Unravelling the Neurocognitive Mechanisms Underlying Counter-Conditioning

Jette de Vos[1,2]
Supervisors: Maxime Houtekamer[1,2], Dr. Marijn Kroes[1,2]

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands
[2]Radboud University Medical Centre, Donders Institute for Brain, Cognition and Behaviour, The Netherlands

Stress- and anxiety-related disorders are devastating for patients and a major burden on society. Although effective, a robust number of patients does not improve or relapse following exposure-based treatments. This highlights the need to find alternatives, like counter-conditioning. In counter-conditioning, the aversive outcome is replaced by a rewarding outcome instead of merely omitting the aversive outcome. Here we aim to unravel the effectiveness and neurocognitive mechanisms underlying counter-conditioning. We hypothesize that counter-conditioning either 1) enhances extinction learning, 2) integrates a fear and reward memory, or 3) overwrites the fear memory by a reward memory. We tested this in a two-day between-subjects study. On day 1, participants were fear conditioned to one of the two stimulus categories (animal vs. object pictures). Next, half of the participants underwent counter-conditioning (N = 10) during which novel trial-unique exemplars from the fear conditioned category were paired with a reward using monetary incentive delay. The other participants (N = 10) received standard extinction training. On day 2, return of fear responses and episodic memory for items from conditioning and counter-conditioning/extinction were tested. Results (pupil dilation, skin conductance (SCR), heart rate, reaction times, episodic memory, fMRI (functional magnetic resonance imaging) indicate feasibility of the design to address our hypothesis that one of the three above mentioned mechanisms supports the effectiveness of counter-conditioning. Moreover, category-specific conditioned SCRs seem to reduce faster during counter-conditioning than extinction, supporting the idea of counter-conditioning as a promising alternative to extinction. In the counter-conditioning group, episodic memory is enhanced for emotionally salient items from both the conditioning and counter-conditioning phase, which is in line with the hypothesis that counter-conditioning integrates a fear and reward memory.

Keywords: fear conditioning, extinction learning, counter-conditioning, fMRI

Corresponding author: Jette de Vos; E-mail: jette-de-vos@hotmail.com

Learning what to fear is a very valuable ability in everyday life. Imagine, if a dog bites you while walking through the park, a fear memory is formed of the association between the park and the possibility of getting bitten. This association will help you to be more cautious in the park and therefore to avoid getting bitten another time. If the dog were to move out of the neighbourhood, however, it would also be helpful to learn that the park is safe again and that you can have a relaxed walk in it in the future. Both learning aspects are important: the learning of the fear association and, if situations change as in the example above, the learning of the safety association. In practice, however, the first aspect, the fear association, turns out to be more persistent than the second aspect, the safety association. Clinical disorders such as specific phobias are examples of instances of difficult-to-eradicate fear associations (Vervliet, Craske, & Hermans, 2013). Disorders such as these are mostly treated with exposure-based therapies. However, a considerable number of patients who undergo treatment does not benefit. In addition, a large number of patients relapses after treatment and the fear association returns (Craske & Mystkowski, 2006), creating the need for finding alternative methods that are more effective in preventing the return of the fear association. Counter-conditioning may be such a promising method as supported by results from previous studies using counter-conditioning (Kerkhof, Vansteenwegen, Baeyens, & Hermans, 2011; Kaag et al., 2016; Karel et al., 2019). However, why it might be more effective remains unknown.

Conditioning and extinction

In laboratory settings, fear learning is mostly studied with Pavlovian classical conditioning paradigms (Pavlov, 1927). In classical conditioning, an intrinsically neutral stimulus (the conditioned stimulus, CS), for example a sound, is coupled with an aversive or appetitive stimulus (the unconditioned stimulus, US), for example an electrical shock. After (repeated) coupling of this CS and US, the CS gets to serve as a predictor of the aversive outcome (i.e., the shock). As a consequence, the CS, which initially did not evoke a response by itself, can evoke a response (the conditioned response, CR) now. The CR can be the same as the response evoked by the US (the unconditioned response, UR), although not necessarily. Conditioned associations can be both formed with aversive and appetitive unconditioned stimuli. Therefore, the conditioning paradigm can be used to study either memories with a positive or negative associative value, depending on the valence of the US. As we will focus on fear conditioning in the current study, most examples used in the rest of this paper will address aversive rather than appetitive conditioning. The stimulus that is coupled with the US is referred to as the CS+; Stimuli in the same paradigm that are not coupled with the US are referred to as CS-.

Research investigating the process of learning that a CS no longer predicts a threat predominantly consists of studies implementing an extinction paradigm. Extinction paradigms aim to reduce conditioned fear responses. During extinction, the CS is repeatedly presented again, but now without the coupling with the US (Pavlov, 1927). The extinction paradigm has been found to be successful in reducing the CR (for a review, see Vervliet et al., 2013). In fear conditioning studies, repeated exposure to the CS without the aversive US does indeed initially reduce multiple aspects of the conditioned fear response (e.g., reduced startle-response and reduced skin conductance response to the CS). Unfortunately, the extinction procedure is less effective in preventing the return of the CR over longer periods (Vervliet et al., 2013). First, if the response to the CS is tested again after some time has passed, the CR is often found to recover. This phenomenon is described as spontaneous recovery (Quirk, 2002). Second, if the subject is confronted with presentations of the US, the response to the CS tends to return; The CR is reinstated with respect to the CS (Rescorla & Heth, 1975). Third, if the response to the CS is tested in a context other than the context in which the extinction procedure took place, again, the CR tends to recover. This phenomenon is known as renewal (Bouton, 2002). In sum, these three phenomena point out the shortcoming of the extinction procedure, namely, that fear may return even after initially successful extinction. This highlights the need for more effective alternatives to induce safety learning (Vervliet et al., 2013).

Several theories aim to explain what mechanisms underlie extinction and thereby aim to explain why extinction is unsuccessful in preventing the return of the CR. One, still widely accepted theory, is that of Bouton (1993) which states that extinction is not the unlearning of the conditioned CS-US association. Instead, according to this theory, an additional inhibitory (safety) association is formed during extinction. This safety memory is thought to have an inhibitory influence on the US memory. When confronted with the CS after extinction, the initial CS-US association and the safety memory will compete for expression. Bouton (1993) argues that if the CS-US association is easier to retrieve than the safety memory, the CR will return in response to the CS. For example, if the CS-US association is triggered by confrontation with the US, this association might be easier to retrieve than the safety memory, causing reinstatement of the CR (Bouton, 2002; Craske, Liao, Brown, & Vervliet, 2012).

Also on a neural level, studies have been looking at the mechanisms of conditioning and extinction. In general, studies looking at the neural underpinnings of conditioning and extinction agree on the central role of the amygdala in forming the CS-US association (Hitchcock, & Davis, 1986; LeDoux, Sakaguchi, & Reis, 1984) during fear conditioning (Quirk & Milad, 2012). In turn, output of the amygdala is projected to the brainstem, which is thought to play a role in evoking the CR (Le Doux, Ruggiero, & Reis, 1985). The prefrontal cortex plays a role in both fear-expression after conditioning and inhibition of this fear expression after extinction. Activation in the dorsal anterior cingulate cortex (dACC) is positively correlated with fear expression, mediated by excitatory stimulation of the amygdala. In extinction, an important role of the ventromedial prefrontal cortex (vmPFC) is proposed to be in the formation of a safety memory and in inhibiting the output of the amygdala to the brainstem during extinction recall, thereby inhibiting the expression of conditioned fear responses (Phelps, Delgado, Nearing, & LeDoux, 2004). Altogether, neurocognitive mechanisms of conditioning and extinction have been studied quite extensively – interest in studying alternatives to extinction is rising, as discussed below.

Preventing the return of the conditioned response

As extinction is not as effective in preventing the return of the CR as one would hope for, efforts have been made to find better alternatives. One such alternative is counter-conditioning. Instead of merely omitting the US when presenting the CS, in counter-conditioning the CS is now coupled with a new unconditioned stimulus. This new stimulus is of opposite valence compared to the initial US. For example, if during conditioning a tone is coupled with a shock, this CS could be coupled with the delivery of food during counter-conditioning.

Correia, McGrath, Lee, Graybiel, and Goosens (2016) studied an aversive-to-appetitive counter-conditioning paradigm in rats. After initial auditory fear conditioning, a subgroup of the animals received reward conditioning. The findings indicate that an extinction procedure followed by a counter-conditioning procedure is more effective in reducing spontaneous recovery of the fear response compared to extinction alone. Similarly, to the findings of Correia et al. (2016), with human subjects counter-conditioning seems to be more effective over the long term as well (Kang, Vervliet, Engelhard, van Dis, & Hagenaars, 2018). After a time interval, the return of threat expectancy is reduced after counter-conditioning: In other words, less spontaneous recovery takes place (Kang et al., 2018). Additionally, several studies have successfully prevented reinstatement of the CR in an appetitive-to-aversive counter-conditioning paradigm (in rodents: Tunstall, Verendeev, & Kearns, 2012; in humans: Kerkhof et al., 2011; Kaag et al., 2016). For example, Karel et al. (2018) showed that cue-conditioned cocaine administration in rats could be successfully counter-conditioned with a footshock, by reducing reinstatement of reward-seeking behaviour. Although recently promising results have been found, results of older studies do not support a difference in effectiveness between counter-conditioning and extinction with regard to preventing the return of fear responses (Brooks, Hale, Nelson, & Bouton, 1995), pointing out the need for more studies to assess the effectiveness of counter-conditioning compared to extinction.

Taken together, both studies looking at appetitive-to-aversive and aversive-to-appetitive show that counter-conditioning is a promising alternative to the common extinction procedures. In addition, some studies have addressed the neural correlates of counter-conditioning (in rodents: Correia et al., 2016; for appetitive-to-aversive counter-conditioning in humans: Bulganin, Bach, & Wittemann, 2014). However, to our knowledge the mechanisms of aversive-to-appetitive counter-conditioning have not yet been assessed in human subjects. Taken together, therefore the aim of the current study is to test the effectiveness of counter-conditioning compared to extinction in preventing the return of fear responses. Moreover, we aim to explore the neurocognitive mechanisms underlying counter-conditioning. In the current study, less return of the (physiological) fear response can be expected after counter-conditioning compared to regular extinction as found in previous studies, supporting the effectiveness of counter-conditioning (Kerkhof et al., 2011; Kaag et al., 2016; Karel et al., 2018). Additionally, based on previous work looking at counter-conditioning and other alternative procedures for extinguishing CRs, we hypothesize one of three neurocognitive mechanisms to underlie counter-conditioning: enhanced extinction, competition or overwriting, which will be explained below.

Hypotheses

The first proposed mechanism is enhanced extinction. Similar processes could underlie both extinction and counter-conditioning. In this case, corresponding to the theory of Bouton (1993) that views extinction as the formation of a novel inhibitory safety memory, counter-conditioning may also induce an additional safety memory regarding the CS, yet a stronger safety memory than standard extinction. Dunsmoor et al. (2019) studied an alternative to extinction, namely: novelty-facilitated extinction (NFE). This paradigm is similar to counter-conditioning, as in that not only the CS is no longer paired with an US, but rather a new stimulus is paired with the CS, namely: a novel neutral stimulus (e.g., a tone). The difference between counter-conditioning and NFE is that the stimulus coupled with the CS is of greater appetitive valence in counter-conditioning than in NFE. Indeed, after NFE compared to extinction, Dunsmoor et al. (2019) found faster fear reduction and increased activation of the vmPFC which is involved in inhibiting the amygdala. Dunsmoor et al. (2019) argue that these results are induced by enhanced attention due to the level of surprise, prompting more learning during the NFE trials than during regular extinction. In the end, this results in a stronger safety memory that might be easier to retrieve and therefore better able to compete with the original CS-US association, leading to more inhibition of the latter.

The second proposed mechanism is competition. Correia et al. (2016) found increased activity of the basolateral amygdala (BLA)-ventral striatum (NAc) circuitry in rodents after counter-conditioning, which enhanced fear reduction. These findings suggest the involvement of reward processing areas (Schultz, Tremblay, & Hollerman, 2000) during counter-conditioning. Therefore, a second mechanism that we propose is that an additional reward memory is formed during counter-conditioning – instead of a safety memory – that competes with the fear memory for expression. Due to the emotional value of this reward memory, this memory might be of similar ease to retrieve, offering better competition with the fear association than a safety memory.

