Tackling fear: Beyond associative memory activation as the only determinant of fear responding

Yannick Boddez[1,2,3], Agnes Moors[1, 4, 5], Gaëtan Mertens[6], & Jan De Houwer[1]

[1]Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

[2]Centre for the Psychology of Learning and Experimental Psychopathology, KU Leuven, Leuven, Belgium

[3]Department of Clinical Psychology and Experimental Psychopathology, University of Groningen, Groningen, The Netherlands

[4]Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Belgium

[5]Centre for Social and Cultural Psychology, KU Leuven, Belgium

[6]Department of Clinical Psychology, Utrecht University, Utrecht, the Netherlands

Correspondence concerning this article can be addressed to Yannick Boddez, Department of Experimental Clinical and Health Psychology, Henri Dunantlaan 2, 9000 Ghent, Belgium.
E-mail: yannick.boddez@ugent.be

Preparation of this article was supported by the Research Foundation - Flanders (FWO) (G022313N, G073317N), the Research Fund of KU Leuven (C14/17/047), Ghent University Methusalem Grant BOF16/MET_V/002, and the Dutch Fund for Scientific Research (NWO Vici) (453-15-005). Yannick Boddez is grateful to Tom Beckers, Peter de Jong, Mathilde Descheemaeker, Dirk Hermans, Sara Scheveneels, and Bram Vervliet for interesting discussions about the topics at hand.

Abstract

For decades already, the human fear conditioning paradigm has been used to study and develop treatments for anxiety disorders. This research is guided by theoretical assumptions that, in some cases indirectly, stem from the tradition of association formation models (e.g., the Rescorla-Wagner model). We argue that one of these assumptions – fear responding as a monotonic function of the associative activation of aversive memory representations – restricts the types of treatment that the research community currently considers. We discuss the importance of this assumption in the context of research on extinction-enhancing and reconsolidation interference techniques. While acknowledging the merit of this research, we argue that unstrapping the straitjacket of this assumption can lead to exploring new directions for utilizing fear conditioning procedures in treatment research. We discuss two determinants of fear responding other than associative memory activation. First, fear responding might also depend on relational information. Second, a recent goal-directed emotion theory suggests that goals might be the primary determinant of the response pattern characterized as fear.

Tackling fear: Beyond associative memory activation as the only determinant of fear responding

Anxiety disorders are linked with substantial impairments in quality of life and with high economic burden. The literature states a prevalence of up to 14% (Wittchen et al., 2011), making it a high-priority target for research. It is commonly assumed that relations between stimuli are crucial in the etiology of anxiety and its disorders (Craske, Hermans, & Vansteenwegen, 2006; Scheveneels, Boddez, & Hermans, 2019). For example, someone may fear certain chemicals because of their (presumed) relation with cancer or a child may fear monsters because it believes that they eat people. In the laboratory, the fear conditioning paradigm is often used to install fears (Field, 2006; Mineka & Zinbarg, 2006). Typically, a conditional stimulus (CS) and an unconditional stimulus (US) are paired and the effect of this regularity on responding to the CS is assessed (De Houwer, Barnes-Holmes, & Moors, 2013). Relying on layman terms, the conditioned responses are typically called “fear responses” (e.g., skin conductance responding, avoidance, verbal reports of feeling fearful) and the US is termed aversive (e.g., an electric shock). This seemingly simple paradigm has become a major force in clinical psychology, psychiatry, affective neuroscience, pharmacology, and genetics (Beckers, Krypotos, Boddez, Effting, & Kindt, 2013). The interest in this paradigm stems from the expectation that fear conditioning studies will provide insight in how to optimize the treatment of anxiety disorders (Scheveneels et al., 2019). Although several meta-analyses demonstrate large effect sizes and high response rates for current treatment (e.g., Hofmann & Smits, 2008; Loerinc et al., 2015), these effects are not always maintained in the long run, leaving room for the further enhancement of clinical treatment (Craske & Mystkowski, 2006).

The causes of the conditioned responses observed in the human fear conditioning paradigm can be considered at the level of environmental events and at the level of mental events. At the level of environmental events, one could state that the spatio-temporal relation between the stimuli (i.e., of the CS and the US) is causing the fear responses (De Houwer, 2011; De Houwer et al., 2013). However, fear conditioning researchers typically do not settle for explanations in terms of observables like regularities between stimuli (Stroebe, 2018). As such, fear conditioning researchers often make additional assumptions about black box mental mechanisms that mediate the relationship between the pairing of the stimuli and the fear responses. A challenge for the mental level of analysis, however, is that multiple mental explanations are compatible with any finite set of observable data (Garcia-Marques & Ferreira, 2011; Goodman, 1955; Lieder & Griffiths, 2019). Despite the multiple candidate mental explanations, very few types of mental explanations received attention in the literature on fear conditioning. More specifically, the currently dominant approach originates from association formation models (e.g., Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981). In what follows, we discuss these models and the way in which they have been adopted in fear conditioning research. More precisely, we discuss their impact on research on extinction-enhancing and on reconsolidation interference techniques. Afterwards, we introduce a theoretical perspective which invokes relational information and goals as determinants of conditioned fear and which may lead to new avenues for utilizing fear conditioning procedures in treatment research.

Association formation models

Association formation models hold that fear conditioning effects are mediated by associations between the memory representations of CSs and USs. An association is typically conceived of as an unqualified link that transmits activation from one representation to another, analogous to the way in which a strip of copper wire conducts electricity (Boddez, De Houwer, & Beckers, 2017; Dacey, 2018; Haselgrove, 2016; Mitchell, De Houwer, & Lovibond, 2009). The nature of the representations often remains undefined but can be understood as the mental imprint of memoranda, comparable to the physical imprint of memoranda (e.g., as a drawing in clay or as photographs; Holland, 1993; Skinner, 1977). Association formation models hold that once an association has been formed, the presentation of the CS results in the activation of its mental representation, which in turn produces an increase in the activation of the representation of the US.

We would like to draw attention to a number of assumptions that in part have their origin in the tradition of these association formation models. Learning theorists have long acknowledged the importance of differentiating between associative strength in memory (i.e., the strength with which the US representation is activated by associations) and conditioned responding (Bouton and Moody, 2004). Nevertheless, in practice, predictions about conditioned responding have typically been made on the basis of the monotonicity assumption. This assumption entails that there is a monotonic relationship between associative strength in memory and the strength of responding (for an early discussion and critique, see Miller et al., 1995). For instance, Rescorla and Wagner (1972; p. 77) provided a formula to calculate associative strength based on training history and then treated associative strength as the only determinant of behavior: "it will generally be sufficient simply to assume that the mappings of Vs [associative strength] into magnitude or probability of conditioned responding preserves their ordering". Other influences on responding (e.g., whether the response has expected utility or has been emitted frequently and/or recently; Moors, 2016) are not considered or, at least, not included in the formalized models. Although the monotonicity assumption was initially a quick fix for models that did not have the aim to deal with the complex topic of behavioral expression (yet), it has rarely been questioned afterwards (for exceptions see Miller et al., 1995; Rescorla, 2001).

