moors, 2019; Scherer et al., 2001) – differ in the hypothesized mechanisms or dimensions responsible for the elicitation and differentiation of emotion (see, e.g., Sander, 2013). Basic emotion theories propose that there exists a limited set of discrete and universal emotions having each a dedicated evolutionary function. Each basic emotion is thought to be characterized by (a) specific, universal antecedent events, (b) distinctive patterns of physiological activation, (c) distinctive and universal expressive signals, and (d) dedicated neural bases (see, e.g., Ekman, 1999). In opposition with this view, dimensional and appraisal theories assert that emotions are underlain by common dimensions or mechanisms rather than

being modular (e.g., Russell, 2003; Sander et al., 2003, 2005, 2018). On the one hand, dimensional theories posit that emotions can be described along a limited number of (orthogonal) dimensions, such as valence and arousal (or activation). Within this family of emotion theories, core affect theories (e.g., Russell, 2003; Russell & Barrett, 1999) notably hypothesize that an emotional episode originates from the attribution of core affect – a neurophysiological state that is consciously accessible and experienced as a feeling integrating the orthogonal dimensions of hedonic (dis)pleasure (i.e., valence) and arousal (Russell, 2003) – to an object. On the other hand, appraisal theories suggest that emotions are elicited and differentiated according to the individual's appraisal of the stimulus event or situation in relation to their current concerns; different appraisal profiles eventually giving rise to different emotions (e.g., Frijda, 1986, 1988; Moors et al., 2013; Sander et al., 2005, 2018; see also chapter 2.3.4, section "Appraisal theories of emotion").

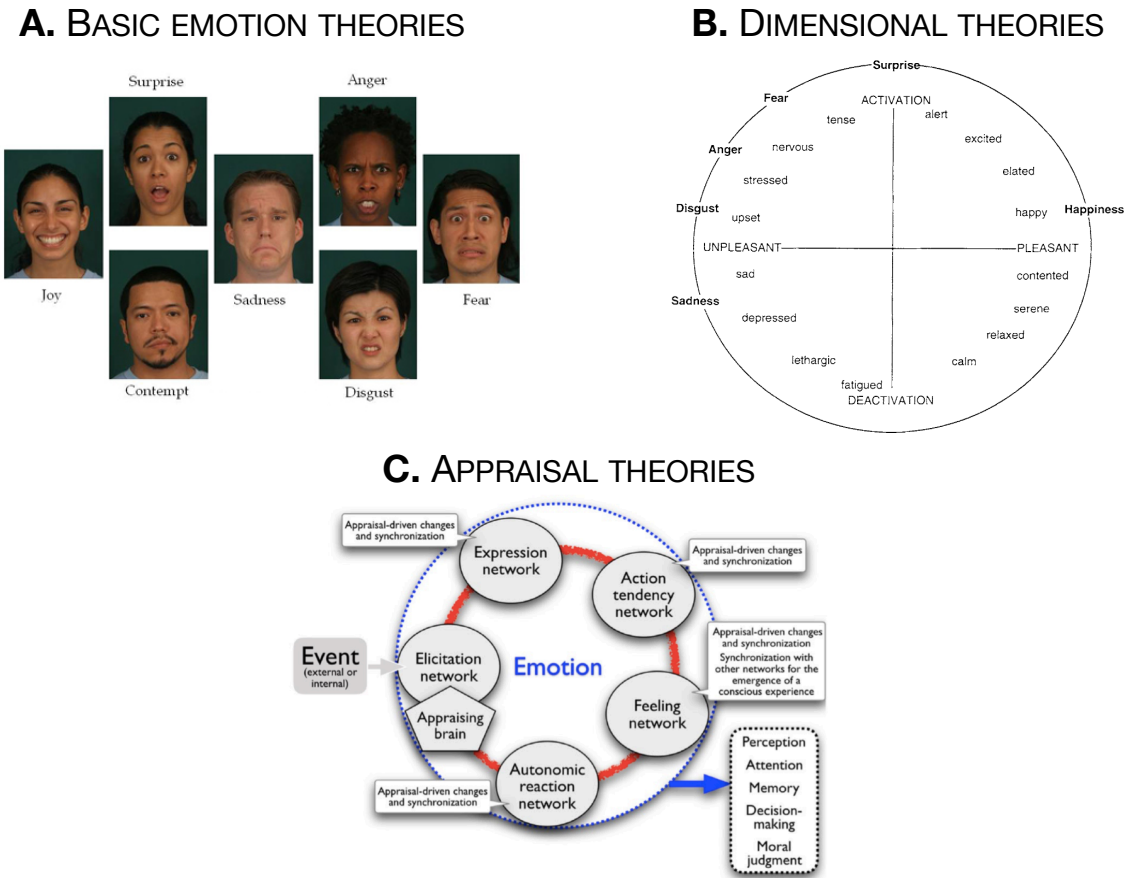

Figure 4.2. Illustrations of the three main families of emotion theories. (A) Basic emotion theories (reproduced from D. Matsumoto & Ekman, 2009). (B) Dimensional or core affect theories (reproduced from Russell & Barrett, 1999). (C) Appraisal theories of emotion (reproduced from Sander et al., 2018).

Correspondingly, these different theories of emotion likewise vary in the proposed mechanisms underlying emotional modulation of cognitive processes, such as attention, learning, memory, or decision-making (see, e.g., Pool, Brosch, et al., 2016; Sander, 2013). Basic emotion theories assume that each basic emotion differentially influences cognitive processes. In particular, fear has been hypothesized to hold a special status compared with other emotions because of its importance to the organism's survival across evolution, as exemplified by the fear module theory for instance (Öhman & Mineka, 2001; see chapter 2.3.2). According to this theory, organisms have been biologically prepared to preferentially process threat-related stimuli from phylogenetic origin, thus selectively biasing attention and learning toward these stimuli (Öhman, Flykt, et al., 2001; Öhman, Lundqvist, et al., 2001; Öhman & Mineka, 2001). In comparison, dimensional or core affect theories posit that core affect influences cognition (Russell, 2003), with valence and arousal being the two candidate dimensions underpinning this influence. In agreement with this perspective, it has been proposed that arousal is the key dimension responsible for the attentional bias for emotional stimuli, irrespective of valence (Anderson, 2005). Arousal has been similarly suggested as the central determinant of memory enhancement for emotional stimuli (e.g., Mather, Clewett, Sakaki, & Harley, 2016; McGaugh, 2004; Sharot & Phelps, 2004). Based on these predictions derived from dimensional theories, the mechanism responsible for emotional modulation of cognitive processes therefore lies in a component of the emotional response. By contrast, appraisal theories state that relevance detection, which is involved in the emotion elicitation process and corresponds to a rapid evaluation of the stimulus after its onset, is the key determinant underlying emotional modulation of cognitive processes (Brosch et al., 2008, 2010, 2013; Montagrin et al., 2013, 2018; Pool, Brosch, et al., 2014, 2016; Sander et al., 2005, 2018; Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018).

In this thesis, we provided evidence suggesting that relevance detection constitutes a key determinant of preferential aversive learning in humans. More specifically, the empirical data from Studies 1 and 2 showed that both threat-relevant and positive relevant stimuli are preferentially associated with an aversive outcome during Pavlovian aversive conditioning, thereby suggesting the existence of a similar functioning for negative threatening and positive rewarding stimuli in emotional learning. In this respect, these results are congruent with the view held by core affect and appraisal theories that emotions, along with their influence on cognitive processes, do not function independently from each other, but are rather underlain by shared dimensions or mechanisms. Our findings notably concur with prior empirical evidence

in the fields of emotional attention and emotional memory. For instance, it has been demonstrated that attention is not only biased toward threatening stimuli with negative valence, but also orients swiftly and involuntarily toward positive stimuli that are affectively relevant in a preferential manner (e.g., Brosch et al., 2007, 2008; Pool, Brosch, et al., 2014, 2016; see also Becker, Anderson, Mortenson, Neufeld, & Neel, 2011). Similarly, emotional stimuli have been reported to induce memory facilitation regardless of whether they are pleasant or unpleasant (e.g., Hamann, Ely, Grafton, & Kilts, 1999; Sharot & Phelps, 2004), and to specifically depend on the stimulus' relevance for the individual's goals (Montagrin et al., 2013, 2018; Montagrin & Sander, 2016). Moreover, visual working memory has likewise been shown to be enhanced both by threatening (e.g., angry and fearful faces; e.g., M. C. Jackson, Wu, Linden, & Raymond, 2009; Sessa, Luria, Gotler, Jolicœur, & Dell'acqua, 2011) and positive (i.e., happy faces; Spotorno, Evans, & Jackson, 2018) facial emotional expressions. At variance with the predictions derived from core affect theories, our results however suggest that arousal does not appear to provide a likely and sufficient factor to satisfactorily account for the emergence of learning biases reported in Studies 1 to 3. Rather, they highlight the key role of the interaction between the stimulus at stake and the individual's concerns in triggering learning biases, as is the case for the modulation of attentional bias (Pool, Brosch, et al., 2016) and for the occurrence of memory facilitation effects (Montagrin et al., 2013, 2018).

Taken together, the findings reported in this thesis align with converging evidence from previous research in affective sciences supporting the appraisal theories' assumption according to which relevance detection is a key mechanism in emotion elicitation as well as in the direct modulation of cognitive processes, without this influence being mediated by the emotional response. The present work hence lends support to the predictions of appraisal theories over those made by basic emotion and dimensional theories in terms of the hypothesized mechanisms underlying emotional modulation of cognitive processes. It further illustrates how the approach suggested by appraisal theories might be particularly useful for investigating the relations between emotions and other psychological constructs, such as attention, learning, memory, and decision-making. In this perspective, appraisal theories of emotion therefore offer a valuable theoretical framework in the study of the (cognitive) mechanisms responsible for the elicitation and differentiation of emotions, as well as their modulatory influence on other psychological processes.

4.2.4. Conceptualization of emotional learning impairments in specific affective disorders

Emotional learning impairments are considered to play a central role in the etiology, maintenance, treatment, and relapse of specific affective disorders, such as anxiety-, phobia-, and addiction-related disorders (see, e.g., Duits et al., 2015; Lissek et al., 2005; Martin-Soelch et al., 2007; Milad & Quirk, 2012; Mineka & Zinbarg, 2006; Seligman, 1971). In particular, a hallmark of these disorders is that they are often easily acquired and can sometimes result from a single traumatic event or a single exposure to an addictive substance, while being typically highly persistent and associated with deficient or weakened extinction learning (e.g., Duits et al., 2015; Milad & Quirk, 2012; Mineka & Zinbarg, 2006; Seligman, 1971). In that sense, such characteristics mirror to some extent the findings of enhanced Pavlovian aversive conditioning to affectively relevant stimuli reported in the present thesis (Stussi, Pourtois, et al., 2018; see also Öhman & Mineka, 2001), thereby suggesting that the relevance detection framework might be suitable to explain the mechanisms contributing to impaired emotional learning.

Whereas the fear module theory postulates that phobias arise from a preferential automatic activation of the fear module jointly determined by biological preparedness and previous aversive experiences in the situation independently of cognitive processes (Mineka & Öhman, 2002; Öhman & Mineka, 2001) – thus explaining in particular the high occurrence of phobia to evolutionarily threat-relevant stimuli, the relevance detection approach deriving from appraisal theories conversely suggests that alterations in emotional learning likely stem from appraisal biases in the subjective evaluation of the stimulus event or the situation's relevance to the organism. In other words, this framework suggests that contradictory, inadequate, and/or involuntary relevance appraisal could be at the heart of the development of altered emotional learning commonly observed in specific affective disorders. More specifically, appraisal theories incorporate the key notion that appraisal processes can occur at different levels of processing (Leventhal & Scherer, 1987), being often automatic and/or implicit rather than deliberative and/or explicit (see, e.g., Arnold, 1960; Moors, 2010). On this basis, a defining feature of dysfunctions in emotional learning may potentially reside in conflicts between implicit and explicit relevance appraisals, which may eventually lead to seemingly irrational and maladaptive emotional responses. For instance, a phobic person or a person suffering from post-traumatic stress disorder might appraise the source object of their disorder as highly threatening, and hence highly relevant for their survival at an implicit level, even though they may be aware that this object is in fact inoffensive or no longer poses a threat for their survival

at a more explicit level. Similarly, "irrational" reward-seeking behaviors that typically characterize addictions could also be possibly underlain by a biased implicit relevance appraisal of cues that have been associated with the addictive reward and/or of the reward itself. Such biases in implicit relevance appraisal might confer these cues an enhanced ability to trigger peaks of wanting to obtain the associated reward despite it is not expected to be liked and is no longer experienced as pleasant by the individual, thus leading to maladaptive reward-seeking behaviors (see, e.g., Pool, Sennwald, et al., 2016).

The relevance detection framework therefore assigns a critical function to affective and cognitive biases in the development of emotional learning impairments, while stressing the importance of considering various levels of processing at which they operate (see also Oyarzún, Càmara, Kouider, Fuentemilla, & de Diego-Balaguer, 2019; Taschereau-Dumouchel et al., 2018; Taschereau-Dumouchel, Liu, & Lau, 2018). Importantly, biases in relevance detection are additionally thought to originate from individual dispositions, environmental contingencies, or the interplay thereof (see Figure 4.1), which suggests the potential involvement of genetic, epigenetic, and/or environmental factors in their emergence (see also Åhs et al., 2018). As affective and cognitive biases in emotional learning are considered to hinge upon inter- and intra-individual differences, the existence of such biases could possibly provide a mechanistic basis that might underlie individual vulnerability and resilience factors in the development of emotional disorders following a traumatic or highly rewarding and reinforcing event (see, e.g., Kalisch et al., 2017; Lonsdorf & Merz, 2017). Accordingly, the relevance detection model of emotional learning could provide a fruitful approach to contribute to an improved understanding of the nature, extent, and determinants of emotional learning impairments associated with specific affective disorders in future research, thus hopefully aiding in developing, validating, and tailoring new individualized and targeted treatments for these conditions (see also Lonsdorf & Merz, 2017).

4.3. LIMITATIONS & FUTURE PERSPECTIVES

In this thesis, we presented empirical studies suggesting that relevance detection is a key determinant of preferential emotional learning in humans. Nonetheless, this work had several limitations, which highlight where new efforts are needed to be made, as well as open new perspectives. Hereafter, we outline some of these limitations and elaborate on new avenues for future research with the aim of further delineating and characterizing the role of relevance detection in emotional learning.

4.3.1. From manipulating to measuring affective relevance

An important limitation of the current thesis pertains to the measurement of stimulus' affective relevance. We indeed did not systematically attempt to measure, either directly or indirectly, the affective relevance value of the various stimuli that we used across our different studies. Because affective relevance is best manipulated than measured, we preferred to base our experiments principally on the manipulation of stimulus' affective relevance according to both strong a priori theoretical grounds and prior empirical findings. Although some attempts have been made at indexing the relevance of a stimulus through the assessment of its affective impact or effect on the individual (Scherer, Dan, & Flykt, 2006; see also Ewbank, Barnard, Croucher, Ramponi, & Calder, 2009), the construct of affective relevance is in fact difficult to adequately measure at a quantitative level, and a suitable, validated instrument that is able to provide a reliable, sensitive, and valid indicator of the stimulus' affective relevance is still lacking.

In Studies 2 and 3, we collected subjective ratings of relevance by asking participants to rate to what extent the different stimuli used were important to them. Nevertheless, these subjective ratings are limited in that (a) they do not offer a pure measure of relevance as they are often influenced or overshadowed by other factors, such as the stimulus' valence, positive stimuli being generally evaluated as subjectively more relevant than negative stimuli despite holding comparable, or even lesser, relevance to the organism from a theoretical standpoint (see chapters 3.2.5 and 3.3.6; see also Montagrin et al., 2013); and (b) they essentially tap into explicit and conscious processes, and may consequently not necessarily reflect relevance detection processes occurring at a more implicit level (see Grandjean, Sander, & Scherer, 2008), and could be contaminated by demand characteristics (e.g., social desirability effects).

Taking this latter consideration into account, we also aimed to measure inter-individual differences in the affective relevance of the various stimuli used in Study 2 by assessing participants' implicit associations between these stimuli and the attribute of importance (versus unimportance) during a Go/No-go Association Task (Nosek & Banaji, 2001). More precisely, we postulated that individuals appraising a given category of stimuli as more relevant to their concerns would associate these stimuli more easily and rapidly with the attribute of importance than that of unimportance in comparison with individuals who do not have this tendency. However, the Go/No-go Association Task that we implemented probably did not provide a direct indicator of the affective relevance or importance value of the three different stimulus categories used. This task indeed indicated that participants more easily associated happy faces with the attribute of importance versus unimportance than angry and neutral faces, whereas the sensitivity index for angry faces was descriptively lower than that for neutral faces, although this difference was not statistically significant. It thus seems that the Go/No-go Association Task likely captured participants' preferences or liking toward the stimulus categories (see chapter 3.2.5; see also Nosek & Banaji, 2001) rather than a pure implicit measure of relevance or importance.

Therefore, the development and validation of reliable and sensitive measures of affective relevance reflecting both implicit and explicit relevance detection processes would ideally represent an important future perspective to offer a more fine-grained assessment of the relevance detection's role in emotional learning, along with the modulatory influence of individual differences in relevance detection exerted on this learning process. These measures could notably allow for testing specific hypotheses regarding the parametric modulation of learning biases during Pavlovian conditioning as a function of the stimulus' level of relevance to the organism (see Figure 4.1). In the absence of such measures, further research should alternatively carefully consider and model the major concerns of the individual when investigating the mechanisms underpinning learning biases in humans – for instance by manipulating their motivational state (see, e.g., Pool, Brosch, et al., 2014; Pool, Pauli, Kress, & O'Doherty, 2019), selecting stimuli individually for each participant (see chapter 3.1.3), or preselecting participants as a function of specific individual dispositions – and rely on a priori theory-driven considerations and extant empirical work to categorize stimuli as affectively relevant.

4.3.2. Preferential Pavlovian aversive conditioning: From a threat-based defensive response to a more general enhanced preparatory response?

In the present thesis, we used skin conductance response as the main dependent variable of the conditioned response. Whereas this measure is a well-established and widely employed psychophysiological indicator of Pavlovian conditioning in humans (e.g., Lonsdorf et al., 2017; see chapter 2.2.3, section "Conditioned response measures"), it is considered a non-specific measure of autonomic activation or anticipatory arousal. In this context, the question arises as to whether the effects of preferential Pavlovian aversive conditioning to affectively relevant stimuli irrespective of their valence observed in Studies 1 to 3 are specifically related to fear or threat, or more generally reflect facilitated and enhanced preparatory responses that are not specific to threat (e.g., an orienting response to significant stimuli; see Bradley, 2009).