The third proposed mechanism is overwriting. We hypothesize that the initial CS-US association could be overwritten by a reward memory formed during counter-conditioning, due to reconsolidation-like processes. In reconsolidation procedures, a consolidated memory is reactivated and, thereby, brought to a labile state again. This labile state opens a window of opportunity for alternations to be made to this memory. If those changes are then reconsolidated, this results in a permanently altered memory (Alberini & LeDoux, 2013). Several studies suggest reconsolidation as an alternative to extinction for reducing CRs (Schiller, Monfils, Raio, Johnson, LeDoux, & Phelps, 2010; Kroes, Dunsmoor, Lin, Evans, & Phelps, 2017). In the current counter-conditioning paradigm, the initial fear association might still be in a labile state when the counter-conditioning takes place. Therefore, the memory might still be sensitive to alterations. This leads to our third hypothesis that the second reward association is not only formed coexisting with the original fear association, but rather interacts with this initial fear association during consolidation, overwriting this association. This results in only the newly acquired reward association getting consolidated.

The current study

In the current study, we aim to assess the neurocognitive mechanisms underlying counter-conditioning. Specifically, as mentioned above, we hypothesize that enhanced extinction, competition, or overwriting underlies counter-conditioning. To test our hypotheses, we ran a counter-conditioning paradigm in humans. In brief, in a between-subject functional magnetic resonance imaging (fMRI) study, all participants first acquired categorical fear-conditioned memories. Subsequently, half of the participants underwent a regular extinction and the other half a reward counter-conditioning paradigm. A day later, we tested for group differences in the return of threat CRs and episodic memory for items from the acquisition and extinction or counter-conditioning experience. An outline of the design can be found in Figure 1.

In our experiment, we used a categorical fear conditioning paradigm for a number of reasons I will explain. CRs tend to generalize to stimuli related to the CS (Dunsmoor & Murphy 2015). Humans generalize to stimuli that are both physically similar and conceptually related to the CS. This generalization creates the opportunity to use a categorical fear conditioning paradigm, shown to be feasible in other studies (Dunsmoor, Kragel, Martin, & LaBar, 2014; Kroes et al., 2017). The advantage of this paradigm is that unique pictures can be used over all trials and it allows testing episodic memory

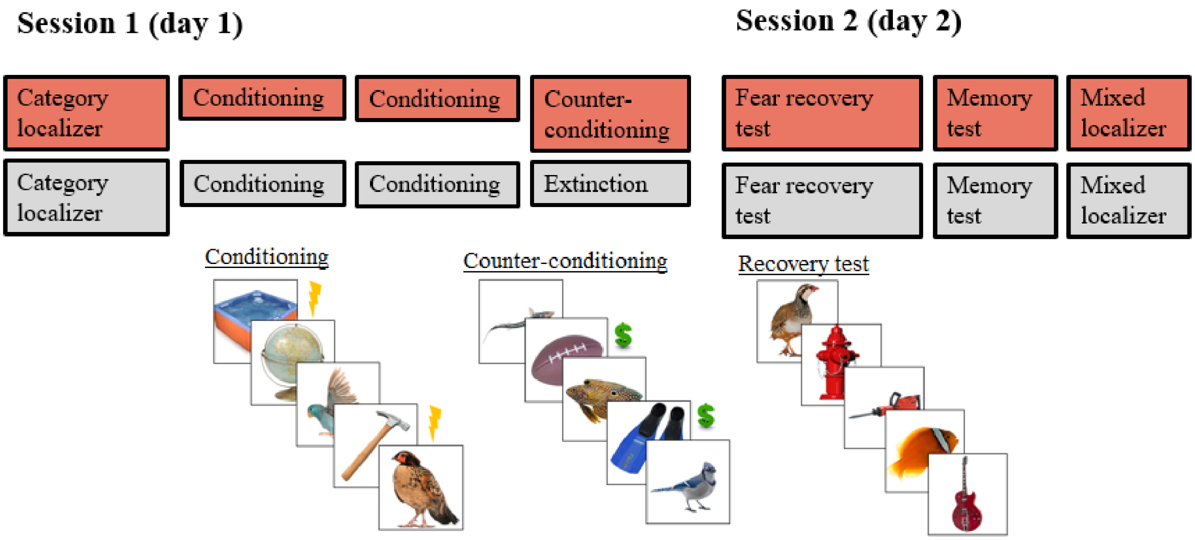


Figure 1. Experimental design. On day 1, after the category localizer, participants did a categorical fear conditioning task. The CS+ category stimuli (animals or objects) were followed by an electrical shock in 50% of the trials. This phase was segmented by a short break halfway through the trials. Next, half of the participants did an extinction procedure, the other half did a counter-conditioning task. In all tasks, trial-unique pictures were used, so on the second day the episodic memory for previously used items could be tested. In this task, participants were presented all pictures used in the conditioning and counter-conditioning or extinction task, mixed with the same number of new pictures. Participants had to indicate if they thought the pictures were new or old (previously seen in another task). In the mixed localizer, neutral abstract stimuli were used to asses brain areas involved in reward and fear learning, as well as general arousal related areas.

for items from the different parts of the experiment, effectively obtaining a retrospective index of memory for different points in time. In the current study, one category consists of pictures of animals and the other of pictures of objects, as animate (in this case the animals) and inanimate (in this case the objects) stimuli are represented by intrinsically different neural organizations (Dunsmoor et al., 2014). Animate stimuli are related to activity in lateral fusiform gyrus (FFG), posterior superior temporal sulcus (pSTS) and inferior occipital regions. Inanimate stimuli are related to activity in medial FFG, posterior middle temporal gyrus and middle occipital cortex (Martin, 2007). Taken together, using a categorical fear conditioning paradigm with the above mentioned categories gives the opportunity to study memory related mechanisms of counter-conditioning and asses changes in representations of stimuli, compared to well established representations of those stimuli under regular circumstances. Differences between extinction and counter-conditioning in those memory related mechanisms and changes in representations will help shed light on the mechanisms underlying counter-conditioning.

As a counter-conditioning procedure a modified version of a monetary incentive delay task was used. In animal studies looking at counter-conditioning, the animals are usually first food deprived in order to make food rewards a meaningful incentive. However, for ethical and pragmatic constraints, food deprivation was not an option in the current study with human subjects. Alternatively, a monetary incentive delay was implemented (MID; Knutson, Westdorp, Kaiser, & Hommer, 2000). By getting participants actively involved in obtaining a reward during the counter-conditioning paradigm, the MID is found to provide enough incentive to instate CRs, as shown following pilot data collection within the current study design. Thus we use an instrumental conditioning procedure as a counter-conditioning procedure against a Pavlovian fear conditioning association. One may argue that mixing Pavlovian and instrumental fear memories is not a clean procedure. Yet, we argue that nearly every appetitive conditioning procedure has an instrumental component (e.g., an animal needs to approach food or lick to swallow sugar water) and this is simply unavoidable.

To induce event segmentation and allow distinct episodic memory traces to be consolidated for fear and safety, the phases (early and late conditioning and counter-conditioning or extinction) of the MID task are separated with a short break. This is in line with the study of Dunsmoor et al. (2018).

In general, we hypothesize that the counter-conditioning paradigm is more effective than the extinction paradigm. More specifically, we expect to find less return of conditioned fear responses in the second session for the counter-conditioning group compared to the extinction group, as measured with both subjective and psychophysiological measures of the fear response. Moreover, we hypothesize that the three proposed mechanisms, that potentially underlie counter-conditioning, would be reflected in distinguishable results from the episodic memory task and from the neural outcome measure. First, if we would find increased vmPFC activity after counter-conditioning in the above mentioned design, suggesting increased reduction of the fear memory, our findings would support the hypothesis that counter-conditioning induces enhanced inhibition of the fear memory compared to extinction. If this is the case, furthermore, similar cognitive mechanisms are expected to underlie counter-conditioning and extinction. Episodic memory for stimuli coupled to the US (CS+ items) tends to be better than for CS- items, as observed in previous conditioning studies (Patil, Murty, Dunsmoor, Phelps, & Davachi, 2017; Dunsmoor, Martin, & LaBar, 2012; Dunsmoor & Murphy, 2015). In the case of enhanced extinction, we expect to find similar results as in those previous studies, in both the extinction and counter-conditioning group.

A second possibility is that a reward memory is formed during counter-conditioning, which competes with the original fear memory. As this reward memory has a meaningful emotional valence on its own, it may induce enhanced retroactive memory consolidation (Patil et al., 2017; Dunsmoor et al., 2012; Dunsmoor, Murty, Davachi, & Phelps, 2015), contributing to stronger competition with the initial CS-US association. If this is the case, we expect to find both involvement of fear and reward learning (neural) mechanisms. As found by Correia et al. (2016), involvement of the ventral striatum is expected after counter-conditioning, next to neural activity related to the conditioned fear memory. This would represent the presence of both a fear and the reward memory which compete after counter-conditioning. Furthermore, in the case of competition, both stimuli coupled with the initial US and stimuli coupled with reward during counter-conditioning are expected to be better remembered than CS- items, due to the fact that CS+ items are associated with the emotional valence of the initial US and the reward used in counter-conditioning.

In contrast to the two formerly proposed mechanisms, for our third hypothesis of overwriting, we expect no involvement of brain areas expressing or inhibiting the fear response. However, involvement of areas related to reward learning, like the ventral striatum, is expected. Moreover, compared to the other two mechanisms, no memory benefits are expected for items related to the US of conditioning, due to the fact that this emotional association is overwritten by the new, opposite emotional association during counter-conditioning. More specifically, it is expected that CS+ items from the counter-conditioning phase will be better remembered than CS- items from this phase, however, no such difference is expected for items from the conditioning phase, in contrast to the two previously mentioned mechanisms.

In summary, the current study aims to asses counter-conditioning as a promising alternative to general extinction for reducing undesired CRs. So far, it has been shown in animal and human studies that counter-conditioning can be more effective compared to standard extinction in preventing the return of the CR. This study aims to extend the literature about counter-conditioning by unravelling the neurocognitive mechanisms underlying this increased effectiveness. This will be done by comparing physiological fear responses, neural activation and episodic memory in a categorical fear conditioning paradigm between groups that either undergo standard extinction or counter-conditioning. We hypothesize that counter-conditioning either results in increased inhibition of the initial fear memory, competition between the fear and reward memory, or overwriting of the fear memory by the reward memory. Shedding light on the effectiveness and mechanisms of counter-conditioning can, in the end, help to improve clinical treatments of undesired conditioned associations found in the clinical population.

Methods

All study proceedings, methods and analyses strategies were preregistered on the Open Science Framework (www.osf.io/).

Participants

Up to this point, 20 Dutch speaking participants took part in this study at the time of analyses (13 female, 7 male). All were healthy adults aged between 18 and 35 years, with normal or corrected-to-normal vision. All participants met the inclusion criteria for MRI (magnetic resonance imaging). People with a history in psychiatric and/or neurological disorders, and people with (by medication induced) altered endocrine and/or hormonal functioning (except by hormonal contraception) were excluded from participation. Participants were recruited via an online recruitment system ('Radboud Research Participation System') or by personal approach. Participants received 10 € per hour in return for participation and could earn a certain extra bonus during part of the tasks. Participants were randomly assigned to either the counter-conditioning (7 females, 3 males, age: 22.0 ± 2.06) or extinction (6 females, 4 males, age: 22.3 ± 2.45) group. One participant from the extinction group did not come to the second session of the experiment. Participants with missing or incomplete data for a critical test were excluded from analyses for that test but will still be included in tests on other data, in accordance with our preregistration. The study was approved by the local ethical review board (CMO region Arnhem-Nijmegen) and in accordance with the Declaration of Helsinki.

Stimuli

A set of 220 animal and 220 object pictures was used for this study. All pictures were luminance equalized, including the grey background. The order of the pictures was pseudorandomized for each even and odd participant couple (participant 1 and 2, 3 and 4, 5 and 6, etc.), so that the participants of this couple had the same order of pictures but for each new couple, a new order was created. The category serving as the CS+ was as well randomly determined per even and odd couple. The first 32 pictures, from both categories, of this pseudorandom order were used for the categorical localizer task, the next 40 were used in the categorical fear conditioning task. In a similar fashion, 40 pictures were assigned to the counter-conditioning or extinction task and 20 pictures to the fear recovery test. In the episodic memory test, the pictures used in the acquisition and extinction phase were used together with 80 new pictures from each category. Phase scrambled versions of the pictures were created for the functional stimulus category localizer task (Reinders, Den Boer, & Büchel, 2005), for a control block in this task. For the mixed valence localizer task, three coloured squares were used. Those squares were also luminance equalized including their background.

Materials

All tasks described here were performed in the MRI scanner. Three of the tasks were performed in the first session, taking place on the first testing day. The other three tasks were completed in the second session, which took place on the day after the first session (as illustrated in Fig. 1).