Two further assumptions support the monotonicity assumption, in that they provide a more detailed answer to what gets activated in memory and therefore determines responding. First, the abstraction assumption holds that what is learned from the repetition of a similar learning event over the course of various trials is summarized in a single association. More precisely, experiencing CS-US pairings on a series of trials is assumed to result in an increase of associative strength between the representation of the CS and the representation of the US. This is nontrivial because one could also come up with an associative learning model in which information provided in different trials is stored separately (i.e., episodically; e.g., Holland, 1993; Dunsmoor & Kroes, 2019; Schmidt, De Houwer, & Rothermund, 2016). To clarify with a daily-life example: If your cat scratches you on three different occasions, association formation models assume that the associative strength between the representation of the cat and the representation of scratching is updated rather than that these three episodes with their specific details would be remembered separately. It may be of further interest that, strictly speaking, the term "remembering" can only be used in a metaphorical sense from the perspective of association formation models, because this term implies reference to the past (i.e., to past episodes) whereas the mere activation of a representation is silent with respect to referencing the past or the future (Jozefowiez, 2018). We will leave this matter for now but return to it in the section on relational information.

A second assumption which supports the monotonicity assumption is the reproduction assumption, which implies that reactivation, or remembering, is reproductive rather than reconstructive in nature. That is, the representations are presumably stored as complete and ready-to-use entities that can be accurately recollected via associations. For example, the taste of a madeleine would activate representations of what previously went together with eating it (e.g., drinking tea or the presence of one's partner). This is again nontrivial, because there are also arguments that remembering is an ad hoc reconstructive process that is influenced by a multitude of factors, such as beliefs and motivations (Loftus, 1975, 2005; Loftus & Hoffman, 1989; Loftus, Miller, & Burns, 1978; Shaw & Porter, 2015).

Part of the fear conditioning community interested in treatment optimization has adopted these assumptions. This is most clearly recognized in the idea that tackling a single association stored in memory is the (only) way to go if one wants to remediate fear responding. Researchers seem to agree on this general aim, even if they differ in opinion on how this is done best. For example, proponents of inhibitory learning theory (e.g., Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014) argue that tackling associations is best achieved by installing a competing inhibitory association, whereas proponents of reconsolidation interference techniques (e.g., Kindt, 2018) aim to remove/update the original CS-US association altogether. Khalaf et al. (2018, p 1239) summarize this difference in opinion as follows: “Whether fear attenuation is mediated by inhibition of the original memory trace of fear with a new memory trace of safety or by updating of the original fear trace toward safety has been a long-standing question in neuroscience and psychology alike”. We discuss the essentials of both approaches and their adherence to the discussed assumptions below.

Glossary

An **association** is a mental link that transmits activation between representations in the way a piece of copper wire would.

The **monotonicity assumption** entails that there is a monotonic relationship between associative strength with which the US is activated and the strength of responding.

The **abstraction assumption** entails that a repetition of learning events results in a strengthening of the association between the CS and US representation rather than that the various events with their episodic details would be stored separately.

The **reproduction assumption** entails that reactivation via associations, or remembering, is reproductive rather than reconstructive in nature.

A **proposition** is a mental structure that contains information on how representations are related (e.g., A *predicts* B; A *causes* B; A *prevents* B).

The **goal-directed account** holds that goal-directed processes are a crucial determinant of conditioned fear.

Inhibitory learning theory

Inhibitory learning theory was developed in the context of research that makes use of a laboratory model of exposure therapy, namely the extinction procedure (Bouton, 2002; Scheveneels, Boddez, Vervliet, & Hermans, 2016). At the observable level, the extinction procedure entails CS-only presentations, resulting in a decrease of the fear responses that were previously established by pairing the CS with the US. At the mental level, inhibitory learning theory holds that an inhibitory CS-US association is acquired during these CS-only trials. This inhibitory association (characterized as a negative value; e.g., -0.5) is supposed to counteract the original association which drives the fear response (characterized as a positive value; e.g., +0.5), resulting in low to no fear responding when both these associations are activated (because of a summed associative value close to 0; Figure 1). In other words, when the US representation becomes activated by the excitatory association — and therefore comes to mind — there will be conditioned responding. In contrast, when this activation is counteracted by the inhibitory association — and the US therefore does not come to mind — there will be no responding. So, in line with the monotonicity assumption, the absence of fear responding is explained by an inactivated US representation. For that reason, one could use the term "non-permanent amnesia" to characterize the inhibitory learning mechanism. If the representation of the US remains completely inactive, it indeed seems hard to imagine that there could be any thinking or remembrance of the US.

It is of note that advocates of inhibitory learning theory sometimes use "inhibitory association" in a less strict sense. For example, it may remain unspecified which aspect of the US representation is inactivated by the inhibitory link (e.g., all of it or only its motor component; Holland, 1993). Some authors also use the term "inhibitory association" to refer to propositional beliefs concerned with safety (e.g., "soap prevents illness"). However, as we will discuss below, there are crucial differences between associations and propositional beliefs. Relatedly, acquisition of an inhibitory link is sometimes equated with acquisition of a "safety memory". Nonetheless, according to a strict interpretation of inhibitory learning theory, the US representation just gets activated or not without episodic remembrance of unreinforced trials. In the context of this manuscript, we will conceptualize inhibitory learning theory in the strict way in which it has originally been conceptualized to account for extinction learning in the tradition of association formation models (e.g., Vervliet, Craske, & Hermans, 2013).