In the human Pavlovian aversive conditioning literature, the conditioned response is usually conceived as a fear (e.g., McNally, 1987; Öhman & Mineka, 2001; Watson & Rayner, 1920) or defensive (LeDoux, 2012, 2014; LeDoux & Daw, 2018; Phelps & LeDoux, 2005) response. In particular, the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories clearly posit that the effects of enhanced Pavlovian aversive conditioning to evolutionarily threat-relevant stimuli originate from heightened or intense fear in response to these stimuli as a result of specific evolutionary contingencies, which thereby automatically triggers threat-based psychophysiological and reflexive defensive responses in a readily and persistent manner (see Figure 2.5). Hence, experiments designed to test these theories have generally been based on the assumption that the psychophysiological measures (e.g., skin conductance response) collected during Pavlovian aversive conditioning are direct indices of fear or threat (see, e.g., McNally, 1987; but see, e.g., Maltzman, 1977, for a different view; see also Paré & Quirk, 2017). According to this assumption, it could thus be argued that affectively relevant stimuli provoke enhanced fear- or threat-related responses in aversive contexts as opposed to stimuli with less relevance. Although we cannot completely rule out this possibility, we do not think that affectively relevant stimuli have a higher propensity to elicit threat responses during Pavlovian aversive conditioning per se. Indeed, subjective liking ratings collected after the Pavlovian conditioning procedure in Study 1 revealed that positive relevant stimuli (i.e., baby faces and erotic stimuli) previously associated with an aversive outcome were still evaluated as positive overall, and as more pleasant than previously reinforced neutral (i.e., neutral faces and colored squares) and negative threat-relevant (i.e., angry faces and snake images) stimuli (CSs+). In Study 2, these subjective ratings likewise

indicated that happy faces previously paired with the unconditioned stimulus were evaluated as more pleasant than both neutral and angry face CSs+. Additionally, Study 3 showed no statistical differences between the goal-relevant and goal-irrelevant stimuli in terms of liking ratings after Pavlovian aversive conditioning. Whereas it is worth noting that these subjective ratings should be considered with caution as they were collected solely after extinction but not after acquisition, they critically suggest that positive affectively relevant and goal-relevant stimuli did not elicit overall increased fear or threat compared with neutral, less relevant stimuli at the subjective level. Accordingly, we propose that the effects of preferential Pavlovian aversive conditioning to affectively relevant stimuli likely relate to the elicitation of enhanced preparatory responses to these stimuli, which in turn enable the organism to more readily shape appropriate behaviors, rather than enhanced threat-specific responses.

Nonetheless, further research is needed to pinpoint whether the phenomena of enhanced Pavlovian aversive conditioning in humans are specifically related to threat or more generally relate to heightened anticipatory responses for behavior preparation and execution. In that regard, it would be interesting to explore whether affectively relevant stimuli also produce preferential Pavlovian aversive conditioning independently of their valence when using other indices of the conditioned response than skin conductance response (see also Study 4, chapter 3.4). In particular, the startle eyeblink reflex potentiation has been suggested to be sensitive to valence (Lang et al., 1990), as well as to be specific for fear states and aversive conditioning (e.g., Hamm & Vaitl, 1996; Lipp et al., 1994; but see Bradley et al., 2018, for a recent study suggesting that the startle reflex can be potentiated by anticipation of both aversive and appetitive events). Future studies could therefore assess whether threat-relevant and positive relevant stimuli lead to a similar modulation pattern of the startle eyeblink reflex during Pavlovian aversive conditioning as is the case for skin conductance response, or, alternatively, whether the startle reflex would be specifically potentiated by threat-relevant stimuli but not by positive relevant stimuli, hence reflecting a dissociation between these psychophysiological measures. The inclusion of online trial-by-trial fear ratings could also potentially provide valuable information for disentangling whether positive relevant stimuli, like threat-relevant stimuli, might elicit higher fear ratings; although it should be considered that such ratings may affect the acquisition and extinction of a conditioned response at the physiological level by influencing participants' awareness of the contingencies between the conditioned stimuli and the unconditioned stimulus (see, e.g., Öhman & Mineka, 2001; but see Sjouwerman, Niehaus, Kuhn, & Lonsdorf, 2016). Another interesting line for future research would be to develop

online measures of action tendency during Pavlovian aversive conditioning (see Beckers et al., 2013), which could contribute to determining whether affectively relevant stimuli, as opposed to stimuli with less relevance, elicit enhanced preparatory responses facilitating action during Pavlovian aversive learning regardless of their valence.

4.3.3. On the generality of relevance detection: From aversive to appetitive conditioning

A core prediction of the relevance detection model of emotional learning is that stimuli that are detected as highly relevant to the organism's current concerns are preferentially associated with aversive and appetitive outcomes during Pavlovian conditioning independently of their valence and evolutionary status per se. In line with this prediction, the relevance detection model postulates the existence of a general mechanism underlying preferential emotional learning in humans that is shared not only across negative and positive stimuli, but also across aversive and appetitive contingencies. In the empirical part of this thesis (see chapter 3), we however only examined and systematically tested the role of relevance detection in emotional learning using Pavlovian aversive conditioning procedures. We therefore cannot be sure that our results can generalize to Pavlovian appetitive conditioning, and whether preferential emotional learning to affectively relevant stimuli is restricted to aversive contingencies or extends to appetitive contingencies as well remains to be investigated.

The choice of focusing on Pavlovian aversive conditioning procedures was primarily driven by the need to proceed incrementally in order to determine whether relevance detection is a key psychological mechanism underlying preferential emotional learning in humans. The major theoretical models of emotional learning, such as the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories, posit that only threat-related stimuli encountered by the species across their evolution benefit from preferential emotional learning, thereby putting emphasis on Pavlovian aversive conditioning processes. Accordingly, we decided to first establish whether relevance detection represents a key determinant of preferential Pavlovian aversive conditioning. In addition, Pavlovian appetitive conditioning has been considerably less studied than Pavlovian aversive conditioning in humans (see, e.g., Martin-Soelch et al., 2007), notably because of the difficulty in finding appropriate appetitive unconditioned stimuli that can trigger physiological responses that are comparably as intense as the ones elicited by aversive unconditioned stimuli such as electric stimulations (Hermann

et al., 2002; Martin-Soelch et al., 2007) and a possible lack of sensitivity of psychophysiological measures generally used to detect physiological changes induced by appetitive conditioning (Stussi, Delplanque, et al., 2018). For these reasons, we considered that it was necessary to develop a new psychophysiological measure that could provide a sensitive indicator of human Pavlovian appetitive conditioning before examining the role of relevance detection in appetitive learning more specifically.

In Study 4, we addressed this question and showed that the postauricular reflex constitutes a valid and sensitive psychophysiological measure of Pavlovian appetitive conditioning in humans. This reflex further appears to be more sensitive to appetitive contingencies than both the startle eyeblink reflex and skin conductance response, two of the most commonly employed psychophysiological indices of Pavlovian appetitive conditioning (e.g., Andreatta & Pauli, 2015; Hermann et al., 2000; see chapter 2.2.3, section "Conditioned response measures"). Thus, Study 4 might provide a blueprint on which future studies could be built to assess the generality of a relevance detection mechanism in Pavlovian appetitive conditioning using the postauricular reflex as a valid psychophysiological index of the appetitive conditioned response. In particular, future research should test (a) whether affectively relevant stimuli, as opposed to neutral stimuli with less relevance, benefit from preferential Pavlovian appetitive conditioning irrespective of their valence and evolutionary history as such, and (b) whether the occurrence of such appetitive learning biases depends on the organism's current concerns. Interestingly, a recent study (Pool et al., 2019) has shown that preparatory conditioned responses during Pavlovian appetitive conditioning critically rely on the outcome relevance to the organism's current motivational state. In this study, hungry participants first learned to associate neutral cues with the subsequent delivery of a food reward (CS+) or no reward (CS-) during an acquisition phase. The food reward then either was devalued through a satiation-induced procedure or was not devalued. Following this procedure, the Pavlovian cues were presented again to participants under extinction. Results indicated that the preparatory conditioned response, as measured with pupil dilation, to the Pavlovian cue previously paired with the food reward instantly extinguished when the food reward was devalued, whereas it persisted for several trials during extinction when the food reward was still valued. In this respect, this study provides initial evidence that preferential Pavlovian appetitive learning might hinge upon the stimulus' affective relevance to the organism's current concerns or motivational state, as suggested by the relevance detection model (see Figure 4.1).

4.3.4. Beyond Pavlovian conditioning: The impact of learning biases on instrumental behavior and decision-making

In this work, we primarily focused on examining the role of relevance detection in Pavlovian (aversive) conditioning. As outlined in the theoretical part (see chapter 2.1), emotional learning is nonetheless not limited to Pavlovian conditioning, but encompasses other fundamental forms of learning, such as instrumental conditioning. The Pavlovian and the instrumental (which is often subdivided into a habit system and a goal-directed system) systems have been suggested to be essential in controlling behavior and modulating decision-making in humans (O'Doherty et al., 2017; Rangel et al., 2008). On the one hand, the Pavlovian system assigns affective value to environmental stimuli or preparatory and consummatory behaviors on the basis of innate predispositions, prior learning, or the interplay between them. On the other hand, the instrumental system flexibly assigns value to actions based on their previous reinforcement history in a given context (i.e., habitual system) together with the organism's current goals (i.e., goal-directed system; e.g., Lindström et al., 2015; Rangel et al., 2008). According to this framework, behavior can thus be understood as arising from the functional interactions between these different valuation systems. Importantly, these interactions can result either in an agreement between the Pavlovian and the instrumental systems when the two systems assign high value to the same set of stimuli or actions, or in a conflict when the Pavlovian system assigns high value to a specific set of stimuli or behaviors while the instrumental one assigns high value to another set of actions in a given context (Rangel et al., 2008). Whereas agreements between the Pavlovian and the instrumental systems are thought to usually lead to adaptive decision-making and behavior, conflicts between them can entail poor decision-making and maladaptive behavior (Rangel et al., 2008; see also Breland & Breland, 1961).

In this light, an important and interesting avenue for future research will be to investigate how Pavlovian learning biases to affectively relevant stimuli can influence decision-making and instrumental behavior, as well as characterize this influence. Of particular interest, a study by Lindström et al. (2015) demonstrated that threat-relevant stimuli can either promote or impinge on adaptive instrumental behavior depending on how Pavlovian biases to these stimuli relate to the environment. More specifically, threat-relevant stimuli enhanced adaptive instrumental behavior when they were reliable predictors of danger (i.e., electric stimulation), whereas they disrupted it when they were unreliable predictors of danger; this pattern being consistent across evolutionarily ancient (i.e., snakes and threatening faces) and

novel (i.e., guns and outgroup faces) threat-relevant stimuli. At the computational level, the Pavlovian influence of threat-relevant stimuli was characterized by a bias reflecting competition between the Pavlovian and the instrumental valuation systems: Threat-relevant stimuli downweighed the instrumental value of the action of choosing these stimuli, thereby increasing the probability of avoiding them both when they reliably and unreliably predicted danger. These findings show that Pavlovian biases associated with threat-relevant stimuli acquired through learning have a powerful influence on instrumental behavior, being capable of affecting adaptive and maladaptive behavior according to current environmental contingencies. Nonetheless, this study only examined the impact of threat-relevant stimuli on instrumental behavior, without considering affectively relevant stimuli that are positively valenced. Accordingly, further studies are warranted to elucidate whether the influence of Pavlovian learning biases on decision-making and instrumental behavior is confined to threat-relevant stimuli or more generally extends to positive relevant stimuli as well, as could be conjectured in the light of the results gathered in the current thesis.

4.3.5. The role of relevance detection in emotional learning: From psychological to brain mechanisms

In line with the predictions of the relevance detection model (Stussi et al., 2015, in press; Stussi, Pourtois, et al., 2018), results from the empirical studies of the present thesis demonstrated that enhanced Pavlovian aversive conditioning in humans is not confined to threat-related stimuli, but can also occur to positive rewarding stimuli and initially neutral stimuli that acquired affective relevance. Additional computational analyses using reinforcement learning models (Li et al., 2011; Pearce & Hall, 1980; Rescorla & Wagner, 1972) moreover suggested the involvement of a common computational correlate underpinning learning biases during Pavlovian aversive conditioning in response to both threat-relevant and positive biologically relevant stimuli, thus providing further support for the existence of a general mechanism underlying preferential Pavlovian aversive conditioning in humans that is common across stimuli with affective relevance to the organism independently of their valence (Stussi, Pourtois, et al., 2018). However, as the experimental part of this thesis only included behavioral experiments using psychophysiological measurement, it is important to note that whether the learning biases to negative and positive affectively relevance stimuli are underpinned by shared or distinct mechanisms at the neural level remains to be established.

For instance, it could be argued that enhanced Pavlovian aversive conditioning to threat-relevant stimuli is triggered by a fear module on the one hand, whereas preferential Pavlovian aversive conditioning to positive relevant stimuli relies on another concurrent module dedicated to processing positive, appetitive, or reward-related stimuli with high relevance on the other hand. Contrasting with this view, increasing evidence from neuroimaging has shown that the amygdala, which plays a pivotal role in emotional learning in general and Pavlovian conditioning in particular (e.g., Büchel et al., 1998; Delgado et al., 2006; LaBar et al., 1998; LeDoux, 2012; LeDoux & Daw, 2018; Phelps & LeDoux, 2005), and has been conceived as implementing a fear module (Öhman, 2005; Öhman & Mineka, 2001), is not preferentially activated by threat-related stimuli in a selective manner, but is more generally involved in the processing of stimuli that are deemed relevant to the organism (Cunningham & Brosch, 2012; Pessoa & Adolphs, 2010; Sander et al., 2003; Sergerie et al., 2008), including positive or rewarding stimuli (Gottfried et al., 2003; Sergerie et al., 2008). Additionally, the amygdala has been shown to be a core brain region of the motivational neural circuits underpinning both aversive and appetitive reinforcement learning (Averbeck & Costa, 2017; see also Prévost et al., 2013), being crucially implicated in the computation of both prediction error (i.e., corticomedial nuclei; Boll et al., 2013; see also Niv & Schoenbaum, 2008) and stimulus' associability (i.e., basolateral nuclei; Boll et al., 2013; Li et al., 2011). In addition to the amygdala, the striatum has been identified as another core neural substrate of Pavlovian conditioning, which critically contributes to the computations of prediction-error signals, mainly in the appetitive domain, but also in the aversive one (e.g., Delgado, 2007; Delgado et al., 2008; Li et al., 2011; O'Doherty et al., 2003). This line of neurobiological evidence thereby suggests that the occurrence of preferential Pavlovian learning to both negative and positive relevant stimuli could possibly hinge upon (partially) shared brain networks.

Investigating the neural correlates underlying preferential emotional learning in humans will therefore be fundamental in determining whether the emergence of learning biases to both negative and positive affectively relevant stimuli depend on a shared neural mechanism rather than on different and non-overlapping brain structures. In this perspective, a beneficial endeavor could be to combine the use of a Pavlovian aversive conditioning paradigm comparing the preferential learning to threat-relevant (e.g., angry faces) and positive relevant (e.g., baby faces) stimuli versus neutral, less relevant stimuli (e.g., neutral faces) with neuroimaging methods and computational modeling using model-based functional magnetic resonance imaging (e.g., Boll et al., 2013; Li et al., 2011; O'Doherty et al., 2003; O'Doherty,

Hampton, & Kim, 2007). This multimodal approach appears especially promising for elucidating whether preferential Pavlovian aversive conditioning is underlain by shared neural structures or circuits, including in particular the amygdala and the striatum, across negative and positive stimuli with affective relevance to the organism, as well as providing insights into how such preferential learning is eventually implemented in the human brain. As such, this line of research could help delineate more precisely the psychological, computational, and neural mechanisms responsible for the attribution of a predictive and distinctive emotional value to specific stimuli and behaviors, along with the role of relevance detection therein, thereby contributing to an improved and refined understanding of emotional learning in humans.

4.4. CONCLUSION

In conclusion, the present thesis suggests that preferential Pavlovian aversive conditioning in humans is driven by a general mechanism of relevance detection that is not specific to threat, and hence contributes to establishing and characterizing the role of relevance detection in emotional learning. Relevance detection constitutes a flexible and adaptive mechanism enabling the organism to swiftly and dynamically learn preferentially environmental stimuli that are affectively relevant to their current concerns. Importantly, the relevance detection model allows for accommodating and reinterpreting the extant evidence available in the human conditioning literature on preferential Pavlovian aversive conditioning to threat-relevant stimuli, as these stimuli are highly relevant for the organism's survival, which arguably represents one of the highest prioritized concerns. Ultimately, this model could also contribute to accounting for the high flexibility and large inter-individual differences seen in human emotional learning across varying contexts and situations, as well as some impairments in this process that typically precede or follow the onset and maintenance of specific affective or emotional disorders, such as anxiety-, phobia-, and addiction-related disorders. Although the role of relevance detection remains to be further established and characterized in appetitive learning and at the neural level, the relevance detection framework – and more generally appraisal theories of emotion – provides a promising approach to foster new and better insights into the basic mechanisms underlying emotional learning.

5. REFERENCES

Aaron, R. V., & Benning, S. D. (2016). Postauricular reflexes elicited by soft acoustic clicks and loud noise probes: Reliability, prepulse facilitation, and sensitivity to picture contents. *Psychophysiology, 53*, 1900-1908. https://doi.org/10.1111/psyp.12757

Adolphs, R., Tranel, D., & Buchanan, T. W. (2005). Amygdala damage impairs emotional memory for gist but not details of complex stimuli. *Nature Neuroscience, 8*, 512-518. https://doi.org/10.1038/nn1413

Agras, S., Sylvester, D., & Oliveau, D. (1969). The epidemiology of common fears and phobias. *Comprehensive Psychiatry, 10*, 151-156. https://doi.org/10.1016/0010-440X(69)90022-4

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*, 430-437. https://doi.org/10.1016/j.neubiorev.2018.10.017

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior, 33*, 17-34. https://doi.org/10.1007/s10919-008-0059-5

Amin, J. M., & Lovibond, P. F. (1997). Dissociations between covariation bias and expectancy bias for fear-relevant stimuli. *Cognition and Emotion, 11*, 273-289. https://doi.org/10.1080/026999397379926

Anderson, A. K. (2005). Affective influences on the attentional dynamics supporting awareness. *Journal of Experimental Psychology: General, 134*, 258-281. https://doi.org/10.1037/0096-3445.134.2.258

Andreatta, M., & Pauli, P. (2015). Appetitive vs. aversive conditioning in humans. *Frontiers in Behavioral Neuroscience, 9*, 128. https://doi.org/10.3389/fnbeh.2015.00128

Arnaudova, I., Krypotos, A.-M., Effting, M., Boddez, Y., Kindt, M., & Beckers, T. (2013). Individual differences in discriminatory fear learning under conditions of ambiguity: A vulnerability factor for anxiety disorders? *Frontiers in Psychology, 4*, 298. https://doi.org/10.3389/fpsyg.2013.00298

Arnold, M. B. (1960). *Emotion and personality*. New York, NY: Columbia University Press.