Functional stimulus category localizer

This task is based on the categorical localizer task used by Voogd, Fernandez, and Hermans (2016a). Note that we look at animals vs. objects instead of animals vs. fruits/vegetables, similar to Dunsmoor et al. (2014). fMRI data from this task is used to assess which areas are involved in processing pictures from the two categories in general. This task consists of three different block types. One block with pictures of animals, one with pictures of objects and one with phase-scrambled versions of pictures of animals and objects. Each block is repeated 12 times in a random order, with the random order of the blocks of the first half of the task being mirrored for the second half. Picture order within each block is randomized. Additionally, 12 rest blocks are included, consisting of a blank screen with a fixation cross. The blocks consist of 32 pictures being presented for 625 ms. The task for the participant is to watch the pictures and pay attention to a white circle which is presented superimposed on one of the pictures in half of the blocks. Once they see this circle, they have to react with a button press. With regard to the episodic memory task, pictures will only be used once in the tasks prior to the episodic memory test. Therefore, pictures used in this part of the experiment are not used again in other parts of the experiment.

Categorical fear conditioning task

This task is a combination of the categorical fear conditioning paradigm as described by Dunsmoor et al. (2012) and the MID as described by Knutson et al. (2000). The choice for the addition of aspects of the MID is based on reasons mentioned above with regard to the counter-conditioning task. For the purpose of keeping most tasks as similar as possible, these MID aspects are also incorporated into the acquisition phase, in the categorical fear conditioning task. In the practice version of this task, a green square is presented. After a variable interval, a cue (a white circle) appears superimposed on this square at a random location. The task of the participant is to press a button as soon as this cue appears. This is repeated ten times with variable inter-trial intervals (ITI).

In the actual version of this task, participants are first presented with a picture from one of the two categories (animals or objects). Figure 2 contains a representation of the setup of this task. After a variable interval (2.5 - 4 s) the cue appears on top of this picture, indicating that the participant has to

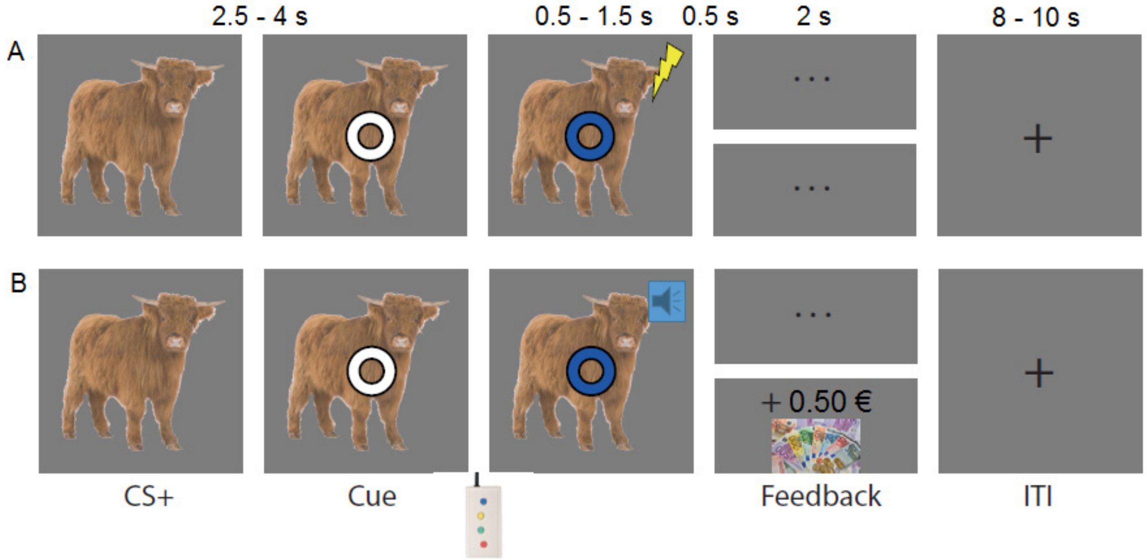


Figure 2. Categorical conditioning and counter-conditioning task including MID aspects. A. Representation of the MID aspect in the task used during the acquisition phase. In this phase 50% of the items from the CS+ category were coupled with a shock. B. Representation of the counter-conditioning task. The extinction group did the same task, however, they never received the monetary reward.

react by pressing the button as quickly as possible within the response window of 1 s. After pressing the button, the cue changes colour as feedback for the participant that their key press was effective. The cue remains on screen for 0.5 - 1.5 more seconds. In 50% of the trials from the CS+ category, a shock was administered during this time window. Subsequently, the neutral feedback of three dots (…) appears for 2 s after the picture and cue disappeared. This is followed by a variable inter-trial-interval (8 – 10 s) during which a fixation cross is presented. This is repeated 80 times with 40 trial-unique animal pictures and 40 trial-unique pictures of objects in a random order. In the middle of the conditioning phase, there is a short break of 10 s (as in Dunsmoor et al., 2018).

Counter-conditioning/extinction task

This task is used to compare two different strategies to extinguish a conditioned fear response. This task is largely the same as the categorical fear conditioning task: participants first see a picture and have to press a button as soon as the cue appears superimposed on the picture. The number of stimuli as well as the timings of the trials are the same as in the categorical fear conditioning task. In the extinction group, the pictures are no longer coupled with a shock. In the counter-conditioning group, none of the pictures are coupled with a shock either. However, now about 70% of the pictures from the CS+ category are coupled with a 0.50 € reward and the message that they receive this reward because their response was fast enough. This percentage is set based on reinforcement rate pilot experiments that showed to induce effective conditioned responses. In order to create a smooth transition from the conditioning phase to the counter-conditioning phase, the first two CS+ trials are not rewarded. From the third trial onwards, the participants are able to receive rewards. The initial response time threshold for obtaining a reward is set to 10% faster than participants' responses during practice. During the remaining trials, this response threshold is altered dynamically in order to approach a 70% reinforcement rate. Note that this is a higher reinforcement rate than during fear acquisition (50%), but pilot studies indicated a higher reinforcement rate to be necessary to induce appetitive conditioned responses.

Fear recovery test

First, participants are presented 15 unique new stimuli from each of the two categories, again with the task to press the button as soon as possible after the cue appears on top of the picture. During this part of the task, the participants receive neither shocks nor rewards, therefore spontaneous recovery of the fear response can be assessed. Critically, pilot studies indicated that reward anticipation did not evoke anticipatory physiological responses.

Timings of the stimuli in this task are the same as in the previously described tasks. Next, we test for reinstatement of the fear response: participants receive three unannounced electrical shocks, followed by an additional five unique new stimuli per category, which are not followed by a reward or shock, to test for fear recovery in relation to those stimuli.

Episodic memory test

This task is used to get an indication of the episodic memory performance for CS+ and CS- items, which are used in different phases of the experiment. Participants are presented with all the pictures shown during the conditioning and counter-conditioning/extinction phase (80 pictures per category in total), one at a time. These are mixed in a random order with an equal amount of new pictures from the two categories. In total, this task consists of 320 items. The task of the participants is to indicate on a 6-point Likert scale how sure they are that a certain picture is an old (1) or new picture (6). They had to do this within 3 s after stimulus onset.

Mixed valence localizer

This task is used to localize brain activation evoked by arousal ([CS+shock & CS+reward] vs. [CS-]) and valence ([CS+shock] vs. [CS+reward]). This part consists of three novel neutral stimuli presented 20 times each in a random order. The stimuli used in this task are abstract figures (squares) in three different colours compared to the pictures of real life animals and objects as in the previously described tasks. One of the stimuli was followed by a shock 50% of the time. The second stimulus was followed by reward in 70% of the time, while the third stimulus was never followed by a reward or shock. The shock and reward rate as well as the timings of the events in this task are the same as in the rest of the experiment.

Valance, arousal and contingency awareness measures

Before conditioning, before extinction, after extinction, before the fear recovery test and after the fear recovery test, participants are asked to rate how negative or positive they experience pictures from the animal and object category at that moment in general (valence) and how arousing they experience the pictures to be in general (calm vs. excited) on a 9-point Likert scale. In addition, participants are asked to estimate for every different phase (conditioning, extinction, spontaneous recovery and reinstatement) what percentage of CS+ and CS- category images were paired with shocks. For participants in the counter-conditioning group, reward estimation and contingency awareness is also assessed for the different tasks.

Procedure

After receiving general information about the procedure and signing the informed consent form, participants were taken into the scanner. First, the shock intensity level was calibrated with an ascending staircase procedure starting with a low voltage (near a perceptible threshold) to reach a level deemed “maximally uncomfortable without being painful” by the participant. Participants were instructed that the intensity of the shocks would stay the same over the course of the experiment. Second, the peripheral psychophysiological and MRI measurements were prepared before the participant could start with the tasks.

On the first day, participants started with the functional stimulus category localizer task. The next task was the categorical fear conditioning task, preceded by practice trials of this task. In this acquisition phase, participants were fear conditioned by pairing trial-unique images of one category (animals or objects) with a mild electrical shock half of the time. Participants were assigned to either one of the fear conditioned categories by counterbalancing. Before starting the tasks, participants were told a contingency exists between the category of the stimuli and the outcome of the trials. Participants were instructed that their instrumental responses had no influence on the shocks they received. Moreover, they were instructed that for the rewards they could receive throughout the task, a relation exists between their reaction time in response to the cue and the possibility of receiving a reward. For half of the participants, the third task of the first session was an extinction phase and the other half received counter-conditioning.

At the start of the second session, participants were reminded that the shock intensity during this session was the same as set during the first session. Furthermore, they were reminded of the non-existing relation between reaction times and the possibility of receiving a shock, as well as of the existing relation between their responses and the possibility of receiving a reward. The first task of the second session was a test for the return of the conditioned fear response. Next, participants did an episodic memory test. The last task of the experiment was the mixed valence localizer paradigm.

Subjective valence and arousal ratings were sampled at the following time points: before and after the acquisition phase, after the extinction phase and before and after the fear recovery test. At the end of both sessions, participants estimated the contingency between the stimulus categories and the possibility of receiving a shock or reward.

Physiological measurements

Skin conductance responses

Electrodermal activity was measured using two Ag/AgCL electrodes attached to the middle- and forefinger of the left hand with a BrainAmp MR system and BrainVision Recorder software (Brain Products GmbH, Munich Germany). Before analysis, data were first cleaned from radio frequency (RF) artefacts and high frequencies (de Voogd, Fernández, & Hermans, 2016a, 2016b). Next, SCR responses were automatically scored and manually checked using Autonomate (Green, Kragel, Fecteau, & LaBar, 2014) implemented in Matlab 7.14 (MathWorks). This procedure is in line with previously published procedures (de Voogd et al., 2019). SCR amplitudes were calculated as the maximum through-to-peak deflection in the latency window of 0 – 12 s after trial onset for each presentation of categorical stimuli. After pilot experiments with the current setup, it was decided to use this long latency window (instead of 0 - 4 s which is used in most studies) because SCRs were found to arise relatively late. This might be due to the fact that in the current MID version of a classical conditioning task, the shock could only come relatively late after the trial onset (after the button press). Only trials in which no shock or reward happened were taken into account. SCRs were calculated for the following tasks: the categorical fear conditioning task, extinction/ counter-conditioning, the fear recovery test and the mixed valence localizer.

Pupil dilation

The extent of pupil dilation in reaction to viewing the CS images was measured with a MR-compatible eye-tracker from SensoMotoric Instruments in the same tasks as the SCR measurements. This device was attached to the scanner bed. The pupil diameter was sampled at a rate of 50 Hz. Data were analysed using in-house software (Hermans, Henckens, Roelofs, & Fernandez, 2013) implemented in Matlab 7.14 (MathWorks), which was based on methods described previously by others (Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003). Eyeblink artefacts were identified and linearly interpolated 100 ms before and 100 ms after each identified blink. After interpolating missing values, time series were band-pass filtered at 0.05 to 5 Hz and z-scored (by subtracting the mean and dividing by the standard deviation) within each participant and run to account for between-subject variance in overall pupil size (including variance related to corrective lenses). Event-related pupil diameter responses were calculated by averaging pupil diameter during the 1 - 4 s period after stimulus onset and dividing by the averaged 1 s pre-stimulus pupil diameter (-1 - 0 s) (de Voogd et al., 2016a). The average of pupil dilation response (PDR) was computed per condition, phase and participant.

Respiration and heart rate

Respiration was measured using a respiration belt placed around the participant's diaphragm. Heart rate was measured using a pulse-oximeter. Raw pulse and respiratory data were processed offline using in-house software for visual artefact correction and peak detection, and were used for retrospective image-based correction (RETROICORplus) of physiological noise artefacts in BOLD (blood-oxygenation level dependent)-fMRI data (Glover, Li, & Ress, 2000). In line with procedures previously described by de Voogd et al. (2016a), processed pulse and respiratory data were used to specify fifth-order Fourier models of cardiac and respiratory phase-related modulation of the BOLD signal, yielding ten nuisance regressors per modality. An additional six regressors of no interest were calculated for heart rate frequency, heart rate variability, abdominal circumference, respiratory frequency, respiratory amplitude and respiration volume per unit time.