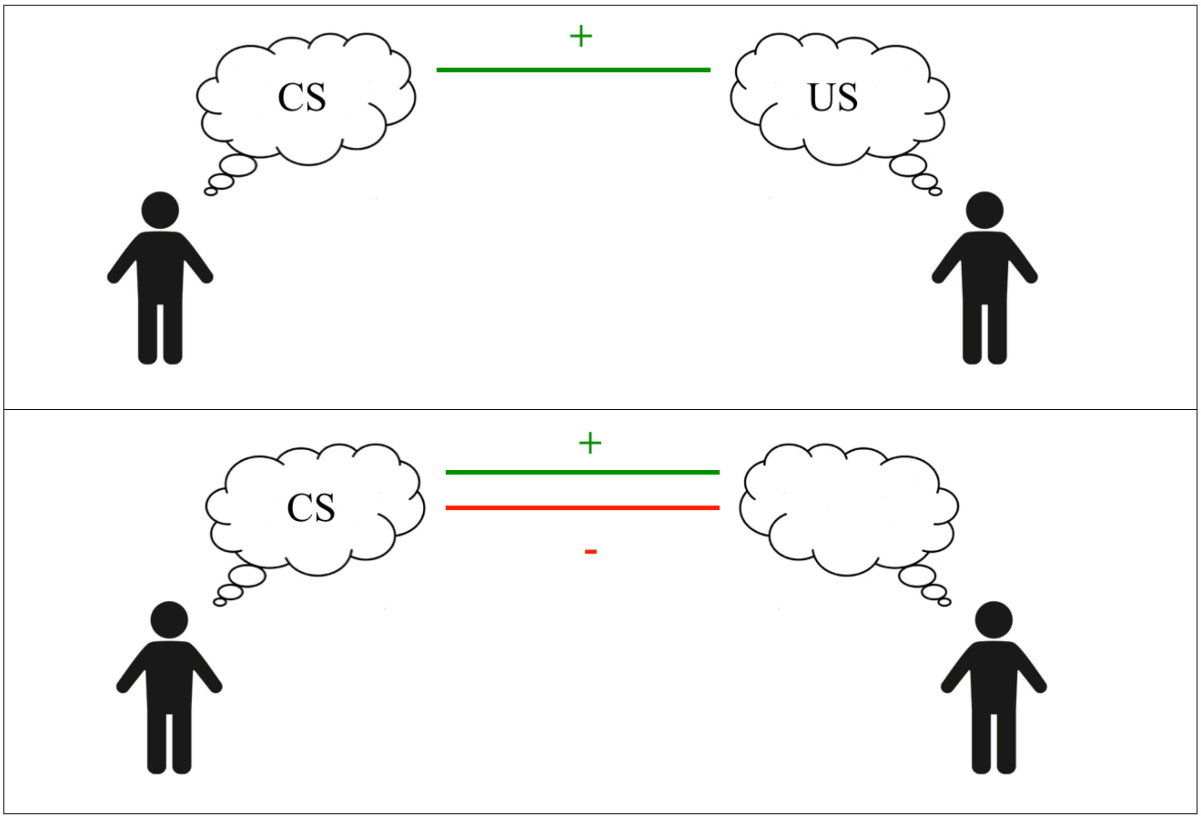

Figure 1. Schematic presentation of inhibitory learning theory. During CS-US trials (upper panel), an excitatory association is formed (green line; +). During the CS-only trials (lower panel), an additional inhibitory association (red line, -) is formed. Excitatory and inhibitory associations counteract each other so that the presentation of the CS no longer results in the activation of the US representation after extinction. If the US representation remains inactive, then there can be no thinking or remembrance of it. To account for renewal effects, it is assumed that the inhibitory association is only effective in the extinction context.

The strong implications of adhering to the monotonicity assumption can be explored by analyzing the renewal effect, which propelled the popularity of inhibitory learning theory.

If a CS and a US first co-occur in context A and the CS is then successfully extinguished in context B, the CS typically comes to elicit renewed fear responding when it is presented in context A again or in a novel context C (Bouton, 2002; Vervliet, Baeyens,Van den Bergh, & Hermans, 2013). Inhibitory learning theory explains renewal by assuming that retrieval of the new inhibitory association is modulated by context and that it is strongest in the context in which it was acquired. So, participants are said to display renewed fear responding because the excitatory association brings the US to mind, while the inhibitory association remains inactive (Figure 1). Note that this implies that participants would fail to consider what happened during the extinction phase right before the test phase (sometimes as recent as the duration of an intertrial interval of 10 seconds; Scheveneels, Boddez, De Ceulaer, & Hermans, 2019). An alternative (and perhaps more intuitive) explanation for the renewal effect is that participants consider both the acquisition and extinction trials and infer it to be likely that the CS at test will again be followed by the US given the contextual information (Boddez, Baeyens, Hermans, & Beckers, 2011; Raes, De Houwer, Verschuere, & De Raedt, 2011). It is further of note that the abstraction and reproduction assumptions can also be recognized in the inhibitory learning account of renewal. What is learned over the course of acquisition trials is abstracted in a single excitatory association and what is learned over the course of extinction trials is abstracted in a single inhibitory association. The interplay between the excitatory and the inhibitory association subsequently determines whether or not the US representation will be reproduced and thus whether or not there will be fear responding.

The strong implications of adhering to the monotonicity assumption are also revealed when using inhibitory learning theory to account for the subjective experience (i.e., the covert responses) of patients partaking in exposure therapy. Consider a patient who received successful exposure treatment for flying phobia. Although he still thinks about dying in a plane crash (US), he nonetheless boards the plane because he managed to persist during therapy and believes his family deserves a holiday (cf. Williams, Kinney, & Falbo, 1989). The subjective experience of this patient would be at odds with a strict interpretation of inhibitory learning theory. Strictly speaking, this theory holds that during therapy an inhibitory association between being on an airplane and dying in a plane crash would have been learned, which would cause the patient to not or barely think about dying in a plane crash when being on the plane (i.e., non-permanent amnesia). Later in this manuscript, we will discuss how a goal-directed model may provide a more plausible account of this type of subjective experience.

Another example of the monotonicity assumption can be found in the treatment of people suffering from dog phobia (to which the first author of this manuscript regularly assists). While most patients manage to approach dogs at the end of the exposure session(s), hitherto none of them seemed to have temporarily forgotten whatever outcome they were fearing from doing so, as evidenced by enduring verbal reports of doubt and obvious approach-avoidance response conflict. If an inhibitory association does not cancel thoughts about an aversive outcome but the patient still performs the desired behavior (i.e., approaching the dog), then therapy may not (only) rely on the successful installation of an inhibitory association. That is not to say, of course, that, therapy may never result in an eventual reduction of such thoughts. Whether and under which conditions that happens is an empirical question. Nevertheless, fear-reducing treatments are often a-priori considered to be interventions that either “weaken the fear memory or augment the extinction memory” (Huff, Hernandez, Blanding, & Labar, 2014; p. 835; Wellman et al., 2014). Note that the use of the singular form in this quote (i.e., the fear memory and the extinction memory) again suggests adherence to the abstraction and reproduction assumptions. Indeed, it suggests that various events are abstracted in a single memory (trace) which determines fear responding upon its reproduction.

Reconsolidation theory

Reconsolidation interference techniques have been promoted as more promising than extinction-based treatment, because they would allow to completely erase the associations driving fear responding instead of only temporarily counteracting them by installing inhibitory associations. As such, reconsolidation interference techniques might be used to overcome the risk of later relapse (e.g., Kindt, 2018).

At the observable level, reconsolidation interference techniques are procedures that involve an intervention in the timeframe close to the presentation of a fear-conditioned CS, which, when successful, lead to a reduction of (return of) fear responding. For example, it has been suggested that administration of propranolol (i.e., a β-adrenergic receptor antagonist) before presenting a fear-conditioned CS can reduce fear responding and its return (Kindt, Soeter, & Vervliet, 2009; but see Schroyens, Beckers, & Kindt, 2017). Similarly, playing Tetris after presentation of a CS has been reported to reduce fear responding (de Voogd, Hermans, & Phelps, 2018; James et al., 2015; but see Chalkia, Vanaken, Fonteyne, & Beckers, 2019). Finally, inserting a break after a single CS-only presentation before continuing with regular extinction training has been reported to reduce return of fear after extinction training (Monfils, Cowansage, Klann, & Ledoux, 2009; Schiller et al. 2010; but see Luyten & Beckers, 2017).

At the mental level, reconsolidation theory states that presenting the CS reactivates the corresponding associations and representations. A theoretical assumption is that reactivated memories have to be placed back in long-term memory after use, because they would otherwise perish. The intervention (e.g., administering propranolol) is supposed to prevent this placing back, which would effectively cause memories to perish and become unusable

(Beckers & Kindt, 2017; Elsey, Van Ast, & Kindt, 2018). To fully grasp this mental process theory, it might be helpful to appreciate that “reconsolidation” is a metaphor that appeals to our idea that certain entities, such as ice, can melt and then solidify again. So, the reactivation trial induces the melting of the ice and the reconsolidation interference technique aims to prevent reconsolidation of the ice after having been melted. The crucial aspect of the metaphor is that once the ice is in a melted state (i.e., water), it is vulnerable to evaporation. Similarly, it is assumed that reconsolidation interference techniques keep memories in a non-consolidated state, which would make them vulnerable and eventually evaporate.