Ashton, M. C., Lee, K., & Paunonen, S. V. (2002). What is the central feature of extraversion? Social attention versus reward sensitivity. *Journal of Personality and Social Psychology, 83*, 245-252. https://doi.org/10.1037/0022-3514.83.1.245

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2018). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0000535

Atlas, L. Y., & Phelps, E. A. (2018). Prepared stimuli enhance aversive learning without weakening the impact of verbal instructions. *Learning & Memory, 25*, 100-104. https://doi.org/10.1101/lm.046359.117

Austin, A. J., & Duka, T. (2010). Mechanisms of attention for appetitive and aversive outcomes in Pavlovian conditioning. *Behavioural Brain Research, 213*, 19-26. https://doi.org/10.1016/j.bbr.2010.04.019

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience, 20*, 505-512. https://doi.org/10.1038/nn.4506

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*, 1225-1229. https://doi.org/10.1126/science.1224313

Bacigalupo, F., & Luck, S. J. (2018). Event-related potential components as measures of aversive conditioning in humans. *Psychophysiology, 55*, e13015. https://doi.org/10.1111/psyp.13015

Balaz, M. A., Kasprow, W. J., & Miller, R. R. (1982). Blocking with a single compound trial. *Animal Learning & Behavior, 10*, 271-276.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science, 1*, 28-58. https://doi.org/10.1111/j.1745-6916.2006.00003.x

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science, 269*, 1115-1118. https://doi.org/10.1126/science.7652558

Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. (2011). The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General, 140*, 637-659. https://doi.org/10.1037/a0024060

Beckers, T., Krypotos, A.-M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning? *Biological Psychology, 92*, 90-96. https://doi.org/10.1016.j.biopsycho.2011.12.015

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology, 68*, 2175-2199. https://doi.org/10.1080/17470218.2015.1009919

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*, 289-300. https://doi.org/10.2307/2346101

Benning, S. D. (2011). Postauricular and superior auricular reflex modulation during emotional pictures and sounds. *Psychophysiology, 48*, 410-414. https://doi.org/10.1111/j.1469-8986.2010.01071.x

Benning, S. D., Patrick, C. J., & Lang, A. R. (2004). Emotional modulation of the post-auricular reflex. *Psychophysiology, 41*, 426-432. https://doi.org/10.1111/j.1469-8986.00160.x

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology, 191*, 391-431. https://doi.org/10.1007/s00213-006-0578-x

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron, 86*, 646-664. https://doi.org/10.1016/j.neuron.2015.02.018

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28*, 309-369. https://doi.org/10.1016/S0165-0173(98)00019-8

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neuroscience, 26*, 507-513. https://doi.org/10.1016/S0166-2236(03)00233-9

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist, 71*, 670-679. https://doi.org/10.1037/amp0000059

Bérzin, F., & Fortinguerra, C. F. H. (1993). EMG study of the anterior, superior, and postauricular muscles in man. *Annals of Anatomy – Anatomischer Anzeiger, 175*, 195-197. https://doi.org/10.1016/S0940-9602(11)80182-2

Blanchette, I. (2006). Snakes, spiders, guns, and syringes: How specific are evolutionary constraints on the detection of threatening stimuli? *Quarterly Journal of Experimental Psychology, 59*, 1484-1504. https://doi.org/10.1080/02724980543000204

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology, 42*, 1–15. https://doi.org/10.1111/j.1469-8986.2005.00271.x

Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., & Beckers, T. (2013). Rating data are underrated: Validity of US expectancy in human fear conditioning. *Journal of Behavior Therapy and Experimental Psychiatry, 44*, 201-206. https://doi.org/10.1016/j.jbtep.2012.08.003

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience, 37*, 758-767. https://doi.org/10.1111/ejn.12094

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological Review, 77*, 32-48. https://doi.org/10.1037/h0028589

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52*, 976-986. https://doi.org/10.1016/S0006-3223(02)01546-9

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Sunderland, MA: Sinauer Associates.

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. *Biological Psychiatry, 60*, 352-360. https://doi.org/10.1016/j.biopsych.2005.12.015

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. *Psychophysiology, 46*, 1-11. https://doi.org/10.1111/j.1469-8986.2008.00702.x

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*, 276-298. https://doi.org/10.1037/1528-3542.1.3.276

Bradley, M. M., Codispoti, M., & Lang, P. J. (2006). A multi-process account of startle modulation during affective perception. *Psychophysiology, 43*, 486-497. https://doi.org/10.1111/j.1469-8986.2006.00412.x

Bradley, M. M., Lang, P. J., & Cuthbert, B. N. (1993). Emotion, novelty, and the startle reflex: Habituation in humans. *Behavioral Neuroscience, 107*, 970-980. https://doi.org/10.1037/0735-7044.107.6.970

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*, 602-607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Bradley, M. M., Zlatar, Z. Z., & Lang, P. J. (2018). Startle reflex modulation during threat of shock and "threat" of reward. *Psychophysiology, 55*, e12989. https://doi.org/10.1111/psyp.12989

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist, 16*, 681-684. https://doi.org/10.1037/h0040090

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433-436. https://doi.org/10.1163/156856897x00357

Bramwell, S., Mallan, K. M., & Lipp, O. V. (2014). Are two threats worse than one? The effects of face race and emotional expression on fear conditioning. *Psychophysiology, 51*, 152-158. https://doi.org/10.1111/psyp.12155

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion, 24*, 377-400. https://doi.org/10.1080/02699930902975754

Brosch, T., Sander, D., Pourtois, G., & Scherer, K. R. (2008). Beyond fear: Rapid spatial orienting toward positive emotional stimuli. *Psychological Science, 19*, 362-370. https://doi.org/10.1111/j.1467-9280.2008.02094.x

Brosch, T., Sander, D., & Scherer, K. R. (2007). That baby caught my eye… Attention capture by infant faces. *Emotion, 7*, 685-689. https://doi.org/10.1037/1528-3542.7.3.685

Brosch, T., Scherer, K. R., Grandjean, D., & Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss Medical Weekly, 143*, w13786. https://doi.org/10.4414/smw.2013.13786

Brosch, T., & Sharma, D. (2005). The role of fear-relevant stimuli in visual search: A comparison of phylogenetic and ontogenetic stimuli. *Emotion, 5*, 360-364. https://doi.org/10.1037/1528-3542.5.3.360

Büchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998). Brain systems mediating aversive conditioning: An event-related fMRI study. *Neuron, 20*, 947-957. https://doi.org/10.1016/S0896- 6273(00)80476-6

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*, 313-323. https://doi.org/10.1037/h0054388

Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. E. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296*, 2191. https://doi.org/10.1126/science.1068749

Christianson, S.-A., & Mjörndal, T. (1985). Adrenalin, emotional arousal and memory. *Scandinavian Journal of Psychology, 26*, 237-248. https://doi.org/10.1111/j.1467-9450.1985.tb01161.x

Clark, J. J., Hollon, N. G., & Phillips, P. E. M. (2012). Pavlovian valuation systems in learning and decision making. *Current Opinion in Neurobiology, 22*, 1054-1061. https://doi.org/10.1016/j.conb.2012.06.004

Cook, E. W., III, Hodes, R. L., & Lang, P. J. (1986). Preparedness and phobia: Effects of stimulus content on human visceral conditioning. *Journal of Abnormal Psychology, 95*, 195-207. https://doi.org/10.1037/0021-843X.95.3.195

Cook, M., & Mineka, S. (1989). Observational conditioning of fear to fear-relevant versus fear-irrelevant stimuli in rhesus monkeys. *Journal of Abnormal Psychology, 98*, 448-459. https://doi.org/10.1037/0021-843X.98.4.448

Cook, M., & Mineka, S. (1990). Selective associations in the observational conditioning of fear in rhesus monkeys. *Journal of Experimental Psychology: Animal Behavior Processes, 16*, 372-389. https://doi.org/10.1037/0097-7403.16.4.372

Coppin, G., Delplanque, S., Bernard, C., Cekic, S., Porcherot, C., Cayeux, I., & Sander, D. (2014). Choice both affects and reflects preferences. *Quarterly Journal of Experimental Psychology, 67*, 1415-1427. https://doi.org/10.1080/17470218.2013.863953

Coppin, G., Pool, E., Delplanque, S., Oud, B., Magot, C., Sander, D., & Van Bavel, J. J. (2016). Swiss identity smells like chocolate: Social identity shapes olfactory judgments. *Scientific Reports, 6*, 34979. https://doi.org/10.1038/srep34979

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10*, 294-300. https://doi.org/10.1016/j.tics.2006.05.004

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10-23. https:/doi.org/10.1016/j.brat.2014.04.006

Critcher, C. R., & Ferguson, M. J. (2016). "Whether I like it or not, it's important": Implicit importance of means predicts self-regulatory persistence and success. *Journal of Personality and Social Psychology, 110*, 818-839. https://doi.org/10.1037/pspa0000053

Critchley, H. D., Elliott, R., Mathias, C. J., & Dolan, R. J. (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience, 20*, 3033-3040.

Cunningham, W. A., & Brosch, T. (2012). Motivational salience: Amygdala tuning from traits, needs, values, and goals. *Current Directions in Psychological Science, 21*, 54-59. https://doi.org/10.1177/0963721411430832

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birmbauer, N., & Lang, P. J. (2000). Brain potentials in affective picture processing: Covariation with autonomic arousal and affective report. *Biological Psychology, 52*, 95-111. https://doi.org/10.1016/S0301-0511(99)00044-7

Davey, G. C. L. (1992). An expectancy model of laboratory preparedness effects. *Journal of Experimental Psychology: General, 121*, 24-40. doi.org/10.1037/0096-3445.121.1.24

Davey, G. C. L. (1995). Preparedness and phobias: Specific evolved associations or a generalized expectancy bias? *Behavioral and Brain Sciences, 18*, 289-325. https://doi.org/10.1017/S0140525X00038498

Davis, M., & Whalen, P. J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry, 6*, 13-34. https://doi.org/10.1038/sj.mp.4000812

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*, 1204-1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*, 1704-1711. https://doi.org/10.1038/nn1560

Dawson, M. E., Schell, A. M., & Banis, H. T. (1986). Greater resistance to extinction of electrodermal responses conditioned to potentially phobic CSs: A noncognitive process? *Psychophysiology, 23*, 552-561. https://doi.org/10.1111/j.1469-8986.1986.tb00673.x

Dawson, M. E., Schell, A. M., & Filion, D. L. (2016). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (4th ed., pp. 217-243). Cambridge, UK: Cambridge University Press.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 473-492. https://doi.org/10.3758/s124115-014-0277-8

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853-869. https://doi.org/10.1037/0033-2909.127.6.853

De Peuter, S., Van Diest, I., Vansteenwegen, D., Van den Bergh, O., & Vlaeyen, J. W. S. (2012). Understanding fear of pain in chronic pain: Interoceptive fear conditioning as a novel approach. *European Journal of Pain, 15*, 889-894. https://doi.org/10.1016/j.ejpain.2011.03.002

Delgado, M. R. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences, 1104*, 70-88. https://doi.org/10.1196/annals.1390.002

Delgado, M. R., Jou, R. L., & Phelps, E. A. (2011). Neural systems underlying aversive conditioning in humans with primary and secondary reinforcers. *Frontiers in Behavioral Neuroscience, 5*, 71. https://doi.org/10.3389/fnins.2011.00071

Delgado, M. R., Labouliere, C. D, & Phelps, E. A. (2006). Fear of losing money? Aversive conditioning with secondary reinforcers. *Social Cognitive and Affective Neuroscience, 1*, 250-259. https://doi.org/10.1093/scan/nsl025

Delgado, M. R., Li, J., Schiller, D., & Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*, 3787-3800. https://doi.org/10.1098/rstb.2008.0161

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

Delplanque, S., N'diaye, K., Scherer, K., & Grandjean, D. (2007). Spatial frequencies or emotional effects? A systematic measure of spatial frequencies for IAPS pictures by a discrete wavelet analysis. *Journal of Neuroscience Methods, 165*, 144-150. https://doi.org/10.1016/j.jneumeth.2007.05.030

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193-222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Dichter, G. S., Benning, S. D., Holtzclaw, T. N., & Bodfish, J. W. (2010). Affective modulation of the startle eyeblink and postauricular reflexes in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 40*, 858-869. https://doi.org/10.1007/s10803-009-0925-y

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274-290. https://doi.org/10.1177/1745691611406920

Dillon, D. G., & LaBar, K. S. (2005). Startle modulation during conscious emotion regulation is arousal-dependent. *Behavioral Neuroscience, 119*, 1118-1124. https://doi.org/10.1037/0735-7044.119.4.1118

Domjan, M. (2005). Pavlovian conditioning: A functional perspective. *Annual Review of Psychology, 56*, 179-206. https://doi.org/10.1146/annurev.psych.55.090902.141409

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20*, 425-443. https://doi.org/10.1016/j.tics.2016.03.014

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., …Baas, J. M. P. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety, 32*, 239-253. https://doi.org/10.1002/da.22353

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature, 520*, 345-348. https://doi.org/10.1038/nature14106

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron, 88*, 47-63. https://doi.org/10.1016/j.neuron.2015.09.028

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207-283). Lincoln, NE: University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169-200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 45-60). New York, NY: John Wiley & Sons Ltd. https://doi.org/10.1002/0470013494.ch3

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: A model of attention in associative learning. *Proceedings of the Royal Society of London B: Biological Sciences, 278*, 2553-2561. https://doi.org/10.1098/rspb.2011.0836

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*, 94-107. https://doi.org/10.1037/h0058559

Esteves, F., Dimberg, U., & Öhman, A. (1994). Automatically elicited fear: Conditioned skin conductance responses to masked facial expressions. *Cognition and Emotion, 8*, 393-413. https://doi.org/10.1080/02699939408408949

Esteves, F., Parra, C., Dimberg, U., & Öhman, A. (1994). Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology, 31*, 375-385. https://doi.org/10.1111/j.1469-8986.1994.tb02446.x

Ewbank, M. P., Barnard, P. J., Croucher, C. J., Ramponi, C., & Calder, A. J. (2009). The amygdala response to images with impact. *Social Cognitive and Affective Neuroscience, 4*, 127-133. https://doi.org/10.1093/scan/nsn048

Eysenck, H. J. (1965). Extraversion and the acquisition of eyeblink and GSR conditioned responses. *Psychological Bulletin, 63*, 258-270. https://doi.org/10.1037/h0021921

Fanselow, M. S., & Bolles, R. C. (1979). Triggering of the endorphin analgesic reaction by a cue previously associated with shock: Reversal by naloxone. *Bulletin of the Psychonomic Society, 14*, 88-90. https://doi.org/10.3758/BF03329408

Fanselow, M. S., & Wassum, K. M. (2016). The origins and organization of vertebrate Pavlovian conditioning. *Cold Harbor Spring Perspectives in Biology, 8*, a021717. https://doi.org/10.1101/cshperspect.a021717

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/BF03193146

Fendt, M., & Fanselow, M. S. (1999). The neuroanatomical and neurochemical basis of conditioned fear. *Neuroscience and Biobehavioral Reviews, 23*, 743-760. https://doi.org/10.1016/S0149-7634(99)00016-0

Ferdenzi, C., Roberts, S. G., Schirmer, A., Delplanque, S., Cekic, S., Porcherot, C., …Grandjean, D. (2013). Variability of affective responses to odors: Culture, gender, and olfactory knowledge. *Chemical Senses, 38*, 175-186. https://doi.org/10.1093/chemse/bjs083

Flykt, A., Esteves, F., & Öhman, A. (2007). Skin conductance responses to masked conditioned stimuli: Phylogenetic/ontogenetic factors versus direction of threat? *Biological Psychology, 74*, 328-336. https://doi.org/10.1016/j.biopsycho.2006.08.004

Fox, E., Griggs, L., & Mouchlianitis, E. (2007). The detection of fear-relevant stimuli: Are guns noticed as quickly as snakes? *Emotion, 7*, 691-696. https://doi.org/10.1037/1528-3542.7.4.691

Franken, I. H. A., Huijding, J., Nijs, I. M. T., & van Strien, J. W. (2011). Electrophysiology of appetitive taste and appetitive taste conditioning in humans. *Biological Psychology, 86*, 273-278. https://doi.org/10.1016/j.biopsycho.2010.12.008

Fredrikson, M., Annas, P., Fischer, H., & Wik, G. (1996). Gender and age differences in the prevalence of specific fears and phobias. *Behaviour Research and Therapy, 34*, 33-39. https://doi.org/10.1016/0005-7976(95)00048-3

Fredrikson, M., Hugdahl, K., & Öhman, A. (1976). Electrodermal conditioning to potentially phobic stimuli in male and female subjects. *Biological Psychology, 4*, 305-313. https://doi.org/10.1016/0301-0511(76)90021-1

French, E. G. (1955). Some characteristics of achievement motivation. *Journal of Experimental Psychology, 50*, 232-236. https://doi.org/10.1037/h0041764

Frijda, N. H. (1986). *The emotions*. London, UK: Cambridge University Press.

Frijda, N. H. (1988). The laws of emotion. *American Psychologist, 43*, 349-358. https://doi.org/10.1037/0003-066X.43.5.349

Frijda, N. H. (2009). Concerns. In D. Sander & K. R. Scherer (Eds.), *The Oxford companion to emotion and the affective sciences* (p. 96). Oxford, UK: Oxford University Press.

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*, 500-508. https://doi.org/10.1038/mp.2015.88

Gable, P. A., & Harmon-Jones, E. (2009). Postauricular reflex responses to picture varying in valence and arousal. *Psychophysiology, 46*, 487-490. https://doi.org/10.1111/j.1469-8986.2009.00794.x

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*, 123-124. https://doi.org/10.3758/BF03342209

Gazendam, F. J., Kamphuis, J. H., Eigenhuis, A., Huizenga, H. M. H., Soeter, M., Bos, M. G. N., …Kindt, M. (2015). Personality predicts individual variation in fear learning: A

multilevel growth modeling approach. *Clinical Psychological Science, 3*, 175-188. https://doi.org/10.1177/2167702614535914

Georgiadis, J. R., & Kringelbach, M. L. (2012). The human sexual response cycle: Brain imaging evidence linking sex to other pleasures. *Progress in Neurobiology, 98*, 49-81. https://doi.org/10.1016/j.pneurobio.2012.05.004

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review, 22*, 1320-1327. https://doi.org/10.3758/s13423-014-0890-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1-6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117*, 197-209. https://doi.org/10.1037/a0017808

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior, 43*, 243-250. https://doi.org/10.3758/s13420-015-0176-z

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife, 6*, e23763. https://doi.org/10.7554/eLife.23763

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior, 40*, 255-268. https://doi.org/10.3758/s13420-012-0080-8

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). State versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*, 585-595. https://doi.org/10.1016/j.neuron.2010.04.016

Gottfried, J. A. (Ed.) (2011). *Neurobiology of sensation and reward.* Boca Raton, FL: CRC Press.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2002). Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *Journal of Neuroscience, 22*, 10829-10837.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science, 301*, 1104-1107. https://doi.org/10.1126/science.1087919

Grady, A. K., Bowen, K. H., Hyde, A. T., Totsch, S. K., & Knight, D. C. (2016). Effect of continuous and partial reinforcement on the acquisition and extinction of human conditioned fear. *Behavioral Neuroscience, 130*, 36-43. https://doi.org/10.1037/bne0000121

Gray, H. (1901/1995). *Anatomy: Descriptive and surgical* (15th ed.). New York: Barnes and Noble Books, Inc.

Gray, J. A. (1987). *The psychology of fear and stress* (2nd ed.). New York, NY: McGraw-Hill.

Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multi-level, appraisal-driven response synchronization. *Consciousness and Cognition, 17*, 484-495. https://doi.org/10.1016/j.concog.2008.03.019

Grandjean, D., & Scherer, K. R. (2008). Unpacking the cognitive architecture of emotion processes. *Emotion, 8*, 341-351. https://doi.org/10.1037/1528-3542.8.3.341

Grillon, C. (2002). Startle reactivity and anxiety disorders: Aversive conditioning, context, and neurobiology. *Biological Psychiatry, 52*, 958-975. https://doi.org/10.1016/S0006-3223(02)01665-7

Grillon, C., & Baas, J. (2003). A review of the modulation of the startle reflex by affective states and its application in psychiatry. *Clinical Neurophysiology, 114*, 1557-1579. https://doi.org/10.1016/S1388-2457(03)00202-5

Grillon, C., & Davis, M. (1997). Fear-potentiated startle conditioning in humans: Explicit and contextual cue conditioning following paired versus unpaired training. *Psychophysiology, 34*, 451-458. https://doi.org/10.1111/j.1469-8986.1997.tb02389.x

Grillon, C., Pellowski, M., Merikangas, K. R., & Davis, M. (1997). Darkness facilitates the acoustic startle reflex in humans. *Biological Psychiatry, 42*, 453-460. https://doi.org/10.1016/S0006-3223(96)00466-0

Haaker, J., Golkar, A., Hermans, D., & Lonsdorf, T. B. (2014). A review on human reinstatement studies: An overview and methodological challenges. *Learning & Memory, 21*, 424-440. https://doi.org/10.1101/lm.036053

Haaker, J., Molapour, T., & Olsson, A. (2016). Conditioned social dominance threat: Observation of others' social dominance biases threat learning. *Social Cognitive and Affective Neuroscience, 11*, 1627-1637. https://doi.org/10.1093/scan/nsw074

Hackley, S. A. (2015). Evidence for a vestigial pinna-orienting system in humans. *Psychophysiology, 52*, 1263-1270. https://doi.org/10.1111/psyp.12501

Hackley, S. A., Muñoz, M. A., Hebert, K., Valle-Inclán, F., & Vila, J. (2009). Reciprocal modulation of eye-blink and pinna-flexion components of startle during reward anticipation. *Psychophysiology, 46*, 1154-1159. https://doi.org/10.1111/j.1469-8986.2009.00867.x

Hackley, S. A., Ren, X., Underwood, A., & Valle-Inclán, F. (2017). Prepulse inhibition and facilitation of the postauricular reflex, a vestigial remnant of pinna startle. *Psychophysiology, 54*, 566-577. https://doi.org/10.1111/psyp.12819

Hackley, S. A., Woldroff, M., & Hillyard, S. A. (1987). Combined use of microreflexes and event-related brain potentials as measure of auditory selective attention. *Psychophysiology, 24*, 632-647. https://doi.org/10.1111/ j.1469-8986.1987.tb00343

Hamann, S., Ely, T. D., Grafton, S. T., & Kilts, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature Neuroscience, 2*, 289-293. https://doi.org/10.1038/6404

Hamann, S., Herman, R. A., Nolan, C. N., & Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. *Nature Neuroscience, 7*, 411-416. https://doi.org/10.1038/nn1208

Hamm, A. O., Greenwald, M. K., Bradley, M. M., & Lang, P. J. (1993). Emotional learning, hedonic change, and the startle probe. *Journal of Abnormal Psychology, 102*, 453-465. https://doi.org/10.1037/0021-843X.102.3.453

Hamm, A. O., & Stark, R. (1993). Sensitization and aversive conditioning: Effects on the startle reflex and electrodermal responding. *Integrative Physiological & Behavioral Science, 28*, 171-176. https://doi.org/10.1007/BF02691223

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology, 33*, 698-710. https://doi.org/10.1111/j.1469-8986.1996.tb02366.x

Hamm, A. O., Vaitl, D., & Lang, P. J. (1989). Fear conditioning, meaning, and belongingness: A selective association analysis. *Journal of Abnormal Psychology, 98*, 395-406. https://doi.org/10.1037/0021-843X.98.4.395

Hare, R., Wood, K., Britain, S., & Shadman, J. (1970). Autonomic responses to affective visual stimulation. *Psychophysiology, 7*, 408-417. https://doi.org/10.1111/j.1469-8986.1970.tb01766.x

Hartley, C. A., Fischl, B., & Phelps, E. A. (2011). Brain structure correlates of individual differences in the acquisition and inhibition of conditioned fear. *Cerebral Cortex, 21*, 1954-1962. https://doi.org/10.1093/cercor/bhq253

Hebert, K. R., Valle-Inclán, F., & Hackley, S. A. (2015). Modulation of eyeblink and postauricular reflexes during the anticipation and viewing of food images. *Psychophysiology, 52*, 509-517. https://doi.org/10.1111/psyp.12372

Hermann, C., Ziegler, S., Birmbauer, N., & Flor, H. (2000). Pavlovian aversive and appetitive odor conditioning in humans: Subjective, peripheral, and electrocortical changes. *Experimental Brain Research, 132*, 203-215. https://doi.org/10.1007/s002210000343

Hess, U., Sabourin, G., & Kleck, R. E. (2007). Postauricular and eyeblink startle responses to facial expressions. *Psychophysiology, 44*, 431-435. https://doi.org/10.1111/j.1469-8986.2007.00516.x

Ho, Y., & Lipp, O. V. (2014). Faster acquisition of conditioned fear to fear-relevant than to nonfear-relevant conditional stimuli. *Psychophysiology, 51*, 810-813. https://doi.org/10.1111/psyp.12223

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin, 136*, 390-421. https://doi.org/10.1037/a0018916

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hugdahl, K. (1978). Electrodermal conditioning to potentially phobic stimuli: Effects of instructed extinction. *Behaviour Research and Therapy, 16*, 315-321. https://doi.org/10.1016/0005-7967(78)90001-3

Hugdahl, K., & Johnsen, B. H. (1989). Preparedness and electrodermal fear-conditioning: Ontogenetic vs phylogenetic explanations. *Behaviour Research and Therapy, 27*, 269-278. https://doi.org/10.1016/0005-7967(89)90046-6

Hugdahl, K., & Kärker, A.-C. (1981). Biological vs experiential factors in phobic conditioning. *Behaviour Research and Therapy, 19*, 109-115. https://doi.org/10.1016/0005-7967(81)90034-6

Hugdahl, K., & Öhman, A. (1977). Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 608-616. https://doi.org/10.1037/0278-7393.3.5.608

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. Oxford, UK: Appleton-Century.

Ischer, M. J., Baron, N., Mermoud, C., Cayeux, I., Porcherot, C., Sander, D., & Delplanque, S. (2014). How incorporation of scents could enhance immersive virtual experiences. *Frontiers in Psychology, 5*, 736. https://doi.org/10.3389/fpsyg.2014.00736

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review, 99*, 561-565. https://doi.org/10.1037/0033-295X.99.3.561

Jackson, D. C., Malmstadt, J. R., Larson, C. L., & Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology, 37*, 515-522. https://doi.org/10.1111/1469-8986.3740515

Jackson, M. C., Wu, C.-Y., Linden, D. E. J., & Raymond, J. E. (2009). Enhanced visual short-term memory for angry faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 363-374. https://doi.org/10.1037/a0013895

Janak, P. H., & Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature, 517*, 284-292. https://doi.org/10.1038/nature14188

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jenkins, W. O., & Stanley, J. C., Jr. (1950). Partial reinforcement: A review and critique. *Psychological Bulletin, 47*, 193-234. https://doi.org/10.1037/h0060772

Jin, J., Zelano, C., Gottfried, J. A., & Mohanty, A. (2015). Human amygdala represents the complete spectrum of subjective valence. *The Journal of Neuroscience, 35*, 15145-15156. https://doi.org/10.1523/JNEUROSCI.2450-15.2015

Johnson, G. M., Valle-Inclán, F., Geary, D. C., & Hackley, S. A. (2012). The nursing hypothesis: An evolutionary account of emotional modulation of the postauricular reflex. *Psychophysiology, 49*, 178-185. https://doi.org/10.1111/j.1469-8986.2011.01297.x

Kagerer, S., Wehrum, S., Klucken, T., Walter, B., Vaitl, D., & Stark, R. (2014). Sex attracts: Investigating individual differences in attentional bias to sexual stimuli. *PLoS ONE, 9*, e107795. https://doi.org/10.1371/journal.pone.0107795

Kalat, J. W., & Rozin, P. (1970). "Salience": A factor which can override temporal contiguity in taste-aversion learning. *Journal of Comparative and Physiological Psychology, 71*, 192-197. https://doi.org/10.1037/h0029158

Kalisch, R., Baker, D. G., Basten, U., Boks, M. P., Bonnano, G. A., Brummelman, E., …Kleim, B. (2017). The resilience framework as a strategy to combat stress-related disorders. *Nature Human Behaviour, 1*, 784-790. https://doi.org/10.1038/s41562-017-0200-8

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9-33). Miami, FL: University Press of Miami.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York, NY: Appleton-Century-Crofts.

Kamin, L. J., & Gaioni, S. J. (1974). Compound conditioned emotional response conditioning with differentially salient elements in rats. *Journal of Comparative and Physiological Psychology, 87*, 591-597. https://doi.org/10.1037/h0036989

Kamin, L. J., & Schaub, R. E. (1963). Effects of conditioned stimulus intensity on the conditioned emotional response. *Journal of Comparative and Physiological Psychology, 56*, 502-507. https://doi.org/10.1037/h0046616

Kennedy, S. J., Rapee, R. M., & Mazurski, E. J. (1997). Covariation bias for phylogenetic versus ontogenetic fear-relevant stimuli. *Behaviour Research and Therapy, 35*, 415-422. https://doi.org/10.1016/S0005-7967(96)00128-3

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male*. Philadelphia, PA: W. B. Saunders.

Klinger, E. (1975). Consequences of commitment to and disengagement from incentives. *Psychological Review, 82*, 1-25. https://doi.org/10.1037/h0076171

Klucken, T., Schweckendiek, J., Merz, C. J., Tabbert, K., Walter, B., Kagerer, S., …Stark, R. (2009). Neural activations of the acquisition of conditioned sexual arousal: Effects on contingency awareness and sex. *Journal of Sexual Medicine, 6*, 3071-3085. https://doi.org/10.1111/j.1743-6109.2009.01405.x

Koch, M., Schmid, A., & Schnitzler, H.-U. (1996). Pleasure-attenuation of startle is disrupted by lesions of the nucleus accumbens. *Neuroreport, 7*, 1442-1446. https://doi.org/10.1097/00001756-199605310-00024

Konorski, J. (1967). *Integrating activity of the brain: An interdisciplinary approach*. Chicago, IL: Chicago University Press.

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

Kremer, E. F. (1978). The Rescorla-Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 4*, 22-36. https://doi.org/10.1037/0097-7403.4.1.22

Kringelbach, M. L., Stark, E. A., Alexander, C., Bornstein, M. H., & Stein, A. (2016). On cuteness: Unlocking the parental brain and beyond. *Trends in Cognitive Sciences, 20*, 545-558. https://doi.org/10.1016/j.tics.2016.05.003

Krypotos, A.-M., Effting, M., Arnaudova, I., Kindt, M., & Beckers, T. (2014). Avoided by association: Acquisition, extinction, and renewal of avoidance tendencies toward conditioned fear stimuli. *Clinical Psychological Science, 2*, 336-343. https://doi.org/10.1177/2167702613503139

Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., & Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain, 131*, 2084-2093. https://doi.org/10.1093/brain/awn136

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7*, 54-64. https://doi.org/10.1038/nrn1825

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron, 20*, 937-945. https://doi.org/10.1016/S0896-6273(00)80475-4

LaBar, K. S., LeDoux, J. E., Spencer, D. D., & Phelps, E. A. (1995). Impaired fear conditioning following unilateral temporal lobectomy in humans. *Journal of Neuroscience, 15*, 6846-6855. https://doi.org/10.1523/JNEUROSCI.15-10-06846.1995

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97*, 377-395. https://doi.org/10.1037/0033-295X.97.3.377

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Tech. Rep. No A-8. Gainesville, FL: University of Florida.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30*, 261-273. 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,* 1377-1388. https://doi.org/10.1080/02699930903485076

Larson, C. L., Aronoff, J., & Stearns, J. J. (2007). The shape of threat: Simple geometric forms evoke rapid and sustained capture of attention. *Emotion, 7*, 526-534. https://doi.org/10.1037/1528-3542.7.3.526

LeDoux, J. E. (1994). Emotion, memory and the brain. *Scientific American, 270*, 50-57. https://doi.org/10.1038/scientificamerican0694-50

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life*. New York, NY: Simon & Schuster.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155-184. https://doi.org/10.1146/annurev.neuro.23.1.155

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron, 73*, 653-676. https://doi.org/10.1016/j.neuron.2012.02.004

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

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviors. *Nature Reviews Neuroscience, 19*, 269-282. https://doi.org/10.1038/nrn.2018.22

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57*, 193-243. https://doi.org/10.1080/02724990344000141

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Willis, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*, 1111-1140. https://doi.org/10.1037/bul0000064

Leuchs, L., Schneider, M., & Spoormaker, V. I. (2019). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology, 56*, e13283. https://doi.org/10.1111/psyp.13283

Leventhal, H., & Scherer, K. R. (1987). The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion, 1*, 3-28. https://doi.org/10.1080/02699938708408361

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14*, 1250-1252. https://doi.org/10.1038/nn.2904

Li, L., & Frost, B. J. (1996). Azimuthal sensitivity of rat pinna reflex: EMG recordings from cervicoauricular muscles. *Hearing Research, 100*, 192-200. https://doi.org/10.1016/0378-5955(96)00119-0

Libukman, T. M., Nichols-Whitehead, P., Griffith, J., & Thomas, R. (1999). Source of arousal and memory for detail. *Memory & Cognition, 27*, 166-190. https://doi.org/10.3758/BF03201222

Lindquist, K. A., Wager, T. D., Kobel, H., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences, 35*, 121-143. https://doi.org/10.1017/S0140525X11000446

Lindström, B., Golkar, A., & Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhanced or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems. *Emotion, 15*, 668-676. https://doi.org/10.1037/emo0000075

Lipp, O. V., Blumenthal, T. D., & Adam, A. R. (2001). Attentional modulation of blink startle at long, short, and very short lead intervals. *Biological Psychology, 58*, 89-103. https://doi.org/10.1016/S0301-0511(01)00109-0

Lipp, O. V., Cronin, S. L., Alhadad, S. S. J., & Luck, C. C. (2015). Enhanced sensitization to animal, interpersonal, and intergroup fear-relevant stimuli (but no evidence for selective one-trial fear learning). *Psychophysiology, 52*, 1520-1528. https://doi.org/10.1111/psyp.12513

Lipp, O. V., & Edwards, M. S. (2002). Effect of instructed extinction on verbal and autonomic indices of Pavlovian learning with fear-relevant and fear-irrelevant conditional stimuli. *Journal of Psychophysiology, 16*, 176-186. https://doi.org/10.1027/0269-8803.16.3.176

Lipp, O. V., Kempich, C., Jee, S. H., & Arnold, D. H. (2014). Fear conditioning to subliminal fear relevant and non fear relevant stimuli. *PLoS ONE, 9*, e99332. https://doi.org/10.1371/journal.pone.0099332

Lipp, O. V., & Purkis, H. M. (2006). The effects of assessment type on verbal ratings of conditional stimulus valence and contingency judgments: Implications for the extinction of evaluative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 32*, 431-440. https://doi.org/10.1037/0097-7403.32.4.431

Lipp, O. V., Sheridan, J., & Siddle, D. A. T. (1994). Human blink startle during aversive and nonaversive Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 20*, 380-389. https://doi.org/10.1037/0097-7403.20.4.380

Lissek, S., Pine, D. S., & Grillon, C. (2006). The strong situation: A potential impediment to studying psychobiology and pharmacology of anxiety disorders. *Biological Psychology, 72*, 265-270. https://doi.org/10.1016/j.biopsycho.2005.11.004

Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behaviour Research Therapy, 43*, 1391-1424. https://doi.org/10.1016/j.brat.2004.10.007

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, 77*, 247-285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans – Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience and Biobehavioral Reviews, 80*, 703-728. https://doi.org/10.1016/j.neubiorev.2017.07.007

Lonsdorf, T. B., Weike, A. I., Nikamo, P., Schalling, M., Hamm, A. O., & Öhman, A. (2009). Genetic gating of human fear learning and extinction. *Psychological Science, 20*, 198-206. https://doi.org/10.1111/j.1467-9280.2009.02280.x

Lorenz, K. (1943). Die angeborenen Formen möglicher Erfahrung [The innate forms of potential experience]. *Zeitschrift für Tierpsychologie, 5*, 235-409. https://doi.org/10.1111/j.1439-0310.1943.tb00655.x

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28*, 3-26. https://doi.org/10.1037/0096-1523.28.1.3

Lovibond, P. F., Siddle, D. A. T., & Bond, N. W. (1993). Resistance to extinction of fear-relevant stimuli: Preparedness or selective sensitization? *Journal of Experimental Psychology: General, 122*, 449-461. https://doi.org/10.1037/0096-3445.122.4.449

Lubow, R. E. (1973). Latent inhibition. *Psychological Review, 79*, 398-407. https://doi.org/10.1037/h0034425

Lucas, R. E., Diener, E., Grob, A., Suh, E. M., Shao, L. (2000). Cross-cultural evidence for the fundamental features of extraversion. *Journal of Personality and Social Psychology, 79*, 452-468. https://doi.org/10.1037/0022-3514.79.3.452

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces – KDEF*. Stockholm, Sweden: Karolinska Institutet, Department of Clinical Neuroscience, Psychology Section.

Lykken, D. T., & Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology, 8*, 656-672. https://doi.org/10.1111/j.1469-8986.1971.tb00501.x

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298. https://doi.org/10.1037/h0076778

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning & Behavior, 4*, 186-192. https://doi.org/10.3758/BF03214033

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, NY: Psychology Press.

Mallan, K. M., Lipp, O. V., & Cochrane, B. (2013). Slithering snakes, angry men and out-group members: What and whom are we evolved to fear? *Cognition and Emotion, 27*, 1168-1180. https://doi.org/10.1080/02699931.2013.778195

Mallan, K. M., Sax, J., & Lipp, O. V. (2009). Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology, 45*, 1303-1307. https://doi.org/10.1016/j.jesp.2009.08.001

Maltzman, I. (1977). Orienting in classical conditioning and generalization of the galvanic skin response to words: An overview. *Journal of Experimental Psychology: General, 106*, 111-119. https://doi.org/10.1037/0096-3445.106.2.111

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience, 24*, 897-931. https://doi.org/10.1146/annurev.neuro.24.1.897

Marks, I. M. (1969). *Fears and phobias*. London, UK: Heineman Medical Books.

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences, 21*, 864-877. https://doi.org/10.1016/j.tics.2017.08.007

Martin-Soelch, C., Linthicum, J., & Ernst, M. (2007). Appetitive conditioning: Neural bases and implications for psychopathology. *Neuroscience & Biobehavioral Reviews, 31*, 426-440. https://doi.org/10.1016/j.neubiorev.2006.11.002

Mather, M., Clewett, D., Sakaki, M., & Harley, C. W. (2016). Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behavioral and Brain Sciences, 39*, e200. https://doi.org/10.1017/S0140525X15000667

Matsumoto, D., & Ekman, P. (2009). Basic emotions. In D. Sander & K. R. Scherer (Eds.), *The Oxford companion to emotion and the affective sciences* (pp. 69-73). Oxford, UK: Oxford University Press.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature, 459*, 837-841. https://doi.org/10.1038/nature08028

Mazurski, E. J., Bond, N. W., Siddle, D. A. T., & Lovibond, P. F. (1996). Conditioning with facial expressions of emotion: Effects of CS sex and age. *Psychophysiology, 33*, 416-425. https://doi.org/10.1111/j.1469-8986.1996.tb01067.x

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27*, 1-28. https://doi.org/10.1146/annurev.neuro.27.070203.144157

McKeachie, W. J. (1961). Motivation, teaching methods, and college learning. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln, NE: University of Nebraska Press.