MRI data acquisition

During all tasks whole-brain functional imaging was conducted on a 3.0 Tesla PrismaFit MRI scanner equipped with 32-channel transmit-receiver head coil. The manufacturer's automatic 3D-shimming procedure was performed at the beginning of each experiment. Participants were placed in a light head restraint within the scanner to limit head movements during acquisition.

Functional images were acquired with multi-band multi-echo gradient echo-planar imaging (EPI) sensitive to the BOLD response using the following parameters: 51 oblique transverse slices, slice thickness = 2.5 mm, repetition time (TR) = 1.5 s, flip angle α = 75°, echo times (TE) = 13.4, 34.8

and 56.2 ms, field-of-view (FOV) = 210 x 210 mm^2, matrix size 84 x 84 x 64, fat suppression.

A magnetic (B_0) field map was collected at the start of each session and used to unwarp the echo-planar images (Hutton, Bork, Josephs, Deichmann, Ashburner, & Turner, 2002). A structural image (1 mm isotropic) was acquired using a T1-weighted 3D magnetization prepared rapid gradient echo (MPRAGE) sequence with the following parameters: TR = 2300 ms, TE = 3.03 ms, flip angle α = 8°, 192 contiguous 1 mm slices, FOV = 256 x 256 mm^2, to be used for normalization procedures.

MRI data preprocessing

Data were preprocessed with *fMRIPrep* 1.2.6-1 (Esteban, Markiewicz, et al., 2018; RRID:SCR_016216). First, a reference volume and its skull stripped version were generated using a custom methodology of *fMRIPrep*. A deformation field to correct for susceptibility distortions was estimated based on a field map that was co-registered to the BOLD reference, using a custom workflow of *fMRIPrep* derived from D. Greve's *epidewarp.fsl* script and further improvements of HCP Pipelines (Glasser et al., 2013). Based on the estimated susceptibility distortion, an unwarped BOLD reference was calculated for a more accurate co-registration with the anatomical reference. Head motion parameters with respect to the BOLD reference (transformation matrices and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using *mcflirt* (FSL 5.0.9, Jenkinson, Bannister, Brady, & Smith, 2002). The BOLD time series were resampled onto their original, native space by applying a single composite transform to correct for head motion and susceptibility distortions. These resampled BOLD time series will be referred to as preprocessed BOLD in original space or just preprocessed BOLD. A T2* map was estimated from the preprocessed BOLD by fitting to a monoexponential signal decay model with loglinear regression. For each voxel, the maximal number of echoes with reliable signal in that voxel were used to fit the model. The calculated T2* map was then used to optimally combine preprocessed BOLD across echoes following the method described in Posse et al. (1999). The optimally combined time series was carried forward as the preprocessed BOLD and the T2* map was also retained as the BOLD reference. The BOLD reference was then co-registered to the T1 weighted (T1w) reference using *bbregister* (FreeSurfer) which implements boundary based registration (Greve & Fischl, 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. The BOLD time series were resampled to MNI152NLin2009cAsym standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. For analyses in native space, weighted, realigned and unwarped functional images co-registered to each participants' structural scan are smoothed using a 6 mm FWHM (full width at half maximum) Gaussian.

Statistics

In general, dependent variables were submitted to repeated measures ANOVAs, using SPSS (IBM SPSS Statistics Inc.) to perform the analysis. Effects with a *p*-value lower than .05 were considered significant and were followed up with independent and paired samples t-tests. Partial eta-squares were reported as effect sizes of significant effects. Additionally, means and standard deviations were reported where relevant.

Neuroimaging data were analysed using SPM12 (Wellcome Trust Centre, www.fil.ion.ucl.ac.uk) implemented in MATLAB. Statistical thresholds of second-level whole-brain analyses were set at familywise error (FWE) corrected cluster p < .05. For the first level analysis of the categorical localizer, regressors were made for each block type (animals, objects and scrambled). Onsets and durations of the blocks were saved online during the execution of the task. Additionally, a nuisance regressor was used to account for events of no interest. For the first level analysis of the conditioning, counter-conditioning/extinction task and fear recovery test, regressors were made for the CS+ and CS- items. In order to be able to assess changes in brain activity within a task, the regressors for the conditioning and counter-conditioning/extinction task were split into regressors for CS+ items from the first and the second half of the task; The same was done for the CS- items. Again, a nuisance regressor was added to account for events of no interest. Contrast estimates from the first level analysis were taken into group level analysis.

Results

Valence and arousal ratings

Results of the subjective valence and arousal ratings of the two categories show that participants acquired conditioned responses during the conditioning task. Moreover, in both groups these

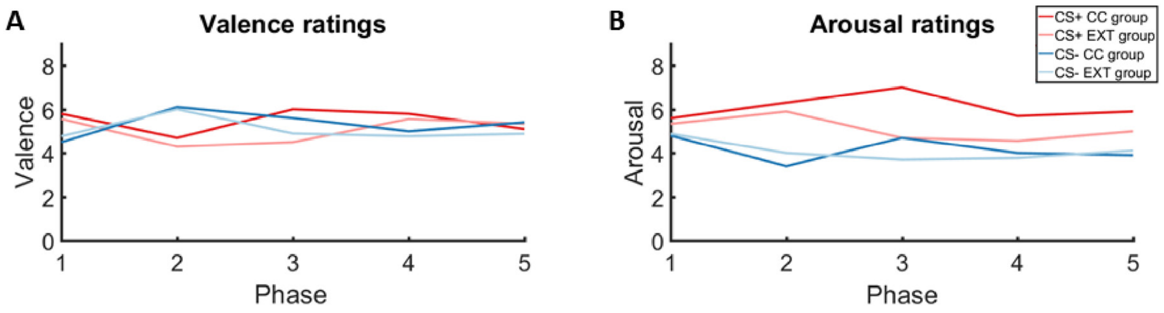


Figure 3. Mean subjective ratings of A. valence and B. arousal over the course of the experiment. Valence and arousal ratings were sampled before (phase 1) and after conditioning (phase 2), after counter-conditioning/extinction (phase 3), and before (phase 4) and after (phase 5) the fear recovery tests on the second day. Ratings were given on a 9-point Likert scale with 1 = very negative or absolutely not arousing, and 9 = very positive or extremely arousing.

conditioned associations were successfully reduced by the extinction and counter-conditioning task.

To test the effect of the conditioning and counter-conditioning (CC)/extinction (EXT) tasks on the evaluation of the two categories, valence and arousal rating were used in a group (EXT, CC) x time point (before conditioning, before extinction, after extinction, before fear recovery test, after fear recovery test) x CS-type (CS+, CS-) 2 x 4 x 2 repeated measures ANOVA. For the valence ratings, this revealed a significant time point*CS-type interaction (F(4, 64) = 3.40, p = .014, η_p^2 = .18). Right after the conditioning phase, there was a significant difference between the ratings for the CS+ (M = 4.50, SD = 1.47) and CS- items (M = 6.10, SD = 2.00). This was not the case for the other time points, as illustrated in Figure 3A and indicated by post-hoc paired-samples t-tests (t(19) = -2.75, p = .013). Thus, both groups show acquisition and later extinction of conditioned differential responses. Neither group shows return of these responses on the second day.

For the arousal ratings, both a main effect of CS-type (F(1, 16) = 6.35, p = .023, η_p^2 = .28) and a significant time point*CS-type interaction (F(4, 64) = 5.46, p = .001, η_p^2 = .25) were found. Overall, the arousal ratings were higher for the CS+ category than for the CS- category. However, when looking at the different time points in a post-hoc paired-samples t-test, this difference was only significant right after the conditioning phase (CS+ ratings: M = 6.05, SD = 1.36; CS- ratings: M = 3.70, SD = 1.78; t(19) = 4.92, p < .001, Fig. 3B). Neither measurement was influenced by the experimental group the participants were in. No main effects or interactions were found for the between-subjects factor group. If the extinction and counter-conditioning tasks would differ in effectiveness in preventing the return of the conditioned responses, differential CS+/CS- ratings would also be expected on the second day in the extinction group. Which would suggest return of the fear association in the extinction group, while this return will then ideally be prevented in the counter-conditioning group.

Reaction time data

Reaction time data from the categorical fear conditioning task were subjected to a group (EXT, CC) x phase (early, late) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANONA. This analysis revealed a main effect of CS-type (F(1, 18) = 4.65, p = .045, η_p^2 = .21). Participants were faster in responding to the target in trials where they could receive a shock (respectively, M = .40, SD = .05; M = .42, SD = .06; t(19) = -2.20, p = .040). There was also a main effect of group (F(1, 18) = 12.74, p = .002, η_p^2 = .41): participants in the counter-conditioning group were in general slower in responding than the ones in the extinction group (respectively, M = .44, SD = .05; M = .38, SD = .03; t(18) = 3.57, p = .002). No interactions were found with the between-subjects factor group; There seems to be a baseline difference in response times between the groups, although, this difference was constant over time and between the CS-types.

To confirm the effectiveness of the MID aspect of the counter-conditioning task to learn the association between the CS+ items category and the possibility of receiving a reward after a fast enough button press, reaction time data of the counter-conditioning and extinction tasks were subjected to a group (EXT, CC) x phase (early, late) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. There was indeed a main effect of CS-type (F(1, 18) = 20.75, p < .001, η_p^2 = .54). In the counter-conditioning/extinction task, participants responded faster to targets in CS+ trials (M = .38, SD = .01)

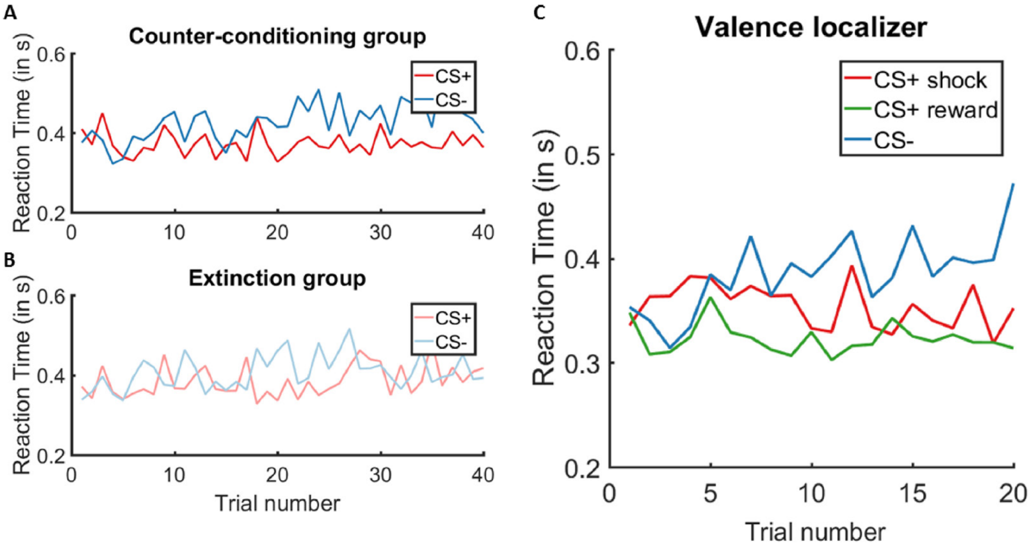


Figure 4. Reaction time data showing mean reaction times of participants in the two groups, illustrated per trial and per trial type (CS+ or CS-). A. & B. Reaction times from the counter-conditioning/extinction task illustrate participants keep responding fast if they can receive a reward in that trial (phase*CS-type*group (F(1, 18) = 12.11, p = .003, ηp2 = .402). C. Reaction times per trial and CS-type from the mixed valence localizer task, illustrating differential reaction times for the three different trial types (reward trials: M = .32, SD = .01; shock trials: M = .35, SD = .01; CS- trials: M = .39, SD = .01).

than they did in CS- trials (M = .42, SD = .01). Additionally, over the course of the task participants responded slower (first half: M = .39, SD = .01; last half: M = .41, SD = .01), shown by a main effect of phase (F(1, 18) = 12.11, p = .003, η_p^2 = .40). Most importantly, there were both a significant phase*CS-type*group (F(1, 18) = 6.30, p = .022, η_p^2 = .26) and a significant CS-type*group (F(1, 18) = 5.00, p = .038, η_p^2 = .22) interaction effect. As illustrated by Figure 4A (showing the data from the counter-conditioning group) and 4B (showing the data from the extinction group), participants in the counter-conditioning group kept responding relatively fast to the CS+ items, while showing a similar decrease in response times for the CS- items as the participants in the extinction group. Thus, participants in the counter-conditioning group clearly learned to differentiate between the CS+ and CS- items over time, as indicated by the reaction times.