A related account holds that reactivating a memory opens a time-window for memory update. So, rather than making the memory trace evaporate, it would be updated. For example, a single CS-only presentation after acquisition would provide a time-window during which the original CS-US association can be changed. If extinction training is administered during this time window, the association would be permanently updated to reflect the new reality that the CS is no longer followed by the US[1] (rather than that a new competing association would be formed as assumed in inhibitory learning theory; De Beukelaar, Woolley, Alaerts, Swinnen, & Wenderoth, 2016; Lee, 2009; also see Orederu & Schiller, 2018; Siegel & Weinberger, 2012).

Reconsolidation theory has not only adopted the general association concept, but the more specific monotonicity assumption as well. The adherence to this assumption most clearly reveals itself in the reverse inference that the absence of fear responding implies the absence of US representations in memory. An illustration of this reverse inference can be

[1] Note that this matches what is assumed in certain association formation models. For example, according to the Rescorla-Wagner model (1972), the CS-US association reduces in strength because of CS-only presentations. However, association formation models are silent with respect to the need to reactivate memory by means of a preceding CS-only trial. At the same time, the consolidation account does not make insightful why the updating would not happen with regular extinction parameters: Why would early extinction trials not open the time-window during which the CS-US association can be updated to reflect what is experienced during later extinction trials (Monfils & Holmes, 2018)?

found in the interpretation of the results of experiments in which subjects were asked whether they remembered or could report the CS-US contingencies after their fear responses had diminished. In an influential study (Soeter & Kindt, 2010), it was demonstrated that propranolol eliminated differential startle responding to the reinforced stimulus (i.e., the CS+) as compared to the nonreinforced control stimulus (i.e., the CS-). Interestingly, however, the manipulation did not affect skin conductance responding and US-expectancy (which is in line with animal findings demonstrating that such procedures do not bring the animal back to a naïve state; Gisquet-Verrier & Riccio, 2018). Although the intact US-expectancies demonstrate that participants remember the co-occurrence of CS and US, the idea of memory erasure was not abandoned. Rather, it was suggested that emotional memory had been erased, while declarative memory was kept intact (Kindt et al., 2009). In other words, multiple memory systems were invoked to save the monotonicity assumption. This demonstrates the dominance of this assumption: If activation of memory representations is the only determinant of responding, then there is simply no other option than to invoke multiple memory representations or systems in the case of response dissociations. Although invoking a dissociation between memory systems certainly provide a straightforward way to explain response dissociations, it might, however, be worth considering other explanations for such dissociations. For example, below, we will discuss that a goal-directed perspective can explain response dissociations by linking different responses to different goals, so without the need to invoke a dissociation between memory systems.

Moreover, the emotion version of memory erasure theory comes with new challenges, beginning with the lack of useful criteria to decide on what constitutes an emotional memory (Duffy, 1941; Russell, 2003). In response to this, reconsolidation interference researchers may of course equate emotional *memory* with the *responses* that are affected by the experimental manipulations (e.g., startle responding). However, this would go against the spirit of the proposal that reconsolidation techniques not only change responses but also the underlying emotional memory (Kindt, 2018).

The monotonicity assumption goes hand in hand with the abstraction and reproduction assumption in the reconsolidation literature as well. The acquisition phase of most reconsolidation experiments comprises multiple fear conditioning trials. However, reconsolidation theory does not posit a need to tackle the episodic memory of each trial separately, presumably because of the assumption that the separate trials get summarized (i.e., abstracted) in a single (emotional) memory trace[2]. The reliance on retrieval cues (e.g., of a single CS presentation) to reactivate memory is also in line with the reproduction assumption which holds that memories are stored as complete and ready-to-use entities that are recollected upon reactivation and can be specifically targeted.

From challenges for association formation models towards a goal-directed model

We argued that an important part of the fear conditioning community adopted the association concept and the monotonicity assumption, which is recognized in the dominant idea that tackling a single association stored in memory is the (only) way to go if one wants to remediate fear responding. We already made clear that this is a strong assumption but will now go a step further and argue that it leaves essential determinants of fear responding uncovered as well. First, fear responding does not (only) depend on activation that spreads via unqualified CS-US associations, but also on information about the nature of the relation between CS and US, suggesting the operation of a propositional mechanism. Second, a recent emotion theory suggests that goal-directed processes might be a crucial determinant of fear responses (Moors, Boddez, & De Houwer, 2017). After fleshing out both points, we put forward an alternative view of conditioned fear as goal-directed behavior. We will argue that

[2] Alternative conceptualizations are possible though. For example, reconsolidation theorists could argue that erasure works according to a Wikipedia analogy: Erasure of a representation makes that this representation disappears from all episodic memories that include it. In terms of the workings of Wikipedia: Deleting a page makes that all links towards this page become dysfunctional.

such an extended theoretical approach may inspire new ways for utilizing fear conditioning procedures in treatment research.

Relational information as a determinant of fear

Associations can differ on only one variable: the strength of the association. Variations in the number of CS-US co-occurrences or the statistical contingency between the CS and US presentations can affect the strength of associations (Holland, 1993), but variations in the type of relation (e.g., referential, predictive, or causal) between CS and US cannot be coded (Mitchell et al., 2009). This is nontrivial because a simple associative architecture therefore does not allow to differentiate remembering a US from predicting a US (Baeyens, Díaz, & Riuz, 2005; Baeyens et al., 1988; Jozefowiez, 2018). Based on what the mental construct of an association allows, Proust could not have known whether what the tea-soaked Madeleine brought to mind was in the past (i.e., was a flashback) or was about to happen in the near future (i.e., was a flashforward). In a recent paper, Jozefowiez (2018; p. 23) summarized this challenge for the association concept as follows: "[…] , let's start with the obvious: remembering the past and predicting the future are two different cognitive processes. A cognitive event cannot be both the retrieved memory of a past event and, at the same time, the expectation of a future event. Otherwise, you would not be able to tell the future from the past as the same mental event could be either one of them (more likely, you would not even have a concept of past and future). Obviously, this is not the case [...]". Translating this problem to the topic at hand, one may remember that a CS was followed by a US without predicting that the US will follow the CS on the current occasion and therefore without generating anticipatory (fear) responding (Zenses, Beckers, & Boddez, 2018). For example, a war veteran who stumbles upon his old gas mask, may remember a chemical attack experienced in the past without entertaining an expectancy of an imminent chemical attack in the future – and without experiencing the accompanying fear for such an attack. Conversely, one may also predict the onset of an aversive event without previous pairings and thus without a remembrance of those pairings (e.g., as is the case when the prediction is based on an inference; Boddez, Bennett, Van Esch, & Beckers, 2017). Hence, it seems essential to differentiate between the mere activation of aversive US representations by associations, on the one hand, and the predictions that drive anticipatory fear responding, on the other hand (Jozefowiez, 2018).