McNally, R. (1987). Preparedness and phobias: A review. *Psychological Bulletin, 101*, 283-303. https://doi.org/10.1037/0033-2909.101.2.283

Mertens, G., Raes, A. K., & De Houwer, J. (2016). Can prepared fear conditioning result from verbal instructions? *Learning and Motivation, 53*, 7-23. https://doi.org/10.1016/j.lmot.2015.11.001

Milad, M. R., Goldstein, J. M., Orr, S. P., Wedig, M. M., Klibanski, A., Pitman, R. K., & Rausch, S. L. (2006). Fear conditioning and extinction: Influence of sex and menstrual cycle in healthy humans. *Behavioral Neuroscience, 120*, 1196-1203. https://doi.org/10.1037/0735-7044.120.5.1196

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: Ten years of progress. *Annual Review of Psychology, 63*, 129-151. https://doi.org/10.1146/annurev.psych.121208.131631

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117*, 363-386. https://doi.org/10.1037/0033-2909.117.3.363

Mineka, S., Davidson, M., Cook, M., & Keir, R. (1984). Observational conditioning of snake fear in rhesus monkeys. *Journal of Abnormal Psychology, 93*, 355-372. https://doi.org/10.1037/0021-843X.93.4.355

Mineka, S., & Öhman, A. (2002). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological Psychiatry, 52*, 927-937. https://doi.org/10.1016/S0006-3223(02)01669-4

Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders. *American Psychologist, 61*, 10-26. https://doi.org/10.1037/0003-066X.61.1.10

Miskovic, V., & Keil, A. (2012). Acquired fears reflected in cortical sensory processing: A review of electrophysiological studies of human classical conditioning. *Psychophysiology, 49*, 1230-1241. https://doi.org/10.1111/j.1469-8986.2012.01398.x

Mobbs, D., & Kim, J. J. (2015). Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Current Opinion in Behavioral Sciences, 5*, 8-15. https://doi.org/10.1016/j.cobeha.2015.06.005

Montagrin, A., Brosch, T., & Sander, D. (2013). Goal conduciveness as a key determinant of memory facilitation. *Emotion, 13*, 622-628. https://doi.org/10.1037/a0033066

Montagrin, A., & Sander, D. (2016). Emotional memory: From affective relevance to arousal. *Behavioral and Brain Sciences, 39*, e216. https://doi.org/10.1017/S0140525X15001879

Montagrin, A., Sterpenich, V., Brosch, T., Grandjean, D., Armony, J., Ceravolo, L., & Sander, D. (2018). Goal-relevant situations facilitate memory of neutral faces. *Cognitive, Affective, & Behavioral Neuroscience, 18*, 1269-1282. https://doi.org/10.3758/s13415-018-0637-x

Moors, A. (2010). Automatic constructive appraisal as a candidate cause of emotion. *Emotion Review, 2*, 139-156. https://doi.org/10.1177/1754073909351755

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review, 5*, 119-124. https://doi.org/10.1177/1754073912468165

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61-64. https://doi.org/10.20982/tqmp.04.2.p061

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393*, 467-470. https://doi.org/10.1038/30976

Morris, J. S., Öhman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 1680-1685. https://doi.org/10.1073/pnas.96.4.1680

Murray, H. A. (1938). *Explorations in personality*. Oxford, UK: Oxford University Press.

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., …Tye, K. M. (2015). A circuit mechanism for differentiating positive and negative associations. *Nature, 520*, 675-678. https://doi.org/10.1038/nature14366

Nasser, H. M., & Delamater, A. R. (2016). The determining conditions for Pavlovian learning: Psychological and neurobiological considerations. In R. A. Murphy & R. C. Honey (Eds.), *The Wiley handbook on the cognitive neuroscience of learning* (pp. 7-46). Chichester, West Sussex, UK: John Wiley & Sons Inc.

Navarrete, C. M., McDonald, M. M., Asher, B. D., Kerr, N. L., Yokota, K., Olsson, A., & Sidanius, J. (2012). Fear is readily associated with an out-group face in a minimal group context. *Evolution and Human Behavior, 33*, 590-593. https://doi.org/10.1016/j.evolhumanbehav.2012.02.007

Navarrete, C. M., 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, 20*, 155-158. https://doi.org/10.1111/j.1467-9280.2009.02273.x

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review, 91*, 328-346. https://doi.org/10.1037/0033-295X.91.3.328

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience, 32*, 551-562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction error. *Trends in Cognitive Sciences, 12*, 265-272. https://doi.org/10.1016/j.tics.2008.03.006

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition, 19*, 625-666. https://doi.org/10.1521/soco.19.6.625.20886

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision-making. *Annual Review of Psychology, 68*, 73-100. https://doi.org/10.1146/annurev-psych-010416-044216

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38*, 329-337. https://doi.org/10.1016/S0896-6237(03)00169-7

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences, 1104*, 35-53. https://doi.org/10.1196/annals.1390.022

Öhman, A. (1979). Fear relevance, autonomic conditioning, and phobias: A laboratory model. In P.-O. Sjödén, S. Bates, & W. S. Dockens, III (Eds.), *Trends in behavior therapy* (pp. 107-133). New York, NY: Academic Press.

Öhman, A. (1986). Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology, 23*, 123-145. https://doi.org/10.1111/j.1469-8986.1986.tb00608.x

Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology, 30*, 953-958. https://doi.org/10.1016/j.psyneuen.2005.03.019

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

Öhman, A., Eriksson, A., & Olofsson, C. (1975). One-trial learning and superior resistance to extinction of autonomic responses conditioned to potentially phobic stimuli. *Journal of Comparative and Physiological Psychology, 88*, 619-627. https://doi.org/10.1037/h0078388

Öhman, A., Erixon, G., & Löfberg, I. (1975). Phobias and preparedness: Phobic versus neutral pictures as conditioned stimuli for human autonomic responses. *Journal of Abnormal Psychology, 84*, 41-45. https://doi.org/10.1037/h0076255

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

Öhman, A., Fredrikson, M., & Hugdahl, K. (1978). Orienting and defensive responding in the electrodermal system: Palmar-dorsal differences and recovery rate during conditioning to potentially phobic stimuli. *Psychophysiology, 15*, 93-101. https://doi.org/10.1111/j.1469-8986.1978.tb01342.x

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmö, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105*, 313-337. https://doi.org/10.1037/0096-3445.105.4.313

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*, 381-396. https://doi.org/10.1037/0022-3514.80.3.381

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

Öhman, A., & Soares, J. J. (1993). On the automatic nature of phobic fear: Conditioned electrodermal responses to masked fear-relevant stimuli. *Journal of Abnormal Psychology, 102*, 121-132. https://doi.org/10.1037/0021-843X.102.1.121

Öhman, A., & Soares, J. J. F. (1998). Emotional conditioning to masked stimuli: Expectancies for aversive outcomes following nonrecognized fear-relevant stimuli. *Journal of Experimental Psychology: General, 127*, 69-82. https://doi.org/10.1037/0096-3445.127.1.69

Öhman, A., & Wiens, S. (2004). The concept of an evolved fear module and cognitive theories of anxiety. In A. S. R. Manstead, N. H. Frijda, & A. H. Fischer (Eds.), *Feelings and emotions: The Amsterdam symposium* (pp. 58-80). Cambridge, UK: Cambridge University Press.

Olsson, A., Carmona, S., Downey, G., Bolger, N., & Ochsner, K. N. (2013). Learning biases underlying individual differences in sensitivity to social rejection. *Emotion, 13*, 616-621. https://doi.org/10.1037/a0033150

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science, 309*, 785-787. https://doi.org/10.1126/science.1113551

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science, 15*, 822-828. https://doi.org/10.1111/j.0956-7976.2004.00762.x

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*, 1095-1102. https://doi.org/10.1038/nn1968

Oyarzún, J. P., Càmara, E., Kouider, S., Fuentemilla, L., & de Diego-Balaguer, R. (2019). Implicit but not explicit extinction to threat-conditioned stimulus prevents spontaneous recovery of threat-potentiated startle responses in humans. *Brain and Behavior, 9*, e01157. https://doi.org/10.1002/brb3.1157

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

Papini, M. R., & Bitterman, M. E. (1990). The role of contingency in classical conditioning. *Psychological Review, 97*, 396-403. https://doi.org/10.1037/0033-295X.97.3.396

Pappens, M., Smets, E., Vansteenwegen, D., Van den Bergh, O., & Van Diest, I. (2012). Learning to fear suffocation: A new paradigm for interoceptive fear conditioning. *Psychophysiology, 49*, 821-828. https://doi.org/10.1111/j.1469-8986.2012.01357.x

Paré, D., & Quirk, G. J. (2017). When scientific paradigms lead to tunnel vision: Lessons from the study of fear. *npj Science of Learning, 2*, 6. https://doi.org/10.1038/s41539-017-0007-4

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107-123. https://doi.org/10.1016/S0042-6989(01)00250-4

Parsons, C. E., Young, K. S., Kumari, N., Stein, A., & Kringelbach, M. L. (2011). The motivational salience of infant faces is similar for men and women. *PLoS ONE, 6*, e20632. https://doi.org/10.1371/journal.pone.0020632

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature, 439*, 865-870. https://doi.org/10.1038/nature04490

Pauli, W. M., Gentile, G., Collette, S., Tyszka, J. M., & O'Doherty, J. P. (2019). Evidence for model-based encoding of Pavlovian contingencies in the human brain. *Nature Communications, 10*, 1099. https://doi.org/10.1038/s41467-019-08922-7

Pauli, W. M., Larsen, T., Collette, S., Tyszka, J. M., Seymour, B., & O'Doherty, J. P. (2015). Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning. *Journal of Neuroscience, 35*, 14220-14233. https://doi.org/10.1523/JNEUROSCI.2277-15.2015

Pavlov, I. P. (1927). *Conditioned reflexes*. London, UK: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review, 87*, 532-552. https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability. In M. L. Commons, R. J. Hernstein, & A. R. Wagner (Eds.), *Quantitative analyses of behavior. Vol. 3. Acquisition* (pp. 241-256). Cambridge, MA: Ballinger.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442. https://doi.org/10.1163/156856897x00366

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a 'low road' to 'many roads' for evaluating biological significance. *Nature Reviews Neuroscience, 11*, 773-783. https://doi.org/10.1038/nrn2920

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Reviews of Psychology, 57*, 27-53. 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*, 897-905. https://doi.org/10.1016/j.neuron.2004.08.042

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*, 175-187. https://doi.org/10.1016/j.neuron.2005.09.025

Pineles, S. L., Vogt, D. S., & Orr, S. P. (2009). Personality and fear responses during conditioning: Beyond extraversion. *Personality and Individual Differences, 46*, 48-53. https://doi.org/10.1016/j.paid.2008.09.003

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2014). Where is the chocolate? Rapid spatial orienting toward stimuli associated with primary rewards. *Cognition, 130*, 348-359. https://doi.org/10.1016/j.cognition.2013.12.002

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2015). Stress increases cue-triggered "Wanting" for sweet reward in humans. *Journal of Experimental Psychology: Animal Learning and Cognition, 41*, 128-136. https://doi.org/10.1037/xan0000052

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin, 142*, 79-106. https://doi.org/10.1037/bul0000026

Pool, E., Delplanque, S., Porcherot, C., Jenkins, T., Cayeux, I., & Sander, D. (2014). Sweet reward increases implicit discrimination of similar odors. *Frontiers in Behavioral Neuroscience, 8*, 158. https://doi.org/10.3389/fnbeh.2014.00158

Pool, E. R., Pauli, W. M., Kress, C. S, & O'Doherty, J. P. (2019). Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in

humans. *Nature Human Behaviour, 3*, 284-296. https://doi.org/10.1038/s41562-018-0572-9

Pool, E., Sennwald, V., Delplanque, S., Brosch, T., & Sander, D. (2016). Measuring wanting and liking from animals to humans: A systematic review. *Neuroscience and Biobehavioral Reviews, 63*, 124-142. https://doi.org/10.1016/j.neurobiorev.2016.01.006

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology, 9*, e1002918. https://doi.org/10.1371/journal.pcbi.1002918

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/

Randich, A., & LoLordo, V. M. (1979). Associative and nonassociative theories of the UCS preexposure phenomenon: Implications for Pavlovian conditioning. *Psychological Bulletin, 86*, 523-548. https://doi.org/10.1037/0033-2909.86.3.523

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9*, 545-556. https://doi.org/10.1038/nrn2357

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

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review, 74*, 71-80. https://doi.org/10.1037/h0024109

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1-5. https://doi.org/10.1037/h0025984

Rescorla, R. A. (1980). *Pavlovian second order conditioning: Studies in associative learning*. Hillsdale, NJ: Erlbaum.

Rescorla, R. A. (1988a). Behavioral studies of Pavlovian conditioning. *Annual Reviews of Neuroscience, 11*, 329-352. https://doi.org/10.1146/annurev.ne.11.030188.001553

Rescorla, R. A. (1988b). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151-160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Holland, P. C. (1976). Some behavioral approaches to the study of learning. In M. R. Rosenzweig & E. L. Bennett (Eds.), *Neural mechanisms of learning* (pp. 165-192). Cambridge, MA: MIT Press.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review, 74*, 151-182. https://doi.org/10.1037/h0024475

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prosky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Rimpel, J., Geyer, D., & Hopf, H. C. (1982). Changes in the blink responses to combined trigeminal, acoustic, and visual repetitive stimulation, studied in the human subject. *Electroencephalography and Clinical Neurophysiology, 54*, 552-560. https://doi.org/10.1016/0013-4694(82)90040-2

Robbins, T. W. (1997). Arousal systems and attentional processes. *Biological Psychology, 45*, 57-71. https://doi.org/10.1016/S0301-0511(96)05222-2

Rodriguez Mosquera, P. M., Fischer, A. H., & Manstead, A. S. R. (2004). Inside the heart of emotion: On culture and relational concerns. In L. Z. Tiedens & C. W. Leach (Eds.), *The social life of emotions* (pp. 187-202). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511819568.011

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience, 35*, 1190-1200. https://doi.org/10.1111/j.1460-9568.2011.07986.x

Rolland, J. P., Parker, W. D., & Strumpf, H. (1998). A psychometric examination of the French translations of the NEO-PI-R and NEO-FFI. *Journal of Personality Assessment, 71*, 269-291. https://doi.org/10.1207/s15327752jpa7102_13

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York, NY: Cambridge University Press.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting or rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. https://doi.org/10.3758/PBR.16.2.225

Rowles, M. E., Lipp, O. V., & Mallan, K. M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology, 49*, 375-380. https://doi.org/10.1111/j.1469-8986.2011.01308.x

Rupp, H. A., & Wallen, K. (2008). Sex differences in response to visual sexual stimuli: A review. *Archives of Sexual Behavior, 37*, 206-218. https://doi.org/10.1007/s10508-007-9217-9

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145-172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodeds, and other things called *emotion*: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*, 805-819. https://doi.org/10.1037/0022-3514.76.5.805

Sander, D. (2013). Models of emotion: The affective neuroscience approach. In J. L. Armony & P. Vuilleumier (Eds.), *The Cambridge Handbook of human affective neuroscience* (pp. 5-53). Cambridge, UK: Cambridge University Press.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*, 303-316. https://doi.org/10.1515/REVNEURO.2003.14.4.303

Sander, D., Grandjean, D., Kaiser, S., Wehrle, T., & Scherer, K. R. (2007). Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology, 19*, 470-480. https://doi.org/10.1080/09541440600757426

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*, 317-352. https://doi.org/10.1016/j.neunet.2005.03.001

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review, 10*, 219-231. https://doi.org/10.1177/1754073918765653

Sandt, A. R., Sloan, D. M., & Johnson, K. J. (2009). Measuring appetitive responding with the postauricular reflex. *Psychophysiology, 46*, 491-497. https://doi.org/10.1111/j.1469-8986.2009.00797.x

Schell, A. M., Dawson, M. E., & Marinkovic, K. (1991). Effects of potentially phobic conditioned stimuli on retention, reconditioning, and extinction of conditioned skin conductance response. *Psychophysiology, 28*, 140-153. https://doi.org/10.1111/j.1469-8986.1991.tb00403.x

Scherer, K. R. (1994). Emotion serves to decouple stimulus and response. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 127-130). New York, NY: Oxford University Press.

Scherer, K. R., Dan, E. S., & Flykt, A. (2006). What determines a feeling's position in affective space? A case for appraisal. *Cognition and Emotion, 20*, 92-113. https://doi.org/10.1080/02699930500305016

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology, 70*, 719-745. https://doi.org/10.1146/annurev-psych-122216-011854

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. New York, NY: Oxford University Press.

Schiller, D., Monfils, M.-H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*, 49-53. https://doi.org/10.1038/nature08637

Schönbrodt, F. D., & Gerstenberg, F. X. R. (2012). An IRT analysis of motive questionnaires: The Unified Motive Scales. *Journal of Research in Personality, 46*, 725-742. https://doi.org/10.1016/j.jrp.2012.08.010

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Review, 95*, 853-951. https://doi.org/10.1152/physrev.00023.2014

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464. https://doi.org/10.1214/aos/1176344136

Sehlmeyer, C., Schöning, S., Zwitserlood, P., Pfleiderer, B., Kircher, T., Arolt, V., & Konrad, C. (2009). Human fear conditioning and extinction in neuroimaging: A systematic review. *PLoS ONE, 4*, e5865. https://doi.org/10.1371/journal.pone.0005865

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

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2*, 307-320. https://doi.org/10.1016/S0005-7894(71)80064-3

Sennwald, V., Pool, E., Brosch, T., Delplanque, S., Bianchi-Demicheli, F., & Sander, D. (2016). Emotional attention for erotic stimuli: Cognitive and brain mechanisms. *The Journal of Comparative Neurology, 524*, 1668-1675. https://doi.org/10.1002/cne.23859

Sennwald, V., Pool, E., Delplanque, S., Brosch, T., Bianchi-Demicheli, F., & Sander, D. (2018). *Inter-individual differences underlie cue-triggered 'wanting' for sexual reward.* Manuscript in preparation.