To get an indication of how participants perceive the categories at the start of the second day compared to the end of the first day, reaction times from the spontaneous recovery task were analysed. Responses on the last four trials from the counter-conditioning/extinction task were compared with the first four trials of the spontaneous recovery task with a group (EXT, CC) x task (Ext/CC, spontaneous recovery) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. This analysis only revealed a main effect of phase (F(1, 17) = 9.65, p = .006, η_p^2 = .36). Participants were faster at the beginning of the second session than at the end of the first (respectively, M = .41, SD = .05; M = .36, SD = .05; t(18) = 3.19, p = .005). There were no differences between the groups. An additional group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 repeated measures ANOVA neither showed a significant influence of CS-type nor of group on the data. This might suggest that in neither group the fear or reward association with the CS+ category was influencing the response behaviour of the participants on the second day. If this would have been the case, in line with the first session, faster responses would have been expected for the CS+ than for the CS- items.

To confirm the effectiveness of the reward learning aspect in the mixed valence localizer task, reaction time data were used for a group (ECT, CC) x phase (first and second half of the trials) x CS-type (CS+ shock, CS+ reward, CS-) 2 x 2 x 3 repeated measures ANOVA. There was a significant main effect of CS-type (F(2, 16) = 22.83, p < .001, η_p^2 = .74). Participants were indeed faster in responding to the targets in CS+ reward trials (M = .32, SD = .04) than they were in CS+ shock (M = .36, SD = .06; t(18) = 3.01, p = .008) and CS- trials (M = .39, SD = .05; t(18) = -6.96, p =< .001). Additionally, participants also responded faster in CS+ shock trials than in CS- trials (t(18) = -3.37, p = .003).

In sum, participants respond faster in trials where a shock or a reward could be received. If neither a reward nor a shock could be received, responses for CS+ and CS- trials are similar, as indicated in the extinction task. On the second day, responses in CS+ and CS- trials are comparable. The fear and reward associations formed on the first day do not seem to influence the response behaviour on the second day. This was the same for both groups, which suggests that similar mechanisms influence the responses on the second day for both groups.

Episodic memory test

Memory performances were analysed to assess whether they varied between tasks, groups and CS-types. In order to do so, per participant four memory scores were computed by subtracting the false alarm rates (pFA; false alarm rate was calculated by dividing the total number of ‘very sure old’, ‘sure old’, ‘probably old’ responses by the total number of responses to new CS+ or CS− items) from the hit rates (pHit; hit rate was calculated by dividing the total number of ‘very sure old’, ‘sure old’, ‘probably old’ responses divided by the total number of responses to old CS+ or CS− items). One memory score was computed for CS+ items previously seen in the conditioning phase and one for CS- items previously seen in the conditioning phase; The same was done for CS+ and CS- items previously used in the extinction/counter-conditioning task. Those memory scores were submitted to a group (EXT, CC) x CS-type (CS+, CS-) x task (conditioning task, extinction/counter-conditioning task) 2 x 2 x 2 repeated measures ANOVA. There was a significant main effect of task (*F*(1, 17) = 10.14, p = .005, η_p^2 = .37): items previously seen in the conditioning phase were better remembered than items from the counter-conditioning/extinction task (respectively, *M* = .38, *SD* = .03; *M* = .32, *SD* = .03). Additionally, there was a main effect of group (*F*(1, 17) = 6.00, *p* = .025, η_p^2 = .26). Participants in the counter-conditioning group (*M* = .42, *SD* = .12) had overall a higher memory score than participants in the extinction group (*M* = .29, *SD* = .10; *t*(17) = 2.45, *p* = .025). As illustrated in Figure 5, there seems to be an enhanced memory performance for the CS+ items in the counter-conditioning group. However, these differences do not reach significance, as indicated by exploratory paired-samples t-tests on the data from this group (items from conditioning: CS+: *M* = .49, *SD* = .18; CS-: *M* = .38, *SD* = .12; *t*(9) = 1.67, *p* = .129; items from counter-conditioning: CS+: *M* = .44, *SD* = .17; CS-: *M* = .36, *SD* = .10; *t*(9) = 1.93, *p* = .085). In sum, data from the memory test tend to support the competition hypothesis, since CS+ items from both phases of the first session tend to be better remembered than CS- items in the counter-conditioning group.

In order to be able to analyse criterion scores of the memory test, hit and false alarm rates were z-transformed. Next, four criterion scores (−0.5*(zHit + zFA)) were computed for each participant and analysed with a group (EXT, CC) x CS-type (CS+, CS-) x task (conditioning task, extinction/counter-conditioning task) 2 x 2 x 2 repeated measures ANOVA. This analysis also yielded a main effect of task (*F*(1, 17) = 8.65, *p* = .009, η_p^2 = .34). A higher criterion was related to items from the counter-conditioning/extinction task (*M* = .10, *SD* = .06) compared to the conditioning

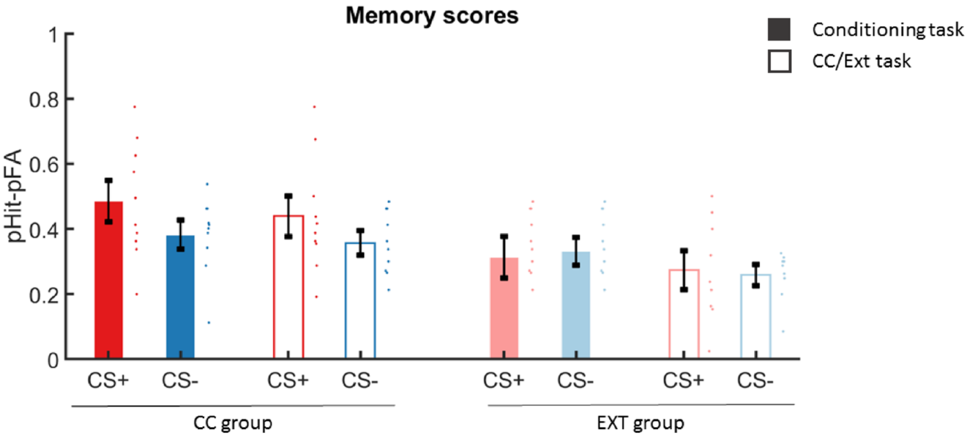


Figure 5. Mean memory scores (Hit rate - False alarm rate), standard errors of the mean (represented by the black bars) and individual memory scores (represented by the scatter plots). Per group memory, scores for CS+ and CS- items previously seen in the conditioning and in the counter-conditioning/extinction task are represented with different bars.

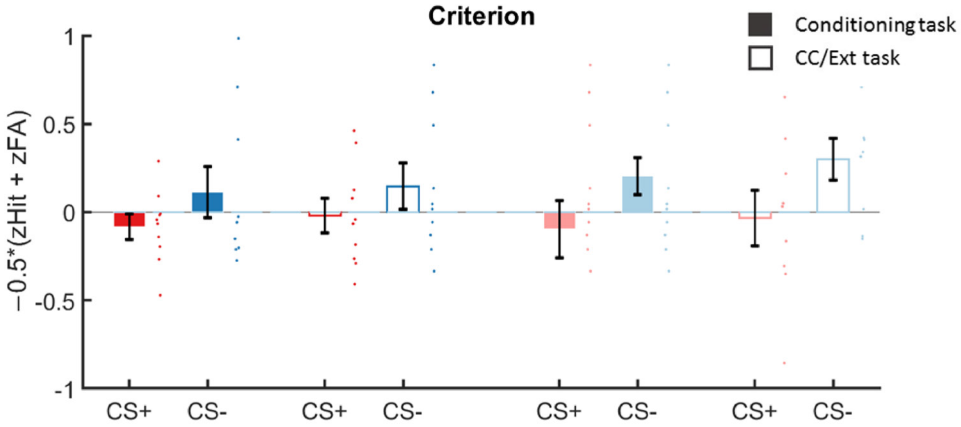


Figure 6. Criterion scores, standard errors of the mean and individual criterion scores from the episodic memory test, presented per group and per task in which the pictures were previously used. Higher criterion scores represent lower likeliness of the participants to state a certain item is old. As illustrated, criterion scores are significantly higher for items from counter-conditioning/extinction.

task (*M* = .03, *SD* = .06). The effect of CS-type was trend significant ($F(1, 17) = 4.23, p = .055, \eta_p^2 = .20$) with a higher criterion for CS- items compared to CS+. In sum, participants are more conservative in stating that an item from the counter-conditioning/extinction task is old rather than new, as they are for items from the conditioning task. Furthermore, participants tend to be more conservative concerning CS- items.

Skin conductance responses (SCR)

Categorical fear conditioning task

To check the effectiveness of the conditioning paradigm, SCR data from this task were analysed using a group (EXT, CC) x phase (early conditioning [first half of the trials], late conditioning [second half of the trials]) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. Only the trials where no shock occurred were taken into account. There was a significant main effect of CS-type ($F(1, 18) = 17.72, p = 0.001, \eta_p^2 = .50$). A post-hoc paired-samples t-test confirmed that SCRs to CS+ items were greater than to CS- items (respectively, *M* = .45, *SD* = .43; *M* = .22, *SD* = .25; $t(19) = 4.21, p < .001$), confirming the effectivity of the paradigm in inducing anticipatory SCRs to the CS+. Additionally, there was a main effect of phase ($F(1, 18) = 8.16, p = .010, \eta_p^2 = .31$): SCRs to both CSs declined at the end of the conditioning task (early: *M* = .43, *SD* = .45; late: *M* = .25, *SD* = .24; $t(19) = 2.90, p = .009$) compared to the first half. There were no group effects: Both groups showed conditioned responses specific to the CS+.

Counter-conditioning phase

To compare the effectiveness of the extinction and counter-conditioning tasks in reducing threat responses, the SCR data were submitted to a group (EXT, CC) x phase (early, late) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. For CS-type, a significant main effect was found ($F(1, 18) = 8.87, p = 0.008, \eta_p^2 = .33$). A post-hoc paired-samples t-test showed that in both groups and over the course of the whole task, the SCRs for CS+ items were greater than for CS- items (respectively, *M* = .22, *SD* = .27; *M* = .16, *SD* = .20; $t(19) = 2.80, p = 0.12$). There was also a significant difference between the groups ($F(1, 18) = 4.48, p = .049, \eta_p^2 = .20$) with participants in the counter-conditioning group having lower SCRs (respectively, *M* = 0.90, *SD* = .10; *M* = .29, *SD* = .28).

Fear recovery test

To test the effect of the counter-conditioning procedure on the return of threat responses during the spontaneous recovery test, SCRs on the last two trials of the extinction or counter-conditioning task were compared with the first two trials of the spontaneous recovery task in a group (EXT, CC) x task (Ext/CC, spontaneous recovery) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. Only a significant main effect of CS-type was found ($F(1, 17) = 5.48, p = 0.32, \eta_p^2 = .24$). Similarly to previous analyses, SCRs for CS+ items were bigger than for CS- items (respectively, *M* = .45, *SD* = .44; *M* = .32, *SD* = .32). However, when separately analysing the first two trials of the spontaneous recovery task in a group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 repeated measures ANOVA, no significant main

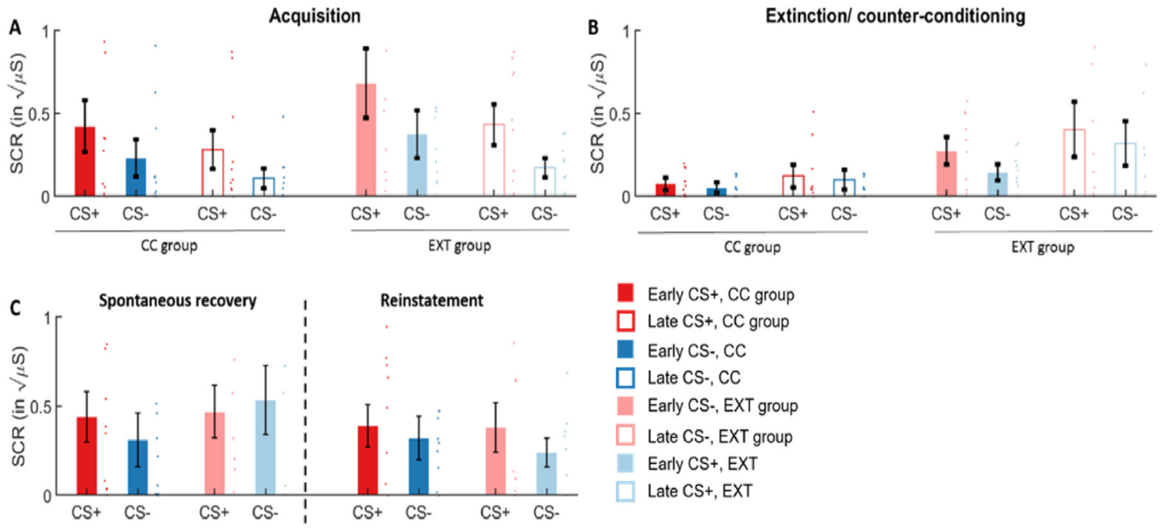


Figure 7. Skin conductance data (SCR) of A. the conditioning task, B. the counterconditioning and extinction tasks for the two groups. C. Shows the data from the first two trials of the spontaneous recovery test at the start of the second session, on the left. On the right, the first trial of the reinstatement test is shown. Figures represent the mean of the square roots of the SCRs, which is measured in micro Siemens. Black bars represent the standard errors of the mean, individual means are represented by the scatter plots.

effect of CS-type remained. As shown by additional paired samples t-tests, neither group shows a differential CS+/CS- SCR at the start of the second session (CC group: CS+: *M* = .30, *SD* = .22, CS-: *M* = .21, *SD* = .18, *t*(9) = 1.78, *p* = .110; EXT group: CS+: *M* = .21, *SD* = .24, CS-: *M* = .20, *SD* = .14, *t*(8) = .11, *p* = .912). Taken together, neither group shows spontaneous recovery of the fear response. The on the first day conditioned increased SCRs to CS+ items did not spontaneously recover after the time interval between the two sessions.