More generally, associations cannot capture the relational information that characterizes propositional beliefs such as "the presence of guard dogs *predicts* vs. *prevents* trouble" and "synthetic soap *predicts* vs. *prevents* illness". Nonetheless, propositional beliefs concerning the presence of danger are exactly the mental events in which the fear conditioning community is interested (e.g., Craske et al., 2014; Pittig, Treanor, LeBeau, & Craske, 2018; Pittig, van den Bergh, & Vervliet, 2016; Weisman & Rodebaugh, 2018). To deal with this challenge, theorists have suggested that the mind contains both associative and propositional representations. Associations would then underlie more complex propositions that contain additional relational information (Gawronski & Bodenhausen, 2014; McLaren et al., 2014). For example, the representation of guard dogs activating the representation of trouble would somehow give rise to the proposition "guard dogs predict trouble." Although this makes sense intuitively, the question remains how the organism can know which type of relation binds the elements: Do guard dogs predict trouble, cause trouble, prevent trouble, or is there still another relation at play (Moors, 2014)? In addition, it is impossible to determine the direction of the relation between the two elements: Does the presence of guard dogs predict trouble or does the presence of trouble predict the presence of guard dogs (Hummel & Holyoak, 2003)? These are challenges that need to be addressed if one takes the stance that experience is represented in both an associative and a propositional way.

Single-process propositional theorists have addressed these challenges in a different way by seeing how far one can get without assuming any role for simple unqualified associations between stimuli. Instead, they hypothesize that the origin of fear responses uniquely lies in propositional beliefs concerning how stimuli are related. Such propositions might stem from CS-US pairings, instructions, or inferences (for detailed discussions see Boddez et al., 2017; De Houwer, 2020; Mitchell et al., 2009). Given that propositions are mental structures that contain relational information (e.g., guard dogs predict trouble; guard dogs cause trouble; guard dogs prevent trouble), this solution provides one way to circumvent the above-mentioned limitation that associations cannot code variations in the type of relation.

A goal-directed process as determinant of fear

We argued that fear responding depends on the nature of the relation between CS and US rather than on the mere activation of a US representation, suggesting the involvement of a propositional mechanism. However, approaches invoking propositions still have an important limitation in common with association formation models. There is still a gap between entertaining propositional beliefs or activating US representations via associations, on the one hand, and displaying a fear response in the form of physiological (e.g., a skin conductance response) and specific motor actions (e.g., running away), on the other hand. Neither of these theoretical approaches has managed to bridge this gap. The criticism that a "model that can explain all but conditioned behavior, is lacking something quintessential" (Baeyens, Vansteenwegen, & Hermans, 2009; p. 199) may therefore apply to both classes of models alike. To explain the occurrence of conditional responses (CR), learning theorists have invoked a transfer from unconditional responses (UR) to the CS, but it has long been understood that this principle goes astray, because responses to the US and CS can differ substantially. For example, a mouse jumps and screams when receiving an electric shock but will freeze and keep quiet when confronted with a CS that signals this shock. There is also no evidence that conditioned responses can be consistently described as compensatory (i.e., opposing the UR; Carter & Tiffany, 1999; Moors, 2017; Rescorla, 1988).

We will now turn to a recent emotion theory (Moors, 2017; Moors et al., 2017; Moors & Fischer, 2019) to address the question how the form of fear responses comes about. More precisely, we will introduce a goal-directed account of conditioned fear.

In common sense explanations of emotional responses, physiological responses and motor actions are said to be caused by emotions (e.g., sweating is caused by fear or aggressive behavior is caused by anger) – end of story. However, according to most contemporary emotion theories, such physiological responses and actions are not caused by, but rather part of, emotional episodes (Moors, 2009). This perspective views emotion as a compound of action tendencies, physiological responses, motor action, and subjective feelings. Action tendencies are a primary building blocks of the emotional episode. An action tendency can be seen as a goal or inclination to act (e.g., to seek safety) and can or cannot result in an eventual bodily action (e.g., running away). Physiological responses further serve to prepare and support this potential action. For example, if one has the tendency to seek safety, the body is getting ready for that challenge, which happens to also produce sweating. Verbalizations, such as subjective fear reports, are assumed to result from a process in which action (tendencies) and physiological responses are integrated and centrally represented as feelings (in nonverbal form) after which they are categorized and labeled with emotion terms (e.g., fearful or sad; Fanselow & Pennington, 2018; Scherer & Moors, 2019).

Emotion theories differ with respect to what sets off the action tendency and with it the emotional episode. For example, according to a certain type of affect program theories (Matsumoto & Ekman, 2009), the perceptual features of the stimulus give rise to the action tendency in a hardwired way (e.g., the sight of a snake and the sensation of an electric shock – or CSs related to them - spur the tendency to seek safety). Appraisal theories, in their turn, hold that an evaluation (i.e., appraisal) of a stimulus (e.g., as a signal for impending danger which is difficult to control) determines the action tendency (Moors, Ellsworth, Scherer, & Fijda, 2013). In contrast, the goal-directed perspective that we propagate here suggest that an evaluation of the expected utilities of one or more action options determines the action tendency that follows. The expected utility of an action options is based on (a) a representation of the value of the outcome (i.e., the goal; e.g., how valuable is it to stay harm-free) and (b) a representation of the contingency between the response and the outcomes, also called the expectancy that the response will lead to the outcome (e.g., how likely running away or pressing a button will lead to the outcome of staying harm-free; Heyes & Dickinson, 1993; Moors et al., 2017). The action option with the highest expected utility (value x expectancy[3]) activates its corresponding action tendency.

Importantly, the goal-directed process does not occur in a vacuum but is best embedded in a cycle. When applied to fear conditioning, the CS would signal an impending discrepancy with a (first) goal (e.g., to stay harm-free). This would activate the (second) goal to reduce this discrepancy and result in the weighing and selecting of action options to do so. The selected action option would activate its action tendency, which – as said – can be understood as a (third) goal or inclination to act.

In sharp contrast to prevailing views, we therefore propose that conditioned fear is not a re-action or reflex (LeDoux & Daw, 2018; Pavlov, 1927) but a product of a goal-directed process. This proposal might be easier to accept in the light of recent evidence that goal-directed processes can be fairly automatic (i.e., can be fast, do not require awareness, etc.;

[3] There might be instances in which the utility of an action (tendency) is low at face value. For example, a chained prisoner might try to break his chains at the sight of his torturer, although such attempts could be considered futile. A goal-directed account can nonetheless account for such cases, because the interaction between expectancy and value is crucial. High value (e.g., to remain harm-free) can therefore compensate for low expectancy (e.g., that he can achieve remaining harm-free by actually breaking his chains) as long as the expectancy is higher than zero.

Moors, 2016). For instance, Bechara, Damasio, Tranel, and Damasio (1997) showed that participants were able to choose options with higher expected utilities without being able to verbally report on these expected utilities. This indicates that people can evaluate expected utilities of action options under conditions of automaticity. There is also preliminary evidence that early action tendencies to fight and flee are determined by expected utilities (Moors et al., 2019).

By proposing that fear is produced by a goal-directed process, we take things a step further than models which assume that goal-directed processes can at best intervene at a later stage to regulate reflex-like action tendencies or emotions (Moors et al., 2017). Nonetheless, our proposal does allow for regulation, be it in the form of competition between multiple goals. Consider, for example, that most participants will not demonstrate whole bodily actions, like running away, when confronted with a CS in a human fear conditioning experiment. Although a goal (e.g., to remain harm-free) is likely to activate those action tendencies and the accompanying physiological responses, competing goals will be at play as well and may play a regulatory role (e.g., staying in the experiment rather than getting up and leaving the laboratory results in the goal of receiving a participant fee and/or making the experimenter happy; Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017).