Sennwald, V., Pool, E., & Sander, D. (2017). Considering the influence of the Pavlovian system on behavior: Appraisal and value representation. *Psychological Inquiry, 28*, 52-55. https://doi.org/10.1080/1047840X.2017.1259951

Sergerie, K., Chochol, C., & Armony, J. L. (2008). The role of the amygdala in emotional processing: A quantitative meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 32*, 811-830. https://doi.org/10.1016/j.neubiorev.2007.12.002

Sessa, P., Luria, R., Gotler, A., Jolicœur, P., & Dell'acqua, R. (2011). Interhemispheric ERP asymmetries over inferior parietal cortex reveal differential visual working memory maintenance for fearful versus neutral face identities. *Psychophysiology, 48*, 187-197. https://doi.org/10.1111/j.1469-8986.2010.01046.x

Severo, M. C., Walentowska, W., Moors, A., & Pourtois, G. (2017). Goal impact influences the evaluative component of performance monitoring: Evidence from ERPs. *Biological Psychology, 129*, 90-102. https://doi.org/10.1016/j.biopsycho.2017.08.052

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience, 27*, 4826-4831. https://doi.org/10.1523/JNEUROSCI.0400-07.2007

Shabel, S. J., & Janak, P. H. (2009). Substantial similarity in amygdala neuronal activity during conditioned appetitive and aversive emotional arousal. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 15031-15036. https://doi.org/10.1073/pnas.0905580106

Sharot, T., & Phelps, E. A. (2004). How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience, 4*, 294-306. https://doi.org/10.3758/CABN.4.3.294

Siddle, D. A. T., & Bond, N. W. (1988). Avoidance learning, Pavlovian conditioning, and the development of phobias. *Biological Psychology, 27*, 167-183. https://doi.org/10.1016/0301-0511(88)90048-8

Sjouwerman, R., Niehaus, J., Kuhn, M., & Lonsdorf, T. B. (2016). Don't startle me – Interference of starle probe presentations and intermittent ratings with fear acquisition. *Psychophysiology, 53*, 1889-1899. https://doi.org/10.1111/psyp.12761

Sjouwerman, R., Scharfenort, R., & Lonsdorf, T. B. (2018, January 4). Individual differences in fear learning: Specificity to trait-anxiety beyond other measures of negative affect, and mediation via amygdala activation. *bioRxiv*, 233528. https://doi.org/10.1101/233528

Smillie, L. D. (2013). Extraversion and reward processing. *Current Directions in Psychological Science, 22*, 167-172. https://doi.org/10.1177/0963721412470133

Smith, C. A., & Pope, K. L. (1992). Appraisal and emotion: The interactional contributions of dispositional and situational factors. In M. S. Clark (Ed.), *Review of personality and social psychology*: *Vol. 14. Emotion and social behavior* (pp. 32-62). Newbury Park, CA: Sage.

Soares, J. J. F., & Öhman, A. (1993). Preattentive processing, preparedness and phobias: Effects of instruction on conditioned electrodermal responses to masked and non-masked fear-relevant stimuli. *Behaviour Research and Therapy, 31*, 87-95. https://doi.org/10.1016/0005-7967(93)90046-W

Sollers, J. J., & Hackley, S. A. (1997). Effect of foreperiod duration on reflexive and voluntary response to intense noise bursts. *Psychophysiology, 34*, 518-526. https://doi.org/10.1111/j.1469-8986.1997.tb01738.x

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review, 81*, 119-145. https://doi.org/10.1037/h0036128

Spector, I. P., Carey, M. P., & Steinberg, L. (1996). The sexual desire inventory: Development, factor structure, and evidence of reliability. *Journal of Sex & Marital Therapy, 22*, 175-190. https://doi.org/10.1080/00926239608414655

Spotorno, S., Evans, M., & Jackson, M. C. (2018). Remembering who was where: A happy expression advantage for face identity-location binding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1365-1383. https://doi.org/10.1037/xlm0000522

Stanley, J., & Knight, R. G. (2004). Emotional specificity of startle potentiation during the early stages of picture viewing. *Psychophysiology, 41*, 935-940. https://doi.org/10.1111/j.1469-8986.2004.00242.x

Stolarova, M., Keil, A., & Moratti, S. (2006). Modulation of the C1 visual-event related component by conditioned stimuli: Evidence for sensory plasticity in early affective perception. *Cerebral Cortex, 16*, 876-887. https://doi.org/10.1093/cercor/bhj031

Stussi, Y., Brosch, T., & Sander, D. (2015). Learning to fear depends on emotion and gaze interaction: The role of self-relevance in fear learning. *Biological Psychology, 109*, 232-238. https://doi.org/10.1016/j.biopsycho.2015.06.008

Stussi, Y., Delplanque, S., Coraj, S., Pourtois, G., & Sander, D. (2018). Measuring Pavlovian appetitive conditioning in humans with the postauricular reflex. *Psychophysiology, 55*, e13073. https://doi.org/10.1111/psyp.13073

Stussi, Y., Ferrero, A., Pourtois, G., & Sander, D. (in press). Achievement modulation modulates Pavlovian aversive conditioning to goal-relevant stimuli. *npj Science of Learning*.

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

Subra, B., Muller, D., Fourgassie, L., Chauvin, A., & Alexopoulos, T. (2017). Of guns and snakes: Testing a modern threat superiority effect. *Cognition and Emotion, 32*, 81-91. https://doi.org/10.1080/02699931.2017.12884044

Sutton, R. S., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Talmi, D., Ziegler, M., Hawksworth, J., Lalani, S., Herman, C. P., & Moscovitch, M. (2013). Emotional stimuli exert parallel effects on attention and memory. *Cognition and Emotion, 27*, 530-538. https://doi.org/10.1080/02699931.2012.722527

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 3470-3475. https://doi.org/10.1073/pnas.1721572115

Taschereau-Dumouchel, V., Liu, K.-Y., & Lau, H. (2018). Unconscious psychological treatments for physiological survival circuits. *Current Opinion in Behavioral Sciences, 24*, 62-68. https://doi.org/10.1016/j.cobeha.2018.04.010

Taylor, K. M., & Boakes, R. A. (2002). Extinction of conditioned taste aversions: Effects of concentration and overshadowing. *Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 55*, 213-239. https://doi.org/10.1080/02724990143000270

Thorndike, E. L. (1898). *Animal intelligence: An experimental study of the associative processes in animals*. New York, NY: Columbia University Press.

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

Tomarken, A. J., Sutton, S. K., & Mineka, S. (1995). Fear-relevant illusory correlations: What types of associations promote judgmental bias? *Journal of Abnormal Psychology, 104*, 312-326. https://doi.org/10.1037/0021-843X.104.2.312

Tooley, M. D., Carmel, D., Chapman, A., & Grimshaw, G. M. (2017). Dissociating the physiological components of unconscious emotional responses. *Neuroscience of Consciousness, 1*, nix021. https://doi.org/10.1093/nc/nix021

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vaitl, D., & Lipp, O. V. (1997). Latent inhibition and autonomic responses: A psychophysiological approach. *Behavioural Brain Research, 88*, 85-93. https://doi.org/10.1016/S0166-4328(97)02310-3

Van Duuren, M., Kendell-Scott, L., & Stark, N. (2003). Early aesthetic choices: Infant preferences for attractive premature infant faces. *International Journal of Behavioral Development, 27*, 212-219. https://doi.org/10.1080/01650250244000218

Van Gucht, D., Baeyens, F., Vansteenwegen, D., Hermans, D., & Beckers, T. (2010). Counterconditioning reduces cue-induced craving and actual cue-elicited consumption. *Emotion, 10*, 688-695. https://doi.org/10.1037/a0019463

Van Gucht, D., Vansteenwegen, D., Van den Bergh, O., & Beckers, T. (2008). Conditioned craving cues elicit an automatic approach tendency. *Behaviour Research and Therapy, 46*, 1160-1169. https://doi.org/10.1016/j.brat.2008.05.010

Vogt, J., De Houwer, J., Moors, A., Van Damme, S., & Crombez, G. (2010). The automatic orienting of attention to goal-relevant stimuli. *Acta Psychologica, 134*, 61-69. https://doi.org/10.1016/j.actapsy.2009.12.006

Walentowska, W., Moors, A., Paul, K., & Pourtois, G. (2016). Goal relevance influences performance monitoring at the level of the FRN and P3 components. *Psychophysiology, 53*, 1020-1033. https://doi.org/10.1111/psyp.12651

Walentowska, W., Paul, K., Severo, M. C., Moors, A., & Pourtois, G. (2018). Relevance and uncertainty jointly influence reward anticipation at the level of the SPN ERP component. *International Journal of Psychophysiology, 132*, 287-297. https://doi.org/10.1016/j.iijpsycho.2017.11.005

Waraczyinski, M., & Perkins, M. (2000). Temporary inactivation of the retrorubral fields decreases the rewarding effect of medial forebrain bundle stimulation. *Brain Research, 885*, 154-165. https://doi.org/10.1016/S0006-8993(00)02908-5

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*, 1-14. https://doi.org/10.1037/h0069608

Weiskrantz, L. (1956). Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *Journal of Comparative and Physiological Psychology, 49*, 381-391. https://doi.org/10.1037/h0088009

Wieser, M. J., Miskovic, M., Rausch, S., & Keil, A. (2014). Differential time course of visuocortical signal changes to fear-conditioned faces with direct or averted gaze: A ssVEP study with single-trial analysis. *Neuropsychologia, 62*, 101-110. https://doi.org/10.1016/j.neuropsychologia.2014.07.009

Yuan, M., Giménez-Fernández, T., Méndez-Bértolo, C., & Moratti, S. (2018). Ultra-fast cortical gain adaptation in the human brain by trial-to-trial changes of associative strength in fear learning. *Journal of Neuroscience, 38*, 8262-8276. https://doi.org/10.1523/JNEUROSCI.0977-18.2018

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Aldridge, J. W. (2009). A neural computational model of incentive salience. *PLoS Computational Biology, 5*, e1000437. https://doi.org/10.1371/journal.pcbi.1000437

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology, 26*, 52-58. https://doi.org/10.1016/j.cub.2015.10.066

Zorawski, M., Cook, C. A., Kuhn, C. M., & LaBar, K. S. (2005). Sex, stress, and fear: Individual differences in conditioned learning. *Cognitive, Affective, & Behavioral Neuroscience, 5*, 191-201. https://doi.org/10.3758/CABN.5.2.191

6. RÉSUMÉ EN FRANÇAIS

INTRODUCTION & PARTIE THÉORIQUE

L'apprentissage émotionnel se réfère au processus mental par lequel un stimulus neutre acquiert une valeur émotionnelle, ou par lequel la valeur émotionnelle d'un stimulus est modifiée. Il s'agit d'une fonction adaptative essentielle permettant de prédire et détecter dans l'environnement des stimuli avec une haute importance pour l'organisme, et ainsi de préparer une réponse appropriée à ces stimuli favorisant la survie et le bien-être.

L'apprentissage émotionnel est principalement étudié par le biais du conditionnement pavlovien (Pavlov, 1927 ; Phelps, 2006). Le conditionnement pavlovien est une forme d'apprentissage associatif fondamentale, présente chez une grande variété d'espèces animales allant d'organismes simples, tels que les mouches à fruits ou les escargots marins, à des organismes plus complexes, tels que les rats ou les êtres humains (LeDoux, 1994). Dans le conditionnement pavlovien, l'organisme apprend à associer un stimulus environnemental (c.-à-d., le stimulus conditionné) avec un stimulus motivationnellement saillant (c.-à-d., le stimulus inconditionné). Après une ou plusieurs associations contingentes avec le stimulus inconditionné, le stimulus conditionné acquiert une valeur prédictive et émotionnelle, et développe la capacité de déclencher une réponse anticipatoire et préparatoire (c.-à-d., la réponse conditionnée ; Pavlov, 1927 ; Rescorla, 1988b).

L'étude du conditionnement pavlovien a grandement contribué à améliorer nos connaissances du fonctionnement et des bases neurobiologiques de l'apprentissage, de la mémoire, et des processus émotionnels, ainsi que de leurs interactions complexes (Büchel et al., 1998 ; Dunsmoor, Murty, et al., 2015 ; Dunsmoor, Niv, et al., 2015 ; LaBar & Cabeza, 2006 ; LaBar et al., 1998 ; LeDoux, 2000, 2012, 2014 ; Phelps, 2006 ; Phelps & LeDoux, 2005). En particulier, le conditionnement pavlovien aversif a permis d'identifier une partie des mécanismes psychologiques et cérébraux impliqués dans le développement, l'expression, la rétention, et la modification des réponses défensives liées à la peur, ainsi que le rôle centrale de l'amygdale dans chacune de ces composantes (p. ex., Büchel et al., 1998 ; Delgado et al., 2006 ; LaBar & Cabeza, 2006 ; LaBar et al., 1998 ; LeDoux, 2000 ; Phelps, 2006 ; Phelps et al., 2004 ; Phelps & LeDoux, 2005) et du cortex préfrontal ventromédial dans la rétention de l'apprentissage d'extinction (Phelps et al., 2004). Les processus de conditionnement pavlovien aversif sont également considérés comme étant cruciaux dans l'étiologie, l'entretien, et le traitement des troubles cliniques liés à la peur, tels que les troubles anxieux et les phobies spécifiques (Lissek et al., 2005 ; Milad & Quirk, 2012 ; Mineka & Zinbarg, 2006 ; Seligman, 1971).

De manière intéressante, tandis que le conditionnement pavlovien aversif a attiré un large intérêt dans le domaine de l'émotion, le rôle du conditionnement pavlovien appétitif a rarement été examiné de façon systématique chez l'être humain en comparaison (p. ex., Martin-Soelch et al., 2007). Cette asymétrie s'explique notamment par le fait que le conditionnement pavlovien appétitif est plus complexe à étudier chez l'humain. Il est en effet difficile de trouver des stimuli appétitifs appropriés étant capable de déclencher des réponses similairement intenses à celles déclenchées par les stimuli inconditionnés aversifs utilisés dans le conditionnement pavlovien aversif, tels que les stimulations électriques (Hermann et al., 2000 ; Martin-Soelch et al., 2007). De plus, il est possible que les mesures psychophysiologiques communément utilisées pour mesurer les réponses conditionnées appétitives ne soient pas suffisamment sensibles pour détecter de manière consistante les changements physiologiques causés par le conditionnement appétitif (Stussi, Delplanque, et al., 2018). Ainsi, le développement et la validation d'indicateurs psychophysiologiques sensibles permettant de mesurer le conditionnement pavlovien appétitif chez l'humain représente un objectif important pour remédier au manque relatif de connaissances concernant ce type d'apprentissage.

La recherche sur le conditionnement pavlovien s'est essentiellement focalisée sur l'exploration des principes généraux de l'apprentissage (Pavlov, 1927), soulignant en particulier le rôle clé de deux mécanismes computationnels dans l'apprentissage associatif : l'erreur de prédiction (c.-à-d., la différence entre la conséquence attendue et la conséquence obtenue ; p. ex., Niv & Schoenbaum, 2008 ; Rescorla & Wagner, 1972 ; Schultz et al., 1997 ; Sutton & Barto, 1998) et l'associabilité du stimulus (p. ex., Mackintosh, 1975 ; Pearce & Hall, 1980 ; voir aussi Li et al., 2011). Néanmoins, cette ligne de recherche a généralement omis de considérer l'importance relative des stimuli en jeu pour l'organisme. Alors que les premiers théoriciens de l'apprentissage ont initialement postulé que tous les stimuli peuvent être associés de manière équivalente peu importe leur nature (p. ex., Estes, 1950 ; Pavlov, 1927 ; Watson & Rayner, 1920), il a cependant été montré que certaines associations sont plus faciles à former et à maintenir que d'autres (Garcia & Koelling, 1966 ; Öhman & Mineka, 2001 ; Seligman, 1970, 1971), reflétant ainsi l'existence de biais d'apprentissage. Etonnamment, les mécanismes sous-jacents à cet apprentissage émotionnel préférentiel restent actuellement mal définis.

Les principaux modèles théoriques de l'apprentissage émotionnel, et plus spécifiquement le modèle de préparation biologique (Seligman, 1970, 1971) et la théorie du module de la peur (Öhman & Mineka, 2001), adoptent une perspective évolutionniste selon laquelle les stimuli menaçants rencontrés par l'espèce au cours de son évolution bénéficieraient

d'un apprentissage émotionnel préférentiel par rapport aux stimuli menaçants d'origine ontogénétique ou aux stimuli non menaçants. En accord avec cette perspective, un ensemble d'études empiriques a montré que les stimuli de menace ayant une origine évolutionnaire, tels que les serpents, les visages de colère, ou les visages de membres de l'hors-groupe, sont plus facilement associés à un événement aversif, et ceci de manière plus persistante, que les stimuli non menaçants, tels que les fleurs, les visages de joie, ou les visages de membres de son propre groupe social (p.ex., Atlas & Phelps, 2018 ; Ho & Lipp, 2014 ; Öhman & Dimberg, 1978 ; Öhman, Eriksson, et al., 1975 ; Öhman et al., 1976 ; Öhman & Mineka, 2001 ; Olsson et al., 2005 ; mais voir Åhs et al., 2018). D'autres études ont également suggéré que les stimuli menaçants d'origine phylogénétique pouvaient être conditionnés aversivement de manière préférentielle indépendamment de la reconnaissance explicite et consciente de ces stimuli (p. ex., Esteves, Dimberg, et al., 1994 ; Esteves, Parra, et al., 1994 ; Öhman & Soares, 1993, 1998) ou d'instructions verbales indiquant que le stimulus inconditionné aversif ne sera plus administré (p. ex., Hugdahl & Öhman, 1977). Combinant ces résultats avec la modèle de préparation biologique, Öhman et Mineka (2001) ont proposé l'existence d'un module de la peur implémenté dans le cerveau humain qui sous-tendrait l'apprentissage émotionnel préférentiel en réponse aux stimuli menaçants d'origine évolutionnaire. Ce module présenterait quatre caractéristiques fondamentales, chacune résultant de contingences évolutionnaires : il serait (a) activé sélectivement par des stimuli ayant été associés avec des conséquences menaçantes au cours de l'évolution, ceci de manière (b) automatique et (c) indépendante des processus cognitifs de haut-niveau, et aurait (d) une implémentation dédiée dans un circuit neural spécifique centré autour de l'amygdale. Ainsi, ces modèles théoriques accentuent l'importance des stimuli négatifs de menace d'origine évolutionnaire dans l'apprentissage émotionnel chez l'humain, et suggèrent que l'apprentissage préférentiel est sous-tendu par un mécanisme spécifique à la menace.

Cependant, le statut privilégié supposé des stimuli menaçants d'origine phylogénétique par rapport aux stimuli menaçants d'origine ontogénétique a été critiqué et remis en cause (voir, p. ex., Davey, 1995 ; Mallan et al., 2013). En effet, des études ont montré que des stimuli représentant une menace culturelle (c.-à-d., armes à feu) peuvent également bénéficier d'un conditionnement pavlovien aversif préférentiel étant similaire à celui obtenu en réponse à des stimuli de menace d'ordre évolutionnaire (c.-à-d., images de serpents ; Hugdahl & Johnsen, 1989 ; Flykt et al., 2007). De plus, il a été montré que la résistance du conditionnement pavlovien aversif en réponse à des stimuli sociaux menaçants est plus malléable que celle en

réponse à des stimuli animaux représentant une menace, ce qui suggère que les biais d'apprentissage engendrés par les stimuli sociaux menaçants pourraient en partie dépendre de facteurs socioculturels plutôt que de facteurs génétiques uniquement (Mallan et al., 2013 ; voir aussi Olsson et al., 2005). Dans leur ensemble, ces résultats suggèrent que le développement de biais d'apprentissage dans le conditionnement pavlovien aversif résulte de l'interaction complexe entre facteurs évolutionnaires et culturels (Davey, 1995).