To test for reinstatement in the second part of the fear recovery test, a reinstatement index was calculated (first trial of reinstatement test - last trial of spontaneous recovery test; similar to Lucas, Luck, & Lipp, 2018) and submitted to a group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 repeated measures ANOVA. Neither group shows a significant reinstatement of the fear response for neither the CS+ nor the CS- items. To explore the differences between CS+ and CS- items separately for the two groups, additional t-tests were conducted on the first two trials of the reinstatement phase. In the counter-conditioning, there was no difference between the CS+ and CS- trials (respectively, *M* = .34, *SD* = .28; *M* = .25, *SD* = .21; *t*(9) = 1.32, *p* = .219). The same holds for the extinction group (respectively, *M* = .26, *SD* = .26; *M* = .28, *SD* = .20; *t*(8) = -.30, *p* = .773). Similarly to the spontaneous recovery, no return of SCRs to the CS+ trials was found after triggering reinstatement of this response. Contrary to expectations, in both parts of the retention tests there tends to be more CS+/CS- differences in the counter-conditioning group. Although not significant, the counter-conditioning phase therefore seems to be less effective in preventing the return of the fear responses.

Valence localizer

SCR data from the mixed valence functional localizer task were subjected to a group (ECT, CC) x phase (first and second half of the trials) x CS-type (CS+ shock, CS+ reward, CS-) 2 x 2 x 3 repeated measures ANOVA. There were both a main effect of CS-type (*F*(2, 16) = 7.37, *p* = .005, η_p^2 = .48) and of phase (*F*(1, 17) = 4.65, *p* = .045, η_p^2 = .22). The anticipatory SCR for CS+ shock items (*M* = .83, *SD* = .84) was bigger than both the SCR for the CS+ reward (*M* = .33, *SD* = .39; *t*(18) = 3.92, *p* = .001) and CS- (*M* = .24, *SD* = .24; *t*(19) = 3.87, *p* = .001). The SCR for CS+ reward and CS- did not differ from each other (*t*(18) = 1.45, *p* = .164); confirming the effectiveness of the fear learning aspect of this task. Moreover, these results support the efficacy of SCR data as a unique metric for anticipatory aversive arousal but not reward anticipation. Similar to the SCR data from the conditioning task, SCRs decreased towards the end of the experiment compared to the beginning (respectively, *M* = .52, *SD* = .52; *M* = 40, *SD* = .43; *t*(18) = 2.21, *p* = .040).

In sum, SCR data support the effectiveness of the categorical fear conditioning task in inducing greater anticipatory SCRs in CS+ trials than in CS- trials. In both groups, this differential SCR is successfully reduced at the end of the counter-conditioning/ extinction task. The counter-conditioning paradigm

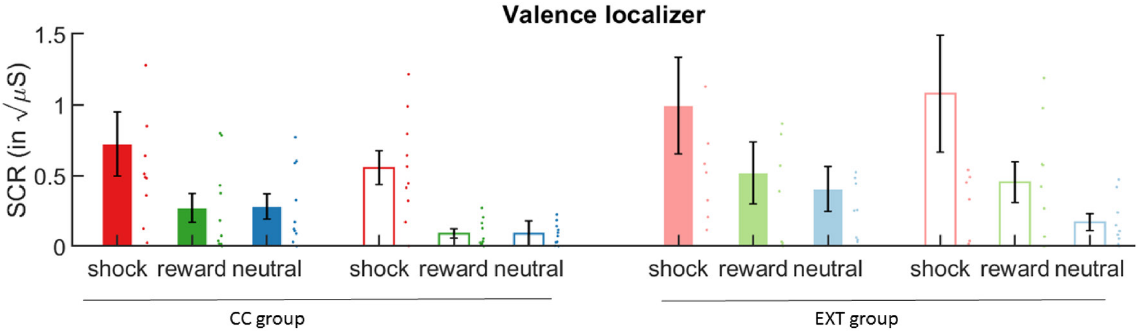


Figure 8. Skin conductance data from the mixed-valence localizer task. Red bars represent the SCR in anticipation of a shock. Green bars represent the anticipatory SCR in trials where a reward could be received. Blue bars show the SCR in CS- trials. As expected, SCRs are higher in anticipation of a shock than of a reward or in neutral trials (CS+ shock: M = .83, SD = .84; CS+ reward: M = .33, SD = .39; CS- M = .24, SD = .24).

seems to reduce fear responses faster. However, no differences were found in effectiveness of preventing the return of the fear responses. On the second day, contrary to expectations, this differential SCR returned in neither group when testing for spontaneous recovery and reinstatement of this fear response.

Pupil dilation data

The pupil dilation responses (PDR) were analysed to assess the effectivity of the categorical fear conditioning task in inducing fear related PDRs. Namely, by submitting the PD (pupil dilation) data to a group (EXT, CC) x phase (early conditioning [first half of the trials], late conditioning [second half of the trials]) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA. This analysis did not reveal significant main effects. However, a significant CS-type*phase*group interaction (*F*(1, 12) = 5.84, *p* = .033, η_p^2 = .33) and a significant CS-type*group (*F*(1, 12) = 7.82, *p* = .016, η_p^2 = .39) were found. These might reflect different starting points of the groups with regard to the different categories. To explore whether the two groups were comparable at the end of the conditioning phase, a group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measures ANOVA was conducted with the data from the second half of the experiment. This yielded a nearly significant main effect of CS-type (*F*(1, 12) = 4.10, *p* = .066, η_p^2 = .25). At the end of the task, anticipatory PDRs for CS+ trials seem to be bigger than for CS- trials (respectively, *M* = .54, *SD* = 1.16; *M* = -.232, *SD* = .59; *t*(13) = 2.00, *p* = .067). So the conditioning phase seems to successfully induce conditioned PDRs in both groups.

For the counter-conditioning/extinction task, the PD data were analysed with a group (EXT, CC) x phase (early EXT/CC, late EXT/CC) x CS-type (CS+, CS-) 2 x 2 x 2 repeated measured ANOVA. This analysis did not reveal significant results, supporting the effectiveness of the counter-conditioning/extinction procedure in reducing conditioned PDRs. To explore whether the absence of an effect of phase, which would be expected if the reduction in conditioned PDRs takes place over the course of the tasks, might be explained by rapid extinction within the first few trials, additional ANOVAs for the separate phases were conducted. This analysis also did not reveal significant results. As described below, data from the valence localizer show a lower PDR for stimuli predicting rewards as for stimuli predicting shocks with PDRs in neutral trials lying in between those two. Since PDRs on day two did not differentiate between the CS-types in both groups, no shock or reward associations that are formed on the first day seem to influence the PDRs on the second day.

To test for spontaneous recovery of the fear response, the last two trials of the counter-conditioning/extinction task and the first two trials of the spontaneous recovery task were submitted to a group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 repeated measures ANOVA. Neither this analysis nor an analysis (group (EXT, CC) x CS-type (CS+, CS-) 2 x 2 repeated measures ANOVA) computed on reinstatement indexes (first trial of reinstatement test - last trial of spontaneous recovery test) of the PD data revealed significant effects.

PD data from the mixed valence localizer were submitted to a group (EXT, CC) x phase (mean of first half, mean of the second half of trials) x CS-type (CS+ shock, CS+ reward, CS-) 2 x 2 x 3 repeated measures ANOVA. This revealed a significant main effect of CS-type (*F*(2, 10) = 5.0, *p* = .031, η_p^2 = .50). Post-hoc paired-samples t-tests show that pupil responses to CS+ shock (*M* = .45, *SD* = .62) were bigger than those for CS+ reward trials (M = -.34, SD = .76; t(12) = 3.19, p = .008). The responses to the CS- items (M = .39, SD = 2.08) did not differ from both CS+ items. Additionally, the above mentioned repeated measures ANOVA revealed a significant main effect of phase (F(1, 11) = 5.52, p = .039, η_p^2 = .33). Responses increased over the course of the experiment (in the first half: M = -.11 SD = .85; in the second half: M = .43, SD = .89; t(12) = -2.37, p = .036).

Taken together, at the end of the conditioning task, there seems to be a conditioned PDR specific to the CS+ category stimuli. Next, no differences between PDRs in the CS+ and CS- trials were left in the counter-conditioning and extinction tasks. No spontaneous recovery or reinstatement seems to take place in the second session, therefore no support was found for a higher effectiveness of the counter-conditioning procedure in preventing fear recovery.

Functional MRI data

Category representation localizer task

BOLD responses to the animals, objects and phase-scrambled blocks were modelled with separate regressors for each participant. Subsequent group level analysis of those responses revealed only a small cluster responsive to animals > objects, namely in the right fusiform gyrus (cluster size = 18 voxels, cluster p < .001, corrected). The opposite contrast of objects > animals revealed two small active clusters, namely in the left fusiform gyrus (cluster size = 9 voxels, cluster p < .001, corrected) and right fusiform gyrus (cluster size = 9 voxels, cluster p < .001, corrected).

Acquisition

Whole brain analysis of the BOLD responses during the categorical fear conditioning task revealed differential activity in several areas for the CS+ > CS- contrast. An overview of those areas can be found in Table 1. Amongst others, activity was found in the right (cluster size = 78 voxels, cluster p < .001, corrected) and left (cluster size = 86 voxels, cluster p < .001, corrected) insular cortex. The reversed contrast of CS- > CS+ did not reveal any significant cluster activity.

Counter-conditioning and extinction task

BOLD responses during the counter-conditioning task were subjected to a second level whole brain analysis. We looked for between-group differences and whether differential CS+ and CS- related activity changed from the first half to the second half of the experiment. Neither differential activity was found between CS+ and CS- items nor did this change over time or differ between the two groups. Exploratory analysis with uncorrected p-values also did not reveal significant effects.

Retention test

Whole brain analyses were conducted on the BOLD responses during the spontaneous recovery and reinstatement tests. No differences were found in activity specific to the CS+ or CS- items. Both categories seem to be retrieved in a similar way with involvement of similar brain areas. Moreover, when exploring group differences in CS+ and CS- related activity, no differences were found either. Neither were there differences when looking at significant activity with an uncorrected p-value.

Mixed valence localizer

To identify regions involved in fear and reward anticipation, BOLD activity in this task was modelled with separate regressors for CS+ shock, CS+ reward and CS- items. Activity was found in the left insular cortex (cluster size = 114 voxels, cluster p < .001, corrected) when looking at the CS+ shock > CS+ reward & CS- contrast, indicating involvement of this area in fear anticipation because it was only active in shock trials and not in neutral trials or trials in which a reward could follow. Activity was found in the superior temporal gyrus (cluster size = 33 voxels, cluster p < .001, corrected) for reward anticipation, represented by the CS+ reward > CS+ shock & CS- contrast.