The goal-directed account of conditioned fear can be considered response-based because it allows for predictions about the type of responses that will occur. More specifically, it postulates that the response with the highest expected utility will be selected (e.g., running away vs. attacking an opponent). This is fundamentally different from the association formation models outlined above. Associative models deal with how associations between stimuli are formed but typically remain silent about which specific responses will be activated. In addition, the goal-directed approach implies that all variables affecting (1) the value of the outcomes and (2) the expectancy that the response will lead to the outcomes can be considered as determinants of responding as well. To illustrate the first point, a goal-directed account predicts that a CS (e.g., being in a plane or seeing a syringe with a lethal substance) could bring to mind the aversive US (e.g., dying) without resulting in fear, provided that the goal to stay harm-free is not valued at that moment in time (e.g., for suicide bombers or for people who requested euthanasia). With respect to the second point, the selection of the type of action would depend on the expected utility of this response in comparison with the expected utilities of other responses. For example, even somebody who values remaining harm-free might not prefer staying at home over climbing a rock without protective equipment if he believes that chances of an accident are equally high at home.

Summary

Before we turn to the evaluation of the goal-directed account of conditioned fear, we summarize its basic premises. In contrast to the dominant view in clinical psychology, which conceptualizes fear responding as the mere reflection of US activation in memory, we presented a view of conditioned fear as goal-directed behavior. According to this view, people can come to entertain a proposition that the CS is predictive of the US due to either exposure to situations in which there is a positive contingency between the CS and the US, instructions, or inference (Boddez et al., 2017). This proposition drives the expectancy of occurrence of the US upon confrontation with the CS. This CS is then hypothesized to function as a discriminatory stimulus that invites the selection of a motor act that will result in reaching the goal to stay harm-free. This leads to the activation of this action's corresponding action tendency. The action tendency can, in turn, go hand in hand with physiological responses (e.g., skin conductance) and possibly whole body motor actions (e.g., leaving the experiment). Awareness of this episode forms the content of feelings and verbal behavior (e.g., reporting fear). Competing goals (e.g., pleasing the experimenter or earning a participant fee) play a regulatory role and will, for instance, make that most participants sit through a human fear conditioning experiment without standing up and leaving. In line with recent evidence, we assumed that all this may occur under conditions of automaticity. We further argued that the goal-directed account allows making predictions about specific responses: the response with the highest expected utility will be selected.

Evaluation of the goal-directed perspective

As said, the goal-directed perspective deviates from prevailing views. We therefore consider it useful to discuss both possible objections against and new implications of the goal-directed view.

Possible objections against the goal-directed perspective. A first argument that critics of the goal-directed perspective may turn to is that the currently dominant models seem to be more in line with neuroscientific evidence. A main appeal of current theories about fear reduction indeed lies in their presumed fit with neuroscientific findings. For example, it is typically observed that extinction effects go hand in hand with activity of the prefrontal areas and the hippocampus in concert with the amygdala activity that is also observed during acquisition training. Although this is still a long way from an inhibitory association counteracting an excitatory one, this pattern is nonetheless taken as neuroscientific evidence for the inhibitory learning theory (e.g., Milad & Quirk, 2012; Milad et al., 2005, 2007; Phelps, Delgado, Nearing, & LeDoux, 2004). Unfortunately, however, the neuroscience-psychology interface is not that clear-cut: Knowledge about the neural level may allow to refute some mental process theories, but a large number of candidate psychological theories remains compatible with any finite set of neuroscientific data. Psychological theory ultimately provides the framework to interpret neuroscientific data and other theories may lead to a different interpretation of the interaction between the prefrontal cortex and the amygdala. For example, from the goal-directed perspective, the interplay between prefrontal areas and the amygdala following CS-only trials might be interpreted as reflecting the weighing of different behavior options (e.g., avoidance or not) when being confronted with an ambiguous CS. Interestingly, the fit between reconsolidation theory and the neuroscientific findings that are supposed to support it has recently been questioned as well (for an extensive review see Gisquet-Verrier & Riccio, 2018).

It has also been suggested that the excitatory and inhibitory associations assumed by inhibitory learning theory can literally be found in the brain in the form of synaptic connections (Kindt, 2018). However, there is no way of knowing whether the changes in the strength of synaptic connections equal associations or not, because of the same reason as mentioned above (i.e., multiple candidate psychological accounts are compatible with any finite set of neuroscientific data). Nonetheless, knowledge about the neural level does constrain theories at the mental level and, in contrast with the claim of the brain containing excitatory and inhibitory associations, there is evidence that the properties of the changes in synaptic transmission align poorly with the properties of associative learning as revealed by behavioral experimentation (for a review, see Gallistel & Matzel, 2013).

It is also worth noting that it is generally accepted that at least some of our behavior is mediated by propositions and goals rather than by associations (e.g., playing chess; Baeyens et al., 2009). Unless one would argue that those behaviors have no foundation in the brain, there is no reason to assume that propositional and goal-directed processes are less neurophysiologically plausible than associative processes.

A second argument against the goal-directed perspective could be that clinical fear is seen as “irrational” or “maladaptive”, whereas propositional beliefs and goals might be taken to imply rationality or even free will. If one a-priori considers every response which is driven by a goal-directed process to be rational, then our approach indeed implies rationality. However, the outcome of a goal-directed process can nonetheless seem irrational. For example, people can make errors when forming or retrieving propositions about CS-US contingencies (e.g., seeing relations where there are none; for a detailed discussion see Boddez et al., 2017), which could make the fear responding seem unnecessary to people who do not entertain these propositions. It is also possible that certain goals remain hidden from the observer. For example, fear (possibly including overt avoidance) caused by a job interview might come at the cost of financial security and might therefore seem irrational, but at the same time it could serve to prevent rejection and secure one's goal for high social status. Relatedly, goal-directed responses will not always serve all aspects of self-interest, because (1) different goals might be at conflict and (2) the assessment of values and expectancies is a subjective matter and people do not have complete knowledge about all behavior options and all their potential outcomes (Moors & Fischer, 2018). A goal-directed approach does not imply free will either (Moors, 2019). Simply put, it is one's phylogenetic and ontogenetic learning history that determines which goals are valued. Action selection, in turn, is determined by utility. From this historical perspective, one does not get to choose freely.

A third argument that could be raised against the goal-directed perspective is that it is unlikely to account for the behavior of non-human animals. The idea then seems to be that the presumed complexity of goal-directed processes can only be handled by the human mind. However, there is a large experimental literature on goal-directed behavior in animals (Heyes & Dickinson, 1993). These experiments examine whether responding of the animal is sensitive to manipulations of (a) the value of the outcome (i.e., the outcome devaluation test) and (b) the expectancy that the response results in the outcome (i.e., the contingency degradation test). Results show that animal behavior often conforms to these criteria of goal-directedness (Balleine, 2019).