En opposition avec les modèles de préparation biologique et du module de la peur, nous proposons ici un nouveau modèle théorique suggérant que l'apprentissage émotionnel préférentiel n'est pas spécifique aux stimuli de menace, mais s'étend à tous les stimuli pertinents pour les préoccupations majeures (*concerns*) de l'organisme, tels que ses besoins, ses buts, ses motifs, ses valeurs, et/ou son bien-être (Frijda, 1986, 1988). Dérivé des théories de l'évaluation cognitive (*appraisal* ; p. ex., Moors et al., 2013 ; Sander et al., 2003, 2005, 2018), ce modèle alternatif soutient que l'apprentissage émotionnel préférentiel est déterminé par un mécanisme général de détection de la pertinence plutôt qu'un mécanisme spécifique à la menace (Stussi et al., 2015, sous presse ; Stussi, Pourtois, et al., 2018). La détection de la pertinence est conceptualisée comme un processus d'évaluation rapide permettant à l'organisme d'évaluer, détecter, et établir si un stimulus rencontré dans l'environnement est pertinent pour ses préoccupations majeures (Frijda, 1986, 1988 ; Pool, Brosch, et al., 2016 ; Sander et al., 2003, 2005). Selon ce modèle, les stimuli menaçants d'origine évolutionnaire sont appris de manière préférentielle non pas parce qu'ils ont été associés avec une menace au cours de l'évolution, mais parce qu'ils sont hautement pertinents pour la survie de l'organisme. Plus spécifiquement, le modèle de détection de la pertinence prédit que les stimuli étant détectés comme pertinents pour les préoccupations majeures de l'organisme bénéficient d'un apprentissage émotionnel préférentiel indépendamment de leur valence et de leur statut évolutionnaire en soi.

Sur la base de cette hypothèse, cinq prédictions peuvent être formulées. Premièrement, des stimuli ayant divers niveaux de pertinence pour l'organisme devraient être appris de manière différentielle, les stimuli étant plus pertinents provoquant un apprentissage émotionnel plus rapide et plus persistant que des stimuli avec un niveau plus bas de pertinence. En adéquation avec cette prédiction, nous avons montré dans une première étude explorant le rôle de la détection de la pertinence dans l'apprentissage émotionnel chez l'humain que des stimuli ayant une plus grande pertinence pour soi (c.-à-d., visages de colère avec regard direct ou visages de peur avec regard dévié) peuvent entraîner un conditionnement palvovien aversif

plus rapide et/ou plus résistant à l'extinction que des stimuli ayant un pertinence pour soi plus basse (c.-à-d., visages de colère avec regard dévié ou visages de peur avec regard direct, respectivement ; Stussi et al., 2015 ; voir aussi Sander et al., 2003, 2007). Deuxièmement et de manière importante, le modèle de détection de la pertinence suggère l'existence d'un mécanisme général sous-tendant l'apprentissage préférentiel étant partagé entre les stimuli émotionnels négatifs et positifs. Ceci implique notamment – même si cela peut paraître contre-intuitif au premier abord – que les stimuli positifs ayant une haute pertinence pour l'organisme devraient également être associés à un événement aversif de manière préférentielle, comme c'est le cas pour les stimuli représentant une menace. Troisièmement, l'hypothèse de la détection de la pertinence postule un mécanisme partagé de l'apprentissage émotionnel non seulement entre les stimuli émotionnels négatifs et positifs, mais aussi entre les contingences aversives et appétitives. Les stimuli détectés comme affectivement pertinents pour l'organisme sont donc supposés bénéficier d'un apprentissage émotionnel préférentiel dans des contextes aversifs, mais également dans des contextes appétitifs. Quatrièmement, l'apprentissage émotionnel préférentiel est considéré comme s'étendant aux stimuli pertinents pour l'organisme au-delà de considérations purement biologiques et évolutionnaires. Cinquièmement, le modèle de détection de la pertinence suggère que l'apprentissage émotionnel est largement affecté par les différences individuelles, la détection de la pertinence étant spécifique à un individu dans une situation donnée. En effet, le processus de détection de la pertinence est inextricablement lié aux préoccupations majeures de l'organisme, dont la saillance et la priorité peuvent changer de manière rapide et flexible en fonction des contingences environnementales en présence (Cunningham & Brosch, 2012 ; Frijda, 1986 ; Sander et al., 2005). Le même stimulus peut donc potentiellement produire un biais d'apprentissage chez un individu donné, mais pas chez un autre, selon si ces individus diffèrent dans leurs préoccupations majeures actuelles, et ainsi dans leur façon d'évaluer la pertinence du stimulus en jeu.

Objectifs de la thèse

Dans le cadre de cette thèse, nous nous proposons ainsi d'examiner si la détection de la pertinence représente un mécanisme de base sous-tendant l'apprentissage émotionnel préférentiel chez l'être humain. Plus spécifiquement, le but central de cette thèse est de tester de manière systématique la prédiction théorique postulant que les stimuli détectés comme pertinents pour les préoccupations majeures de l'organisme bénéficient d'un conditionnement pavlovien préférentiel (c.-à-d., plus rapide et/ou plus persistant) indépendamment de leur

valence et de leur statut évolutionnaire. Dans cette optique, nous avons défini cinq objectifs principaux. Le premier et objectif primordial est de déterminer si, à l'instar des stimuli menaçants, les stimuli positifs ayant une pertinence affective pour l'organisme sont également associés préférentiellement à un événement naturellement aversif lors du conditionnement pavlovien aversif. Le deuxième objectif est de caractériser l'influence de la pertinence affective du stimulus sur le conditionnement pavlovien aversif au niveau computationnel. Deux études expérimentales ont été effectuées pour répondre à ces objectifs (études 1 et 2). Le troisième objectif consiste à évaluer si le conditionnement pavlovien aversif préférentiel s'étend aux stimuli détectés comme pertinents pour les préoccupations majeures de l'organisme au-delà de considérations biologiques et évolutionnaires (étude 3). Etant donné qu'un postulat important du modèle de la détection de la pertinence est que l'apprentissage émotionnel préférentiel varie en fonction de différences individuelles dans les préoccupations majeures de l'organisme, le quatrième objectif de la thèse est d'étudier le rôle des différences interindividuelles dans le conditionnement pavlovien aversif préférentiel (études 2 et 3). Enfin, bien que la présente thèse se focalise essentiellement sur le conditionnement pavlovien aversif, le modèle de détection de la pertinence suggère que l'implication d'un mécanisme de détection de la pertinence ne se limite pas aux contingences aversives, mais s'étend également aux contingences appétitives. Comme mentionné précédemment, les processus de conditionnement pavlovien appétitif n'ont toutefois été que rarement étudiés de manière systématique chez l'humain en comparaison aux processus de conditionnement pavlovien aversif (p. ex., Martin-Soelch et al., 2007), possiblement à cause d'un manque de mesure psychophysiologique suffisamment sensible pour permettre de détecter les changements physiologiques provoqués par le conditionnement appétitif (Stussi, Delplanque, et al., 2018). Pour ces raisons, le cinquième et objectif final de cette thèse est lié à des aspects méthodologiques, et se concentre sur le développement et la validation d'une nouvelle mesure psychophysiologique du conditionnement pavlovien appétitif chez l'humain, pouvant être par la suite utilisée dans des recherches ultérieures pour établir si le rôle de la détection de la pertinence dans l'apprentissage émotionnel se généralise à l'apprentissage pavlovien appétitif (étude 4).

PARTIE EMPIRIQUE

Dans la partie empirique de cette thèse, nous rapportons quatre études ayant été menées afin de répondre aux différents objectifs mentionnés ci-dessus, et ainsi établir et caractériser le rôle de la détection de la pertinence dans l'apprentissage émotionnel chez l'humain.

Etudes 1 et 2

Dans l'étude 1, nous avons examiné dans une série de trois expériences si des stimuli menaçants et des stimuli positifs ayant une pertinence biologique pour l'organisme sont associés de manière préférentielle à une conséquence aversive par rapport à des stimuli neutres présentant une pertinence moindre. Pour ce faire, nous avons manipulé la valence et la pertinence de stimuli conditionnés dans un paradigme de conditionnement pavlovien aversif différentiel en utilisant trois catégories distinctes de stimuli : (a) des stimuli biologiquement pertinents et négatifs (c.-à-d., visages de colère dans les expériences 1.1 et 1.2, images de serpent dans l'expérience 1.3), (b) des stimuli biologiquement pertinents et positifs (c.-à-d., visages de bébé dans les expériences 1.1 et 1.2, images érotiques dans l'expérience 1.3), et (c) des stimuli neutres ayant un plus faible niveau de pertinence (c.-à-d., visages neutres dans les expériences 1.1 et 1.2, carrés de couleur dans l'expérience 1.3). Dans chacune des expériences (N = 40 pour les expériences 1.1 et 1.3, N = 60 pour l'expérience 1.2), le paradigme de conditionnement pavlovien aversif différentiel comprenait trois phases contiguës. Pendant la phase initiale d'habituation, tous les stimuli conditionnés étaient présentés sans être renforcés. Dans la phase subséquente d'acquisition, un stimulus (stimulus renforcé, SC+) de chaque catégorie de stimulus conditionné était systématiquement associé de manière contingente à une stimulation électrique (stimulus inconditionné) selon un programme de renforcement partiel, tandis que l'autre stimulus (stimulus non renforcé, SC-) de chaque catégorie n'était jamais associé à la stimulation électrique. Dans la phase suivante d'extinction, la stimulation électrique n'était plus administrée. La réponse électrodermale des participants était mesurée en continu durant toutes les phases du conditionnement. La réponse conditionnée a été opérationnalisée comme la réponse électrodermale au SC+ moins la réponse électrodermale au SC- de la même catégorie de stimulus (voir, p. ex., Olsson et al., 2005) et utilisée comme indicateur de l'apprentissage.

Au travers des trois expériences, les résultats ont révélé une persistance accrue de la réponse conditionnée à la fois aux stimuli menaçants et aux stimuli positifs biologiquement pertinents en comparaison à la réponse conditionnée aux stimuli neutres. Ceci démontre que le conditionnement pavlovien aversif préférentiel n'est pas limité aux stimuli de menace, mais peut également se produire en réponse à des stimuli positifs étant biologiquement pertinents pour l'organisme. Ces résultats suggèrent l'existence d'un mécanisme général sous-jacent à l'apprentissage pavlovien aversif préférentiel qui est partagé entre les stimuli négatifs et positifs ayant une haute pertinence biologique pour l'organisme.

De manière similaire, l'étude 2 visait à déterminer (a) si, comme pour les visages de colère, un conditionnement pavlovien aversif préférentiel pouvait être observé en réponse aux visages de joie par rapport aux visages neutres en utilisant un échantillon plus important (N = 107) que ceux typiquement utilisés dans la littérature sur le conditionnement pavlovien chez l'humain (N = 15-25), et (b) si l'émergence de tels biais d'apprentissage en réponse aux visages de joie est influencée par des différences interindividuelles dans l'évaluation affective de ces stimuli. A cet effet, nous avons dans un premier temps mesurer les différences interindividuelles en termes d'extraversion. Les individus extravertis sont en effet généralement caractérisés par une haute sociabilité, une haute sensibilité à la récompense, ainsi qu'un affect positif élevé (p. ex., Ashton et al., 2002 ; Lucas et al., 2002 ; Smillie, 2013), et montrent une réactivité amygdalienne augmentée en réponse aux visages de joie (Canli et al., 2002). En conséquence, il a été suggéré que ces individus évaluent les visages de joie comme étant plus pertinents pour leurs préoccupations que les individus ayant un plus bas niveau d'extraversion (Sander et al., 2003, 2005). Nous avons de plus mesuré les différences interindividuelles au niveau des associations implicites entre les différents types de visages présentés (colère vs. joie vs. neutres) et l'attribut d'importance (vs. de non-importance) au moyen d'une tâche associative de *Go/No-go* (Nosek & Banaji, 2001), afin d'estimer indirectement l'évaluation affective de ces visages. Un paradigme de conditionnement pavlovien aversif différentiel similaire à celui utilisé dans l'étude 1 a par la suite été implémenté. Dans ce paradigme, des visages de colère, de joie, et neutres ont été utilisés comme stimuli conditionnés. La réponse conditionnée a été définie comme la réponse électrodermale au SC+ moins la réponse électrodermale au SC- de la même catégorie de visage (voir, p. ex., Olsson et al., 2005).

Les résultats ont montré qu'à la fois les visages de colère et les visages de joie ont entrainé une acquisition plus rapide et une résistance à l'extinction plus élevée de la réponse conditionnée par rapport aux visages neutres. Alors que les effets de conditionnement aversif plus rapide en réponse aux visages de colère et de joie étaient de taille modérée, la résistance à l'extinction accrue pour les visages de joie était de taille relativement petite et d'une magnitude moindre que celle pour les visages de colère, ce qui semble cohérent avec la notion que les visages de joie ont un niveau général de pertinence pour l'organisme plus faible que les visages de colère au niveau de groupe (Brosch et al., 2008, 2010 ; Pool, Brosch, et al., 2016). Cet effet était en outre modulé par des différences interindividuelles dans l'évaluation affective des visages de joie, comme indiqué par une plus grande persistance de la réponse conditionnée

chez les individus étant plus rapides pour associer ces visages à l'attribut d'importance que celui de non-importance. A l'inverse, la persistance de la réponse conditionnée aux visages de colère ou neutres n'était pas influencée par des différences interindividuelles dans leur évaluation. Ainsi, les résultats de l'étude 2 s'alignent avec ceux de l'étude 1 en suggérant que la présence de biais d'apprentissage dans le conditionnement pavlovien aversif ne se restreint pas aux stimuli menaçants, mais s'étend plus généralement aux stimuli pertinents pour l'organisme sur le plan affectif.

Des analyses supplémentaires effectuées dans les études 1 et 2 au moyen de modèles d'apprentissage par renforcement (Li et al., 2011 ; Pearce & Hall, 1980 ; Rescorla & Wagner, 1972) ont par ailleurs suggéré que l'influence des stimuli négatifs et positifs ayant une haute pertinence affective pour l'organisme, par rapport aux stimuli neutres, pouvait être caractérisée par un taux d'apprentissage diminué pour l'erreur de prédiction négative (c.-à-d., quand la conséquence obtenue est plus faible que celle prédite ou omise). Dans l'étude 1, les visages de colère et les visages de bébé étaient associés à un taux d'apprentissage relatif à l'erreur de prédiction négative plus faible que les visages neutres. Dans l'étude 2, le taux d'apprentissage estimé en lien avec l'erreur de prédiction négative pour les visages de colère était inférieur à celui estimé pour les visages de joie et les visages neutres, et celui pour les visages de joie était plus faible que celui pour les visages neutres, bien que cette dernière différence n'était que tendancielle après correction pour comparaisons multiples. Etant donné que les taux d'apprentissage modulent l'influence de l'erreur de prédiction sur le conditionnement pavlovien, ces résultats suggèrent que les stimuli affectivement pertinents pourraient biaiser l'apprentissage inhibiteur sous-tendant l'extinction (Dunsmoor, Niv, et al., 2015) à travers un impact diminué de l'erreur de prédiction négative, produisant ainsi une résistance à l'extinction augmentée de la réponse conditionnée à ces stimuli. En revanche, nous n'avons pas trouvé de soutien empirique indiquant que le conditionnement pavlovien aversif plus rapide en réponse aux visages de colère et de joie était sous-tendu par un taux d'apprentissage pour l'erreur de prédiction positive (c.-à-d., quand la conséquence obtenue est supérieure à celle prédite ou n'est pas attendue) associé à ces stimuli plus élevé que pour les visages neutres.

Etude 3

La troisième étude avait pour objectif d'examiner (a) si le conditionnement pavlovien aversif préférentiel peut se produire en réponse à des stimuli étant pertinents pour les préoccupations majeures de l'organisme au-delà de considérations biologiques et évolutionnaires, et (b) si un tel apprentissage préférentiel est modulé par des différences

interindividuelles de motivation. Dans cette étude, les participants ($N = 72$) ont d'abord réalisé une tâche attentionnelle d'indiçage spatial, dans laquelle la pertinence aux buts de stimuli initialement neutres était manipulée expérimentalement, certains stimuli étant pertinents pour les buts de la tâche (*goal-relevant*) et d'autres étant non pertinents (*goal-irrelevant*). Ils ont ensuite été soumis à un paradigme de conditionnement pavlovien aversif différentiel, dans lequel les stimuli pertinents pour les buts et les stimuli non pertinents ont servi de stimuli conditionnés. La réponse électrodermale des participants a été enregistrée en tant qu'index de la réponse conditionnée (c.-à-d., la réponse électrodermale au SC+ moins celle au SC- pour chaque catégorie de stimulus séparément ; Olsson et al., 2005), et leur motivation d'accomplissement a été mesurée pour examiner son influence sur le conditionnement en réponse aux différentes catégories de stimulus.

Les résultats ont montré que la motivation d'accomplissement modulait l'apprentissage pavlovien aversif en réponse aux stimuli pertinents pour les buts versus non pertinents. Les participants ayant une motivation d'accomplissement élevée ont acquis une réponse conditionnée plus rapidement pour les stimuli pertinents pour les buts que pour les stimuli non pertinents, ce qui n'était pas le cas pour les individus ayant un plus bas niveau de motivation d'accomplissement. Toutefois, aucune différence n'a été observée entre les stimuli pertinents pour les buts et les stimuli non pertinents dans la phase d'extinction. Ces résultats suggèrent que les stimuli étant détectés comme pertinents pour l'organisme peuvent provoquer un conditionnement pavlovien aversif facilité, même s'ils ne possèdent pas de valeur menaçante intrinsèque ou de signification évolutionnaire, et que l'apparition d'un tel biais d'apprentissage dépend de manière critique de différences interindividuelles dans les préoccupations majeures de l'organisme, telles que sa motivation d'accomplissement.

Etude 4

La quatrième étude de cette thèse visait à tester et valider si le réflexe postauriculaire – un micro-réflexe musculaire vestigial chez l'humain qui est spécifiquement potentialisé par des stimuli plaisants par rapport à des stimuli déplaisants ou neutres (p. ex., Benning et al., 2004 ; Gable & Harmon-Jones, 2009 ; Hackley et al., 2009 ; Sandt et al., 2009) – représente une mesure psychophysiologique valide du conditionnement pavlovien appétitif humain. A cette fin, nous avons implémenté un paradigme de conditionnement pavlovien appétitif différentiel, dans laquelle un stimulus neutre (SC+) était associé de manière contingente à une odeur plaisante (stimulus inconditionné), tandis qu'un autre stimulus neutre (SC-) n'était jamais

associé à l'odeur plaisante. Nous avons mesuré le réflexe postauriculaire, ainsi que le réflexe de clignement et la réponse électrodermale comme indicateurs de l'apprentissage.

Les résultats de l'étude 4 ont indiqué que le réflexe postauriculaire était potentialisé lors de la présentation du SC+ en comparaison à celle du SC- pendant la phase d'acquisition. Cette potentialisation n'était cependant plus présente lorsque l'odeur plaisante n'était plus administrée durant la phase d'extinction. En revanche, nous n'avons pas trouvé d'atténuation du réflexe de clignement pendant la présentation du SC+ par rapport au SC-, et aucun effet du conditionnement appétitif n'a été observé au niveau de la réponse électrodermale. Ces résultats suggèrent donc que le réflexe postauriculaire est une mesure sensible du conditionnement pavlovien appétitif chez l'humain.