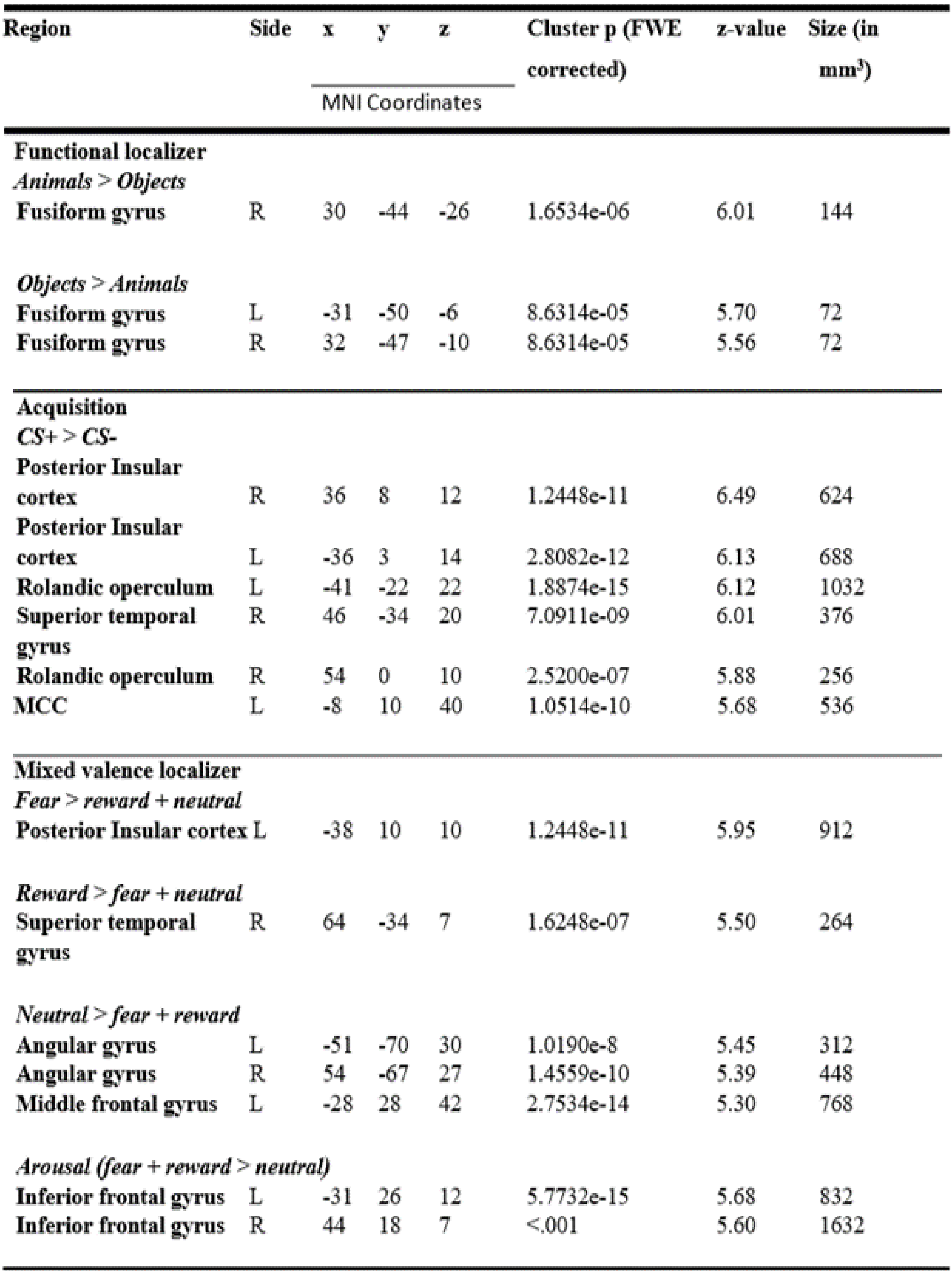
Differential activity in relation to the CS- compared to the other regressors was analysed in order to indicate areas related to situations where no aversive or appetitive outcome was anticipated. Activity was found in frontal middle gyrus (cluster size = 96 voxels, cluster p < .001, corrected). Other involved areas can be found in Table 1.

Lastly, arousal related activity was assessed by a CS+ shock & CS+ reward > CS- contrast. This revealed activity in the bilateral inferior frontal gyrus (left: cluster size = 104 voxels, cluster p < .001, corrected; right: cluster size = 204 voxels, cluster p < .001, corrected).

Discussion

Extinction is a commonly used procedure to reduce undesirable conditioned associations. However, extinction procedures often do not succeed in preventing the return of the extinguished conditioned response (CR). Therefore, the current study aimed to assess the effectiveness of a counter-conditioning paradigm compared to an extinction paradigm. Moreover, this study aimed to assess the neurocognitive mechanisms underlying counter-conditioning and to test the hypotheses that counter-conditioning either induces an enhanced extinction memory, induces a reward memory that competes with the initial fear memory or induces a reward memory that overwrites the fear memory via reconsolidation-like processes. Preliminary results highlight the effectiveness of the study design to

Table 1. Peak voxel coordinates and cluster statistics and size for the localizer paradigm.



Region	Side	x	y	z	Cluster p (FWE corrected)	z-value	Size (in mm^3)
		MNI Coordinates					
Functional localizer							
Animals > Objects							
Fusiform gyrus	R	30	-44	-26	1.6534e-06	6.01	144
Objects > Animals							
Fusiform gyrus	L	-31	-50	-6	8.6314e-05	5.70	72
Fusiform gyrus	R	32	-47	-10	8.6314e-05	5.56	72
Acquisition							
CS+ > CS-							
Posterior Insular cortex	R	36	8	12	1.2448e-11	6.49	624
Posterior Insular cortex	L	-36	3	14	2.8082e-12	6.13	688
Rolandic operculum	L	-41	-22	22	1.8874e-15	6.12	1032
Superior temporal gyrus	R	46	-34	20	7.0911e-09	6.01	376
Rolandic operculum	R	54	0	10	2.5200e-07	5.88	256
MCC	L	-8	10	40	1.0514e-10	5.68	536
Mixed valence localizer							
Fear > reward + neutral							
Posterior Insular cortex	L	-38	10	10	1.2448e-11	5.95	912
Reward > fear + neutral							
Superior temporal gyrus	R	64	-34	7	1.6248e-07	5.50	264
Neutral > fear + reward							
Angular gyrus	L	-51	-70	30	1.0190e-8	5.45	312
Angular gyrus	R	54	-67	27	1.4559e-10	5.39	448
Middle frontal gyrus	L	-28	28	42	2.7534e-14	5.30	768
Arousal (fear + reward > neutral)							
Inferior frontal gyrus	L	-31	26	12	5.7732e-15	5.68	832
Inferior frontal gyrus	R	44	18	7	<.001	5.60	1632

Notes: All coordinates are defined in MNI152 space. All reported statistics are significant at p < .05, cluster corrected.

induce conditioned skin conductance responses (SCRs) and pupil dilation responses (PDRs) and to increase subjective arousal and negative valence ratings with regard to the CS+ category. SCR data from the counter-conditioning and extinction task suggest both procedures are effective in reducing conditioned fear responses. Furthermore, analysis of reaction times recorded during the counter-conditioning and mixed valence localizer tasks indicate that the monetary incentive delay (MID) components of tasks were successful in producing sufficient reward incentive.

We found successful category conditioning with the current task design. After the conditioning phase, subjective ratings, SCRs and PDRs differentiated between CS+ and CS- items. These differences were in the expected directions, namely CS+ items were rated more negative and arousing than CS- items and CS+ trials were related to increased SCRs and PDRs. These results suggest the presence of a learned association between the CS+ category and the possibility of a negative outcome, in this case an electrical shock. The psychophysiological data from the mixed valence localizer, in which fear anticipation responses were measured alongside reward anticipation and neutral trials, support the directions of the findings from the conditioning task.

The reaction time data we collected support the effectiveness of the task design we used for counter-conditioning. In order to create an appetitive US with enough incentive to counteract the initial aversive US, we decided to make the participants actively involved in obtaining the reward rather than passively, which was the case for the aversive US. This was achieved by using an MID task (Knutson et al., 2000). As reaction time data from the counter-conditioning task show, participants' responses become slower over time, however, for the targets in the CS+ trials participants seem to stay motivated to respond fast. Additionally, in the mixed valence localizer the motivation to respond was greater in trials where a reward could be received. Interestingly, even though participants were instructed that their responses would have no influence on the possibility of receiving a shock, participants also show shorter reaction times in CS+ trials in the conditioning phase, as they do for the trials where they could receive a shock in the mixed valence localizer task. This could be explained by a state of increased action preparedness, which is proposed to be induced in the anticipation of threat when responses are available (Gladwin, Hashemi, Van Ast, & Roelofs, 2016). Hashemi et al. (2019) found a similar decrease in response time in relation to physiological threat anticipatory responses.

After the counter-conditioning in one and extinction procedure in the other group, the differential conditioned fear response, that was formed during the conditioning phase, disappeared again. This suggests an effective reduction of the earlier present fear association towards the CS+ category. Again, this was the case for the subjective ratings and the SC and PD data. When looking at the SC data, the CRs seem to fade away faster in the counter-conditioning group than in the extinction group, in line with the hypothesis that counter-conditioning is more effective in reducing fear responses compared to extinction. However, this group difference in effectiveness is not found with regard to the subjective ratings. Neither is this difference found in the PD data, notwithstanding, the PDRs were found to extinguish rapidly in both groups. This ceiling effect in both groups could prevent observing between group differences in effectiveness.

If counter-conditioning and extinction paradigms have a different efficacy in preventing the return of conditioned fear responses, then we hypothesize a reduction in return of subjective and psychophysiological fear responses in the participants undergoing counter-conditioning compared with extinction. However, the data does not support this hypothesis as no group differences were found. These results suggest that the two paradigms have comparable efficacy. An explanation for the absence of a group difference could be that the interval of one day between sessions is not long enough to observe differences in the rate of fear recovery between the groups. Quirk (2002) suggests that fear responses recover gradually with the passing of time after extinction, with an interval of minimum ten days until full recovery in rats and only minor recovery after one day.

Findings from the episodic memory task show that participants in the counter-conditioning group have a better memory for the items they have seen during the first session than the participants in the extinction group. In line with previous studies (Patil et al., 2017; Dunsmoor et al., 2012; Dunsmoor et al., 2015), participants in the counter-conditioning group tend to have higher memory scores for items from the CS+ category than from the CS- category. This difference does not seem to be present in the extinction group. These results from the counter-conditioning group support the competition hypothesis as for both the conditioning and counter-conditioning phase, memory for the emotionally relevant stimulus category seems enhanced. This suggests both the existence of a fear and a reward association (Hamann, Ely, Grafton, & Kilts, 1999). Alternatively, the later formed reward association with the CS+ category may have caused a beneficial memory effect for items from both phases, inducing improved recollection of CS+ items in general (Hamann, 2001). This would explain why in the extinction group no CS+ related memory enhancement seems to be present.

Although data from the episodic memory test imply that both a fear and reward memory exist after counter-conditioning that compete for expression, results from the reaction time data point more in the direction of the enhanced extinction hypothesis to underlie counter-conditioning. As indicated by the conditioning, counter-conditioning and mixed valence localizer tasks, participants respond faster in trials where they could receive a shock or earn a reward, compared to neutral CS- trials. At the start of the second day, during the test for spontaneous recovery and reinstatement, participants respond similarly in CS+ and CS- trials. Moreover, this was the same in both groups, suggesting that similar mechanisms underlie counter-conditioning as well as extinction. In neither group, a fear or reward association with the CS+ category seemed to influence the response behaviour of the participants on the second day.

So far, the brain regions that have been found active in relation to the tasks are in line with the expected regions. In line with earlier studies using a task similar to the currently used category localizer (Dunsmoor et al., 2013; de Voogd et al., 2016a), we found activity in the fusiform gyrus. In the conditioning and mixed valence localizer task, the insular cortex was found to be involved in fear learning, as could be expected based on the role of the insula in monitoring emotions and anticipation of affective stimuli (Nicholson et al., 2016). More data might provide a more reliable indication of which subpart of the insula is involved in those tasks, since the subareas of the insular cortex are found to have distinguishable functionalities (Nelson et al., 2010). An area that was not found to be active during fear learning was the amygdala. This is unexpected, as based on studies naming the amygdala as crucial for acquiring conditioned fear associations (Quirk & Milad, 2012; Fragkaki, Thomaes, & Sijbrandij, 2016), it was expected to play a role in the current study as well. However, multiple other conditioning studies also do not report differential amygdala involvement during fear learning (Mechias, Etkin, & Kalisch, 2010; Bulganin et al., 2014; de Voogd et al., 2016a; Fullana et al., 2016). The proposed distinction between situations with certain and uncertain threats by Fox and Shackman (2019) could shed light on the absence of amygdala involvement. They propose that in situations with a certain threat the amygdala and BNST are involved, while in situations with an uncertain threat more cortical regions, like the insular cortex, are involved. This would indeed be in line with the current paradigm. Since we used a partial reinforcement of 50% of the trials, there was uncertainty of whether or not a shock would follow in the CS+ trials. We indeed found involvement of the insular cortex rather than the amygdala and BNST.