Finally, one may be tempted to argue that the goal-directed perspective is too broad to be falsifiable. However, as discussed in the previous paragraph, experimental tests to identify the involvement of specific goals do exist. Still, an obvious challenge is the large number of potential goals, meta-goals, and their variability over time. Furthermore, the wide range of variables that can potentially affect the expected utility (value * expectancy) of the actions that may serve these goals is not a-priori defined. On the upside, all this does not take away from the usefulness of entertaining a goal-directed perspective. In the section below, we will indeed demonstrate that the goal-directed perspective may inspire new studies and therefore lead to new knowledge. As such, the goal-directed perspective is useful even for researchers who would like to defend the stance that it is too broad to be falsifiable. Allow us to make the humbling comparison with the Rescorla-Wagner model: Many of its predictions have been falsified (for an overview see Miller et al., 1995), but this history of falsification has not reduced the impact of the model, probably because it still provides an elegant account of some of the existing findings and continues to inspire new studies (i.e., it still provides heuristic and predictive value; Beckers & Vervliet, 2009). At least to some extent, the usefulness of a theoretical view thus lies in its capacity to generate falsifiable predictions that can be tested in empirical studies. By doing so, the theory helps us to extend our knowledge and thus our ability to predict and control. This is true regardless of whether falsification of its predictions allows for falsification of the theory itself.

Implications of the goal-directed perspective. As discussed**,** current procedures that attempt to reduce fear are typically described as interventions that “weaken the fear memory or augment the extinction memory” (Huff et al., 2009; p. 835; Wellman, Fitzpatrick, Machida, & Sanford, 2014). One could see such language as mere metaphors that are not necessarily informative about the hidden assumptions that the research community entertains about the relationship between memory and fear responding (i.e., as a monotonic relation). Alternatively, one could argue that there is some risk that a-priori assumptions about the mental mechanisms involved in extinction come to gain control over the research agenda (see

De Houwer, 2011; De Houwer et al., 2013; De Houwer & Hughes, 2017). It is indeed the case that a fair amount of the extinction-enhancing techniques are selected for study because of their presumed effect on activation of the US representation in memory. For example, presentation of an extinction reminder cue during renewal testing is studied because it is supposed to enhance retrieval of the inhibitory association (“extinction memory”; Pittig et al., 2016). Similarly, sleep after extinction learning is studied because of its presumed role in memory consolidation (Kleim et al., 2014) and extinction in multiple contexts is examined because it is thought to facilitate retrieval of the inhibitory association by increasing the chance of overlap between test context and extinction context (de Jong, Lommen, de Jong, & Nauta, 2019; Pittig et al., 2016). To be clear, the research strategy to study treatment techniques because one assumes that they have their effect through memory is in itself legitimate. If such techniques turn out to be successful, then this is a valuable outcome even if the techniques would turn out to rely on other mental mechanisms than memory (e.g., sleep might affect inferential processes or the relative importance of goals such as staying harm-free). At the same time, the consideration that other determinants than associative memory activation could play a role in producing fear could lead to exploring different treatment techniques.

We discussed two other determinants of fear responding than mere memory activation: the operation of a propositional mechanism and of a goal-directed mechanism. Invoking propositions invites new hypotheses about the extinction deficits observed in anxiety patients (Duits et al., 2015). For example, these could be caused by a specific inference rather than by an “inhibitory learning deficit”. More precisely, patients might know very well that the US did not follow the CS on previous trials (i.e., there would be no learning deficit), but still infer that this might not hold on the current trial. This could be so because they entertain the inference rule that what happened on the previous (nonreinforced) trials will not necessarily happen on the current trial. If this would indeed be so, one should find a dissociation between (verbal) measures of remembrance of previous CS-(no)US pairings and of prediction of US-occurrence in anxiety patients. In addition, one could target (confirm or negate; Boddez et al., 2017) this inference rule in an experimental study and assess whether that affects (slows down or speeds up) extinction performance.

Invoking the goal-directed perspective, in turn, uniquely predicts that devaluation of the outcome that drives fear responses (e.g., staying harm-free) would reduce these responses. In an experiment, one could offer a vignette that devalues the importance of living a harm-free life and see how it affects conditioned fear. Alternatively (or additionally), one could try to alter the expectancy that (preparing for) the motor act to leave the experiment serves the goal to stay harm-free. To this end, one could, for example, offer a vignette that describes that leaving the experiment is as risky as staying in it (e.g., one could trip on the way out or get hit by a car). Although it might be challenging to present this information in such way that it will affect action selection under conditions of automaticity, techniques to enhance the impact of such new information are available (e.g., hypnosis; Van Dessel & De Houwer, 2019).

If these experiments would confirm our hypotheses, then one could translate these techniques to clinical practice. Treatment could focus, for instance, on helping overly cautious patients by questioning the high value of staying harm-free relative to that of competing goals and/or by helping them recalibrate the expectancy that avoidance (e.g., of leaving the house) will lead to a harm-free life (e.g., one could also get harmed at home).

We can now also re-analyze the above-discussed case-study of the patient who successfully defeated flying phobia. As said, inhibitory learning theory would state that, during exposure therapy, an inhibitory association between being on an airplane and dying in a plane crash would have been learned, which would cause the patient to not or barely think about dying in a plane crash when boarding a plane (i.e., non-permanent amnesia). However, following therapy, this patient still thought about dying in a plane crash (US) but nonetheless persisted because he managed to persist during therapy and believed his family deserves a holiday (cf. Williams, Kinney, & Falbo, 1989). The goal-directed account would hold that the expected utility of boarding the flight (serving, for example, the goal to have a family vacation abroad) has come to outweigh that of avoiding the flight (serving the goal to stay harm-free) during therapy. An increase in expected utility of flying compared to not flying may be due to (a) an increase in the value of the outcome of flying (i.e., the family vacation abroad has become more valuable), (b) an increase in the expectancy that one can achieve this outcome by flying (i.e., an increase in self-efficacy), (c) a decrease in the value of the outcome of staying harm-free, (d) and/or a decrease in the expectancy that not flying is necessary in order to stay harm-free.

The goal-directed perspective might also inspire new studies about the effects seen in reconsolidation interference experiments (also see Cogan, Shapses, Robinson, & Tronson, 2019). Propranolol is a beta blocker with anxiolytic effects. Administration of propranolol before (or shortly after) the presentation of the fear-conditioned stimulus could therefore make the participant feel calm while there is (still) some degree of mental activation of the CS and the related US (Wagner, 1981). This points to an analogy with systematic desensitization therapy, which entails the confrontation with an either presented or imagined threatening stimulus while the patient is asked to relax (Wolpe, 1958). The rationale behind the relaxation component is that it consists of a response pattern which is antagonistic, or at least incompatible, with fear responding. Although it has been questioned whether this component has added value (beyond only exposure), there are studies that suggest that relaxation plays a facilitative role (Levin & Gross, 1985). Rachman (1968) suggested that, rather than muscular relaxation, a sense of mental calm is what is helping. Not only the beta blocker propranolol may create such a sense of calm, but also other techniques such as playing Tetris (James et al.,

2015) may evoke a response pattern that is incompatible with fear responding. This absence of fear responding could lead participants to infer that they have become relatively immune to the unpleasant effects of being confronted with the CS. Even when the reconsolidation interference technique is applied only (shortly) after the presentation of the CS, they might infer that they have become immune to its unpleasant after-effects (e.g., ruminating or being upset). Given this newly developed immunity, subjects might believe that their goal to stay harm-free is guaranteed even in the absence of avoidance. This would imply that there is no longer a need to activate avoidance tendencies (and the corresponding physiological activity) to reach this goal. This hypothesis also explains why administration of propranolol after exposure to a tarantula in spider phobics reduces later avoidance behavior in response to it (Soeter & Kindt, 2015): Patients might consider themselves to be immune to the effects of being confronted with this stimulus after treatment, which takes away the need to avoid in order to stay harm-free. Some animal findings already suggest that this is a valuable account (Cogan et al., 2019). Interestingly, the goal-directed account also provides a way to understand the reported dissociations between intact US-expectancy and reduced startle responding after propranolol treatment (Soeter & Kindt, 2010). One can hypothesize that the responses to US-expectancy questions serve another goal than physiological responses. The US expectancy rating responses may stem from the epistemic goal to demonstrate one's knowledge of the environment (Niv & Chan, 2011), while the startle response is a measure of the action readiness to avoid in order to remain harm-free.