DISCUSSION GÉNÉRALE

Dans cette thèse, nous avons cherché à établir si la détection de la pertinence est un déterminant clé de l'apprentissage émotionnel chez l'humain. De manière plus spécifique, nous avons testé à travers différentes études empiriques la prédiction théorique dérivée des théories de l'évaluation cognitive (p. ex., Sander et al., 2003, 2005, 2018) selon laquelle les stimuli détectés comme pertinents pour les préoccupations majeures de l'organisme sont appris préférentiellement durant le conditionnement pavlovien, indépendamment de leur valence et de leur histoire évolutionnaire en soi (Stussi et al., 2015, sous presse ; Stussi, Pourtois, et al., 2018).

En lien avec le premier objectif de la thèse, les études 1 et 2 ont montré que, de manière similaire aux stimuli de menace, les stimuli positifs ayant une pertinence affective pour l'organisme sont également associés préférentiellement à un événement naturellement aversif pendant le conditionnement pavlovien aversif. Dans l'étude 1, ces biais d'apprentissage se sont traduits par une augmentation de la résistance à l'extinction de la réponse conditionnée à la fois aux stimuli négatifs menaçants (visages de colère ou images de serpent) et aux stimuli positifs étant biologiquement pertinents pour l'organisme (visages de bébé ou images érotiques) par rapport à la réponse conditionnée aux stimuli neutres et moins pertinents (visages neutres ou carrés de couleur). Dans l'étude 2, un tel apprentissage préférentiel a été mis en évidence par une acquisition facilitée et une persistance accrue de la réponse conditionnée à la fois à des stimuli sociaux liés à la menace (visages de colère) et des stimuli sociaux positifs (visages de joie) en comparaison à des stimuli neutres (visages neutres). En adéquation avec le modèle de

la détection de la pertinence, les études 1 et 2 démontrent donc de manière critique que le conditionnement pavlovien aversif préférentiel ne se limite pas aux stimuli négatifs de menace, mais s'étend aussi aux stimuli positifs étant affectivement pertinents pour l'organisme. Ces résultats, bien que quelque peu contre-intuitifs, corroborent ainsi l'hypothèse selon laquelle les stimuli pertinents sur le plan affectif sont appris de manière préférentielle au cours du conditionnement pavlovien, peu importe leur valence intrinsèque.

Concernant le deuxième objectif de la thèse, les études 1 et 2 ont également indiqué que l'influence de la pertinence affective du stimulus sur le conditionnement pavlovien aversif pouvait se caractériser par un taux d'apprentissage relatif à l'erreur de prédiction négative plus faible. Un tel taux d'apprentissage diminué contribue vraisemblablement à affaiblir l'apprentissage inhibiteur sous-tendant l'extinction (Dunsmoor, Niv, et al., 2015) en réponse aux stimuli étant détectés comme pertinents pour l'organisme, augmentant ainsi la persistance de la réponse conditionnée à ces stimuli par rapport à des stimuli moins pertinents. Globalement, ces résultats fournissent un premier aperçu des mécanismes computationnels par lesquels l'influence de la pertinence affective sur le conditionnement pavlovien aversif opère et contribuent à caractériser le rôle de la détection de la pertinence dans l'apprentissage émotionnel. Etant donné que ces études ne constituent qu'une première tentative pour élucider l'impact de la pertinence affective du stimulus au niveau computationnel, il convient néanmoins de noter que ces mécanismes computationnels restent encore à spécifier plus précisément dans des recherches futures.

Dans l'étude 3, nous avons en outre montré que des stimuli initialement neutres étant devenus pertinents pour les buts des participants étaient plus facilement appris lors du conditionnement pavlovien aversif que des stimuli non pertinents chez les individus ayant une motivation d'accomplissement élevée, ce qui n'était pas le cas pour les individus ayant une motivation d'accomplissement plus basse. Ces résultats indiquent que les stimuli temporairement associés à une haute pertinence aux buts peuvent produire un conditionnement pavlovien aversif facilité même s'ils ne possèdent pas de valeur menaçante préexistante, à condition que les dispositions motivationnelles individuelles soient respectées simultanément. En accord avec les prédictions du modèle de détection de la pertinence, ceci démontre que les stimuli pertinents pour l'organisme au-delà de considérations biologiques et évolutionnaires peuvent aussi bénéficier d'un apprentissage pavlovien aversif accéléré, répondant ainsi au troisième objectif de cette thèse.

Les études 2 et 3 ont de plus mis en évidence le rôle central des différences interindividuelles en termes de préoccupations majeures de l'organisme dans le conditionnement pavlovien aversif préférentiel, conformément au quatrième objectif de ce travail. Les résultats de l'étude 2 ont en effet montré que la réponse conditionnée aux visages de joie durant la phase d'extinction était modulée par des différences interindividuelles dans l'évaluation affective de ces visages, les individus associant plus rapidement les visages de joie à l'attribut d'importance (vs. de non-importance) exhibant une plus grande résistance à l'extinction pour ces visages. Dans la même lignée, l'étude 3 a souligné que le développement de biais d'apprentissage dans le conditionnement pavlovien aversif repose de manière cruciale sur les différences interindividuelles dans l'affect et la motivation.

Finalement, la quatrième et dernière étude présentée dans la partie empirique a permis de répondre au cinquième objectif de la thèse en montrant que le réflexe postauriculaire représente un indicateur psychophysiologique valide du conditionnement pavlovien appétitif chez l'être humain. De surcroît, cette étude suggère que le réflexe postauriculaire est une mesure psychophysiologique étant probablement plus sensible que le réflexe de clignement et la réponse électrodermale, deux des mesures psychophysiologiques étant les plus utilisées comme indicateurs du conditionnement appétitif (p. ex., Andreatta & Pauli, 2015). Dans cette perspective, le réflexe postauriculaire constitue un outil prometteur pour tester de manière systématique dans des études ultérieures si le rôle de la détection de la pertinence dans l'apprentissage émotionnel se généralise aussi à l'apprentissage pavlovien appétitif et ne se limite pas uniquement aux contingences aversives. En particulier, ce réflèxe pourrait permettre de tester l'hypothèse soutenue par le modèle de détection de la pertinence selon laquelle à la fois les stimuli négatifs et positifs étant pertinents pour l'organisme seraient plus facilement associés à une conséquence appétitive, et ceci de manière plus persistante, au cours du conditionnement pavlovien appétitif.

En résumé, l'ensemble des expériences menées dans le cadre de cette thèse remet en question l'idée selon laquelle les stimuli menaçants bénéficient d'un conditionnement pavlovien aversif plus rapide et plus persistant parce qu'ils ont été associés à des menaces au cours de l'évolution de l'espèce. De manière alternative, ces études suggèrent que le déterminant clé sous-tendant l'apprentissage pavlovien aversif préférentiel chez l'humain correspond à un processus de détection de la pertinence plutôt qu'à un mécanisme spécifique à la menace comme on le pensait auparavant. Nos résultats supportent donc l'existence d'un mécanisme général qui est partagé entre les stimuli pertinents pour les préoccupations majeures

de l'organisme, facilitant et renforçant l'apprentissage émotionnel. Ce mécanisme semble particulièrement fonctionnel, car il priorise l'apprentissage des stimuli pertinents en fonction de motivations individuelles spécifiques par le biais d'un apprentissage accéléré et plus persistant, aidant ainsi l'organisme à préparer de manière flexible des réponses appropriées à ces stimuli, ainsi qu'à interagir avec son environnement et s'y adapter.

Implications théoriques

Les résultats empiriques présentés dans cette thèse ont d'importantes implications pour la modélisation théorique des déterminants de l'apprentissage émotionnel préférentiel chez l'humain. En premier lieu, ils préconisent un changement de perspective dans la conceptualisation de mécanismes fondamentaux sous-jacents à l'apprentissage pavlovien aversif préférentiel. Alors qu'il a longtemps été postulé qu'uniquement les stimuli qui ont représenté une menace pour la survie de l'espèce au cours de son évolution sont appris de manière préférentielle lors du conditionnement pavlovien aversif (Öhman & Mineka, 2001 ; Seligman, 1971), nos études révèlent que l'apprentissage pavlovien aversif préférentiel n'est pas restreint aux stimuli menaçants d'origine phylogénétique, mais s'étend plus généralement aux stimuli étant détectés comme pertinents pour les préoccupations majeures de l'organisme indépendamment de leur valence et de leur histoire évolutionnaire, et ceci malgré le fait qu'ils ne possèdent aucune valeur de menace inhérente. De plus, nos résultats soulignent également l'impact central de l'état motivationnel de l'individu sur l'apprentissage émotionnel préférentiel, ainsi que l'importance clé des différences interindividuelles dans la formation et le maintien des biais d'apprentissage. A cet égard, le modèle de détection de la pertinence proposé dans le cadre de ce travail fournit un modèle théorique valide permettant d'offrir une explication mécanistique du rôle des préoccupations majeures de l'organisme et des différences individuelles dans ces construits motivationnels au sein de l'apprentissage émotionnel, contribuant ainsi à mieux comprendre les déterminants de ce processus d'apprentissage.

En deuxième lieu, cette thèse apporte des pistes potentielles pour l'élaboration d'un modèle computationnel révisé de l'apprentissage pavlovien incorporant des taux d'apprentissage distincts pour les erreurs de prédiction positive et négative, ce qui pourrait permettre une modélisation plus adéquate de l'existence de biais d'apprentissage dans le conditionnement pavlovien en réponse à des stimuli spécifiques. Nos données suggèrent notamment que ces taux d'apprentissage devraient être affectés par un paramètre computationnel reflétant la pertinence affective du stimulus pour l'organisme. Etant donné que la pertinence affective d'un stimulus donné est établie en fonction de l'interaction entre le

stimulus et les préoccupations actuelles de l'organisme (p. ex., Sander et al., 2005, 2018), ce paramètre devrait être modulé dynamiquement par l'état motivationnel actuel (ou futur) de l'individu, permettant ainsi à la pertinence affective du stimulus de varier de manière flexible en fonction de celui-ci. Il est toutefois important de mentionner que l'implémentation de telles computations pavloviennes demeure à développer et à tester systématiquement d'un point de vue algorithmique (Dayan & Berridge, 2014).

En outre, les résultats rapportés dans cette thèse sont congruents avec des données convergentes de plusieurs recherches antérieures en sciences affectives soutenant l'hypothèse des théories de l'évaluation cognitive (p. ex., Moors et al., 2013 ; Sander et al., 2003, 2005, 2018) selon laquelle la détection de la pertinence est un mécanisme essentiel dans le déclenchement des émotions, ainsi que dans la modulation des processus cognitifs, tels que l'attention (Brosch et al., 2008 ; Pool, Brosch, et al., 2014, 2016) et la mémoire (Montagrin et al., 2013, 2018). Le présent travail appuie donc les prédictions des théories de l'évaluation cognitive par rapport à celles d'autres familles majeures de théories de l'émotion – comme les théories des émotions de base (p. ex., Ekman, 1972, 1992, 1999 ; Panksepp, 1998) ou les théories dimensionnelles (p. ex., Lindquist et al., 2012 ; Russell, 2003 ; Russell & Barrett, 1999) – en lien avec les mécanismes sous-jacents à la modulation émotionnelle des processus cognitifs. Il montre également en quoi l'approche suggérée par les théories de l'évaluation peut être particulièrement fructueuse pour étudier les relations entres les émotions et d'autres processus psychologiques.

Etant donné que les déficiences d'apprentissage émotionnel sont considérées comme jouant un rôle crucial dans l'étiologie et la persistance de troubles affectifs spécifiques, tels que les troubles de l'anxiété, les phobies, ou les troubles addictifs (voir, p. ex., Duits et al., 2015 ; Lissek et al., 2005 ; Martin-Soelch et al., 2007 ; Milad & Quirk, 2012 ; Mineka & Zinbarg, 2006 ; Seligman, 1971), le modèle de la détection de la pertinence proposé dans cette thèse pourrait également s'avérer bénéfique dans la conceptualisation de ces déficiences et de leurs déterminants. Par exemple, la théorie du module de la peur (Öhman & Mineka, 2001) postule que les phobies ou peurs intenses surviennent de l'activation automatique préférentielle du module de la peur, qui est conjointement déterminée par un processus de préparation biologique et les expériences aversives antérieures dans des situations identiques ou similaires indépendamment des processus cognitifs (Mineka & Öhman, 2002). En comparaison, le modèle de la détection de la pertinence suggère que les altérations d'apprentissage émotionnel découlent plutôt de biais d'évaluation de la pertinence du stimulus ou de la situation en relation

avec les préoccupations majeures de l'organisme. En d'autres termes, cette approche théorique propose que des évaluations de pertinence contradictoires, inadéquates, ou involontaires pourraient être au cœur des dysfonctionnements d'apprentissage émotionnel et conduire à des réponses émotionnelles apparemment irrationnelles et inadaptées. Ainsi, le modèle de détection de la pertinence pourrait contribuer à une meilleure compréhension de la nature, de l'étendue, et des déterminants des déficiences d'apprentissage émotionnel associés à certains troubles affectifs dans les recherches futures, et aider au développement et à la validation de nouvelles interventions thérapeutiques individualisées pour ces troubles (voir aussi Lonsdorf & Merz, 2017).

Limites & perspectives

Le travail développé dans cette thèse présente cependant plusieurs limites, ces dernières ouvrant sur de nouvelles perspectives indiquant où de nouveaux efforts devront être effectués. Une limite importante réside notamment dans la mesure de la pertinence affective du stimulus. En effet, du fait que la pertinence affective est mieux manipulée que mesurée, nous avons préféré fonder nos expériences sur la manipulation de la pertinence affective du stimulus en fonction de bases théoriques solides et de résultats empiriques antérieurs, plutôt que sur sa mesure. Le concept de pertinence affective est difficile à mesurer de manière adéquate à un niveau quantitatif et un instrument validé et approprié permettant de fournir un indicateur fiable, sensible, et valide de la pertinence affective du stimulus fait toujours défaut. Par conséquent, le développement et la validation de tels instruments représenteraient idéalement une perspective future importante pour offrir une évaluation plus fine du rôle de la détection de la pertinence dans l'apprentissage émotionnel, ainsi que de l'influence des différences individuelles dans ce processus.

Bien qu'une hypothèse centrale du modèle de la détection de la pertinence soit que les stimuli détectés comme pertinents pour l'organisme sont associés de manière préférentielle à la fois avec des conséquences aversives et des conséquences appétitives durant le conditionnement pavlovien, nous n'avons qu'examiné et testé systématiquement le rôle de la détection de la pertinence dans des contextes aversifs. Nous ne pouvons donc pas être sûrs que nos résultats peuvent se généraliser au conditionnement pavlovien appétitif, et de plus amples recherches sont nécessaires pour établir si l'apprentissage émotionnel préférentiel en réponse aux stimuli pertinents sur le plan affectif se limite aux contingences aversives ou s'étend également aux contingences appétitives. Dans cette optique, l'étude 4 pourrait fournir une base sur laquelle des études futures pourraient être conçues pour tester la généralité d'un mécanisme

de détection de la pertinence dans le conditionnement pavlovien appétitif en utilisant le réflexe postauriculaire comme indice psychophysiologique de la réponse conditionnée.

Dans la partie empirique de cette thèse, nous nous sommes focalisés principalement sur le rôle de la détection de la pertinence dans le conditionnement pavlovien (aversif). Or, l'apprentissage émotionnel ne se limite pas au conditionnement pavlovien, mais englobe d'autres formes fondamentales d'apprentissage, comme le conditionnement instrumental par exemple. Etant donné qu'il a été montré que les biais d'apprentissage pavlovien provoqués par des stimuli menaçants (c.-à-d., images de serpent, visages de colère, visages de l'hors-groupe, armes à feu) ont une forte influence sur le comportement instrumental adaptatif (Lindström et al., 2015), une piste importante et intéressante pour la recherche future sera d'étudier si et comment les biais d'apprentissage pavlovien induits par les stimuli pertinents pour l'organisme sur le plan affectif peuvent influencer la prise de décision et le comportement instrumental indépendamment de leur valence, ainsi que de caractériser cette influence.

Une dernière limitation de ce travail se situe dans l'absence d'investigation des soubassements cérébraux sous-tendant l'impact de la détection de la pertinence sur l'apprentissage émotionnel préférentiel. En conséquence, le fait de savoir si les biais d'apprentissage en réponse aux stimuli négatifs et positifs étant pertinents pour l'organisme sont sous-tendus par des mécanismes neuraux partagés ou des structures cérébrales strictement distinctes reste à établir. Dans cette optique, un effort bénéfique pourrait être de combiner l'utilisation d'un paradigme de conditionnement pavlovien aversif comparant l'apprentissage des stimuli pertinents relatifs à la menace (p. ex., visages de colère) et positifs (p.ex., visages de bébé) à des stimuli neutres et moins pertinents (p. ex., visages neutres) avec des méthodes de neuroimagerie fonctionnelle et de modélisation computationnelle (p. ex., Boll et al., 2013 ; Li et al., 2011 ; O'Doherty et al., 2003, 2007). Cette approche multimodale semble être particulièrement prometteuse pour déterminer si le conditionnement pavlovien aversif préférentiel repose sur des structures ou des circuits cérébraux communs à la fois pour les stimuli négatifs et positifs ayant une pertinence affective pour l'organisme, et pour fournir des indications sur comment cet apprentissage émotionnel préférentiel est implémenté dans le cerveau humain. Cet axe de recherche pourrait ainsi aider à délimiter plus précisément les mécanismes psychologiques, computationnels, et cérébraux responsables de l'attribution d'une valeur émotionnelle prédictive et distinctive à des stimuli et comportements spécifiques, ainsi que le rôle de la détection de la pertinence dans ce processus, contribuant dès lors à une compréhension affinée de l'apprentissage émotionnel chez l'humain.

Conclusion

En conclusion, la présente thèse suggère que le conditionnement pavlovien aversif préférentiel est déterminé par un mécanisme général de détection de la pertinence qui n'est pas spécifique à la menace, contribuant ainsi à établir et caractériser son rôle dans l'apprentissage émotionnel. La détection de la pertinence représente un mécanisme flexible et adaptatif permettant à l'organisme de déclencher de façon rapide et dynamique un apprentissage préférentiel des stimuli environnementaux étant pertinents pour ses préoccupations majeures actuelles. Le modèle de détection de la pertinence permet notamment d'expliquer et de réinterpréter les résultats empiriques disponibles dans la littérature sur le conditionnement pavlovien humain démontrant un conditionnement pavlovien aversif préférentiel en réponse aux stimuli de menace, étant donné que ces stimuli sont hautement pertinents pour la survie de l'organisme, qui constitue une des préoccupations majeures les plus saillantes. En définitive, ce modèle pourrait également contribuer à expliquer la flexibilité et l'existence d'importantes différences interindividuelles observées dans l'apprentissage émotionnel au travers de différents contextes et situations, ainsi que les déficiences dans ce processus d'apprentissage qui précèdent ou suivent typiquement la survenue et la persistance de troubles affectifs ou émotionnels spécifiques, tels que les troubles de l'anxiété, les phobies, ou les addictions. Bien que le rôle de la détection de la pertinence reste à être déterminé dans l'apprentissage appétitif et au niveau cérébral, le cadre théorique de la détection de la pertinence – et plus généralement les théories de l'évaluation cognitive – offre une approche prometteuse pour promouvoir l'acquisition de nouvelles connaissances sur les mécanismes fondamentaux sous-tendant l'apprentissage émotionnel.