Several analyses are yet to be done when data collection is finished and the total sample size of 48 participants is reached. First, a representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008) will be conducted on the functional MRI data from the episodic memory test. As Dunsmoor et al. (2013) found, representations of CS+ category members encoded during conditioning are more similar than CS- category items. This analysis can help elucidate the mechanisms of counter-conditioning. If counter-conditioning is akin to extinction, a similar effect on the category representations is expected to be found as in the study of Dunsmoor et al. (2013). Another possibility is that due to the reward learning in the counter-conditioning phase, items previously seen in this phase also have increased similarity in representation. Yet another possibility is that this reward learning induces a reward association with the CS+ category that overwrites the fear association. In this case, a representational similarity analysis can support this hypothesis by showing that only items previously seen in the counter-conditioning phase have enhanced representational similarity. Second, alterations in functional connectivity will be assessed, especially to differentiate the hypothesis that extinction mechanisms are enhanced during counter-conditioning from the hypothesis that a reward association is formed that competes with the conditioned fear association. If the former were correct, similar to during extinction increased vmPFC-amygdala activity during extinction recall would be expected (Sierra-Mercado, Padilla-Coreano, & Quirk, 2011; Krabbe, Gründemann, & Lüthi, 2017; Dunsmoor et al., 2019) as a marker of the vmPFC playing a role in regulating the expression of the fear response. If the second hypothesis were correct, however, involvement of an amygdala-ventral striatum circuitry is expected, in line with the findings of Correia et al. (2016).

As mentioned above, a limitation of the used design is the short interval between the extinction or counter-conditioning task and the measurement of spontaneous recovery. Having a longer interval might have allowed more spontaneous recovery to take place (Quirk, 2002) and thereby allowing a better comparison of the effectiveness of counter-conditioning and extinction in preventing the return of this fear response. Another limitation of the current results is that they are based on a subset of the final participant sample. In this paper, preliminary results based on 20 participants are discussed. Our study has a final intended sample size of 48, as indicated during preregistration. More participants are needed to successfully find group level effects of conditioning on all psychophysiological outcome measurements and pinpoint the neurocognitive mechanisms of counter-conditioning. Moreover, a limitation of the current study design is that extinction and counter-conditioning immediately follow-up on the conditioning phase. Maren (2014) suggests that immediate extinction is less effective than extinction that takes place at least a day after conditioning. Nonetheless, preliminary data of this design, as shown in this paper, indicate neither spontaneous recovery nor reinstatement of the fear response. This suggests that the current design is strong enough to induce extinction of the fear response. This could be explained by the fact that the paper of Maren (2014) is focused on rats, in which very high stress levels are induced, which is not the case in human studies. Furthermore, since the interval between conditioning and counter-conditioning and extinction is this short, findings are less translatable to clinical situations, where the extinction-based therapies are likely to take place long after the initial formation of fearful associations. A final limitation is that, for pragmatic reasons, the length of the interval between the two sessions varied between participants. Although the second session was always on the day after the first session, the exact amount of time between them varied. If the effectiveness of the counter-conditioning procedure relies on consolidation dependent mechanisms, this might complicate the results.

In conclusion, the current study looks at the effectiveness and neurocognitive mechanisms of counter-conditioning as an alternative to extinction procedures. Preliminary results indicate the effectiveness of the study design in inducing both fear and reward learning. Tentative effects of the counter-conditioning procedure on the episodic memory tend to support the hypothesis that during counter-conditioning a reward association is established, which competes with the conditioned fear association for expression. If results from the final sample also support the effectiveness of a counter-conditioning procedure, then this underlines counter-conditioning as a promising alternative to extinction. Understanding counter-conditioning will provide neurocognitive insight into fear learning and regulation and may help to improve treatments for stress- and anxiety-related disorders.

References

Alberini, C. M., & LeDoux, J. E. (2013). Memory reconsolidation. *Current Biology, 23*(17), R746-R750.

Bouton, M. E. (1993). Context, time and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114*(1), 80–99.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry, 52*(10), 976–986.

Bouton, M. E., & Bolles, R. C. (1979). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology-Animal Behavior Processes, 5*(4), 368-378.

Brooks, D. C., Hale, B., Nelson, J. B., & Bouton, M. E. (1995). Reinstatement after counterconditioning. *Animal Learning & Behavior, 23*(4), 383-390.

Bulganin, L., Bach, D. R., & Wittemann, B. C. (2014). Prior fear conditioning and reward learning interact in fear and reward networks. *Frontiers in Behavioral Neuroscience, 8*:67.

Correia, S. S., McGrath, A. G., Lee, A., Graybiel, A. M., & Goosens, K. A. (2016). Amygdala ventral striatum circuit activation decreases long-term fear. *eLife,* 1–25.

Craske M. G., Liao B., Brown L., & Vervliet, B. (2012). Role of inhibition in exposure therapy. Journal of Experimental Psychopathology, 3(3), 322–345.

Craske, M. G., & Mystkowski, J. L. (2006). Exposure Therapy and Extinction: Clinical Studies. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Eds.), *Fear and learning: From basic processes to clinical implications* (pp. 217-233). Washington DC, US: American Psychological Association.

de Voogd, L. D., Fernandez, G., & Hermans, E. J. (2016a). Awake reactivation of emotional memory traces through hippocampal–neocortical interactions. *Neuroimage, 134,* 563-572.

de Voogd, L. D., Fernandez, G., & Hermans, E. J. (2016b). Disentangling the roles of arousal and amygdala activation in emotional declarative memory. *Social Cognitive and Affective Neuroscience*, *11*(9), 1471–1480.

de Voogd, L. D., Murray, Y. P., Barte, R. M., van der Heide, A., Fernandez, G., Doeller, C. F., & Hermans, E. J. (2019). The role of hippocampal spatial representations in contextualization and generalization of fear. *bioRxiv preprint.*

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2014). Aversive Learning Modulates Cortical Representations of Object Categories. *Cerebral Cortex, 24*(1), 2859-2872.

Dunsmoor, J. E., Kroes, M. C. W., Li, J., Daw, N. D., Simpson, H. B., & Phelps, E. A. (2019).
Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. *Jorunal of Neuroscience,* 2713-2718.

Dunsmoor, J. E., Kroes, M. C. W., Moscatelli, C. M., Evans, M. D., Davachi, L., & Phelps, E. A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour, 2,* 291-299.

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology, 89*(2), 300-305.

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: how humans generalize fear. *Trends in Cognitive Sciences, 19*(2), 73-77.

Dunsmoor, J. E., Murty, V. P., Davachi, L. & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature, 520*(7547), 345-348.

Esteban, O., Markiewicz, C., Blair, R., Moodie C., Isik, A. I., Aliaga, … & Kent, J., et al. (2018). fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI. *Nature Methods,* 16(1), 111-116.

Fox, A., & Shackman, A. J. (2019). The central extended amygdala in fear and anxiety: Closing the gap between mechanistic and neuroimaging research. *Neuroscience letters, 693*(S1), 58-67.

Fragkaki, I., Thomaes, K., & Sijbrandij, M. (2016). Posttraumatic stress disorder under ongoing threat: a review of neurobiological and neuroendocrine findings. *European Journal of Psychotraumatology,* 7:30915.

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Avila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry, 21* (4), 500-508.

Gladwin, T. E., Hashemi, M. M., Van Ast, V., & Roelofs, K. (2016). Ready and waiting: Freezing as active action preparation under threat. *Neuroscience Letters, 619*, 182-188.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., et al. (2013). The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage, Mapping the connectome, 80*, 105–24.

Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine, 44*, 162-167

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology, 91*(3), 186–93.

Greve, D. N., & Fischl, B. (2009). Accurate and Robust Brain Image Alignment Using Boundary-Based Registration. *NeuroImage, 48* (1), 63–72.

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Science, 5*(9), 394-400.

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*(3), 289-293.

Hashemi, M. M., Zhang, W., Kaldewaij, R., Koch, S. B. J., Jonker, R., Figner, … & Roelofs, K. (2019). Human defensive freezing is associated with acute threat coping, long term hair cortisol levels and trait anxiety. *BioRxiv.*

Hermans, E. J., Henckens, M. J. A. G., Roelofs, K., & Fernandez, G. (2013). Fear bradycardia and activation of the human periaqueductal grey. *NeuroImage, 66*, 278-287.

Hitchcock, J., & Davis, M. (1986). Lesions of the amygdala, but not of the cerebellum or red
nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience, 100*(1), 11–22.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. *Neuroimage, 16*(1), 217-240

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage, 17*(2), 825–841.

Kaag, A., Schluter, R. S., Karel, P., Homberg, J., van den Brink, W., Reneman, L., & van Wingen, G. A. (2016). Aversive Counterconditioning Attenuates Reward Signaling in the Ventral Striatum. *Frontiers in Human Neuroscience, 10*:418.

Kang, S., Vervliet, B., Engelhard, I. M., van Dis, E. A. M., & Hagenaars, M. A. (2018). Reduced return of threat expectancy after counterconditioning versus extinction. *Behaviour Research and Therapy, 108*, 78-84.

Karel, P., Almacellas-Barbanoj, A., Prijn, J., Kaag, A., Reneman, L., Verheij, M., & Homberg, J. (2019). Appetitive to aversive counter-conditioning as intervention to reduce reinstatement of reward-seeking behavior: the role of the serotonin transporter *Addiction Biology.*

Kerkhof, I., Vansteenwegen, D., Baeyens, F., & Hermans, D. (2011). Counterconditioning An Effective Technique for Changing Conditioned Preferences. *Experimental Psychology, 58*(1), 31-38.

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *Neuroimage, 12*(1), 20-27.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Krabbe, S., Gründemann, J., & Lüthi, A. (2017) Amygdala inhibitory circuits regulate associative fear. *Biological Psychiatry, 83*(10), 800-809.

Kroes, M. C. W., Dunsmoor, J. E., Lin, Q., Evans, M., & Phelps, E. A. (2017). A reminder before extinction strengthens episodic memory via reconsolidation but fails to disrupt generalized threat responses. *Scientific reports,* 10858.

LeDoux, J. E., Ruggiero, D. A., & Reis, D. J. (1985). Projections to the subcortical forebrain from anatomically defined regions of the medial geniculate body in the rat. *Journal of Comparative Neurology, 242*(2), 182–213.

LeDoux, J. E., Sakaguchi, A., & Reis, D. J. (1984). Subcortical efferent projections of the medial geniculate nucleus mediate emotional responses conditioned to acoustic stimuli. *Journal of Neuroscience, 4*(3), 683–698.

Lucas, K., Luck, C. C., & Lipp, O. V. (2018). Novelty-facilitated extinction and the reinstatement of conditional human fear. *Behaviour research and therapy, 109*, 68-74.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58,* 25-45.

Mechias, M. L., Etkin, A., & Kalisch, R. (2010). A meta-analysis of instructed fear studies: Implications for conscious appraisal of threat. *Neuroimage, 49*(2), 1760-1768.

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.

Nelson, S. M., Dosenbach, N. U. F., Cohen, A. L., Wheeler, M. E., Schlaggar, B. L., & Petersen, S. E. (2010). Role of the anterior insula in task-level control and focal attention. Brain Structure & Function, 214(5-6), 669-680.

Nicholson, A. A., Sapru, I., Densmore, M., Frewen, P. A., Neufeld, R. W. J., Theberge, J., McKinnon, M. C., & Lanius, R. A. (2016). Unique insula subregion resting-state functional connectivity with amygdala complexes in posttraumatic stress disorder and its dissociative subtype. *Psychiatry Research-Neuroimaging, 250*, 61-72.

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & Memory, 24*(1), 65-69.

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

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron, 43*(6), 897–905.

Posse, S., Wiese, S., Gembris, D., Mathiak, K., Kessler, C., Grosse-Ruyken, M., … & Kiselev, V. G. (1999). Enhancement of BOLD Contrast Sensitivity by Single-Shot Multi-Echo Functional MR Imaging. *Magnetic Resonance in Medicine 42*(1), 87–97.

Reinders, A. A. T. S., Den Boer, J. A., & Büchel, C. (2005). The robustness of perception. *European Journal of Neuroscience, 22*(2), 524-530.

Rescorla, R. A., & Heth, D. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology-Animal Behavioral Processes, 104*(1), 88–96.

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. Prokasy, Classical conditioning II (pp. 64-99). New York, NY: Appleton-Century-Crofts

Schiller, D., Monfils, M. H., Raio, C., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*(7277), 49-U51.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia. *Cerebral Cortex, 10*(3), 272-283.

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage, 20*, 114-124.

Sierra-Mercado, D., Padilla-Coreano, N., & Quirk, G. J. (2011). Dissociable Roles of Prelimbic and Infralimbic Cortices, Ventral Hippocampus, and Basolateral Amygdala in the Expression and Extinction of Conditioned Fear. *Neuropsychopharmacology, 36*(2), 529-538.

Tunstall, B. J., Verendeev, A., & Kearns, D. N. (2012). A comparison of therapies for the treatment of drug cues: Counterconditioning vs. extinction in male rats. *Experimental and Clinical Psychopharmacology, 20*(6), 447-453.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: state of art. *Annual Review of Clinical Psychology, 9,* 215-248.

Quirk, G. J. (2002). Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery. *Learning & Memory, 9*(6), 402–407.