Conclusion

We argued that an important part of the fear conditioning community implicitly adopted the association concept and the monotonicity assumption, which is recognized in the dominant idea that tackling a single association stored in memory is the (only) way to go if one wants to remediate fear responding. We illustrated that this account of fear responding as the mere reflection of US activation overlooks other important determinants (i.e., relational information and the goal to stay harm-free) of the anxiety symptoms that the research community eventually aims to reduce.

An obvious solution would be to evolve to a light version of this assumption in which memory retrieval of the US is just one of many mental processes that may determine the fear response. However, this version is not the dominant view, as illustrated in the literature by (a) the interchangeable use of the terms fear memory and fear responding and by (b) reverse inferences in which the presence / absence of memory is inferred from the level of fear responding.

We introduced the goal-directed account, which provides an explanation for both the expectancy- and utility-based nature of conditioned fear. We hope that it may offer new directions for utilizing fear conditioning procedures in treatment research.

References

Baeyens, F., Díaz, E., & Ruiz, G. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cognition & Emotion, 19*, 245–268.

Baeyens, F., Vansteenwegen, D., & Hermans, D. (2009). Associative learning requires associations, not propositions. *Behavioral and Brain Sciences, 32*(2), 198-199.

Balleine, B. W. (2019). The meaning of behavior: discriminating reflex and volition in the brain. *Neuron, 104*, 47-62.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293-1295.

Beckers, T., & Kindt, M. (2017). Memory reconsolidation interference as an emerging treatment for emotional Disorders: Strengths, limitations, challenges, and opportunities. *Annual Review of Clinical Psychology, 13,* 99–121.

Beckers, T., Krypotos, A. M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning? *Biological psychology*, *92*(1), 90-96.

Beckers, T., & Vervliet, B. (2009). The truth and value of theories of associative learning. *Behavioral and Brain Sciences, 32*, 200-201.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Current Directions in Psychological Science, 26*, 422-428.

Boddez, Y., Baeyens, F., Hermans, D., & Beckers, T. (2011). The hide-and-seek of retrospective revaluation: Recovery from blocking is context dependent in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 230-240.

Boddez, Y., Bennett, M. P., van Esch, S., & Beckers, T. (2017). Bending rules: The shape of the perceptual generalisation gradient is sensitive to inference rules. *Cognition and Emotion, 31*, 1444-1452.

Boddez, Y., De Houwer, J., & Beckers T. (2017). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasonin*g. Oxford UK: Oxford University Press

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

Bouton, M. E., & Moody, E. W. (2004). Memory processes in classical conditioning. *Neuroscience & Biobehavioral Reviews, 28*, 663-674.

Carter, B.L., & Tiffany, S.T., 1999. Meta-analysis of cue-reactivity in addiction research. *Addiction, 94*, 327–340.

Chalkia, A., Vanaken, L., Fonteyne, R., & Beckers, T. (2019). Interfering with emotional processing resources upon associative threat memory reactivation does not affect memory retention. *Scientific reports*, *9*, 4175.

Cogan, E. S., Shapses, M. A., Robinson, T. E., & Tronson, N. C. (2019). Disrupting reconsolidation: memory erasure or blunting of emotional/motivational value. *Neuropsychopharmacology*, *44,*399-417.

Craske, M. G., Hermans, D., & Vansteenwegen, D. (2006). *Fear and learning: From basic processes to clinical implications*. Washington, DC: American Psychological Association.

Craske, M. G., & Mystkowski, J. (2006). Exposure therapy and extinction: Clinical studies. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Eds.), *Fear and learning: Basic science to clinical application* (pp. 213-233). Washington, DC: American Psychological Association.

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

Dacey, M. (2018). Simplicity and the Meaning of Mental Association. *Erkenntnis*, 1-22.

De Beukelaar, T. T., Woolley, D. G., Alaerts, K., Swinnen, S. P., & Wenderoth, N. (2016). Reconsolidation of motor memories is a time-dependent process. *Frontiers in human neuroscience*, *10*, 1-10.

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, *6*(2), 202-209.

De Houwer, J. (2020; in press). Conditioning is More Than Association Formation: On the Different Ways in Which Conditioning Research is Valuable for Clinical Psychology. *Collabra*.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review*, *20*(4), 631-642.

De Houwer, J., & Hughes, S. (2017). Environmental regularities as a concept for carving up the realm of learning research: Implications for Relational Frame Theory. *Journal of Contextual Behavioral Science, 6*(3), 343-346.

de Jong, R., Lommen, M. J. J., de Jong, P. J., & Nauta, M. H. (2019). Using multiple contexts and retrieval cues in exposure-based therapy to prevent relapse in anxiety disorders. *Cognitive and Behavioral Practice, 26*(1), 154-165.

de Voogd, L. D., Hermans, E. J., & Phelps, E. A. (2018). Regulating defensive survival circuits through cognitive demand via large-scale network reorganization. *Current Opinion in Behavioral Sciences, 24*, 124-129.

Duffy, E. (1941). An explanation of "emotional" phenomena without the use of the concept "emotion". *The Journal of General Psychology*, *25*(2), 283-293.

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., ... & Baas, J. M. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and anxiety, 32*(4), 239-253.

Dunsmoor, J. E., & Kroes, M. C. (2019). Episodic memory and Pavlovian conditioning: ships passing in the night. *Current Opinion in Behavioral Sciences*, *26*, 32-39.

Elsey, J. W. B., Van Ast, V. A., & Kindt, M. (2018). Human memory reconsolidation: A guiding framework and critical review of the evidence. *Psychological Bulletin, 144*(8), 797-848.

Fanselow, M. S., & Pennington, Z. T. (2018). A return to the psychiatric dark ages with a two-system framework for fear. *Behaviour research and therapy, 100*, 24-29.

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review, 26*, 857-875.

Gallistel, C. R., & Matzel, L. D. (2013). The neuroscience of learning: Beyond the Hebbian synapse. *Annual Review of Psychology*, *64*, 169-200.

Garcia-Marques, L., & Ferreira, M. B. (2011). Friends and foes of theory construction in psychological science: Vague dichotomies, unified theories of cognition, and the new experimentalism. *Perspectives on Psychological Science*, *6*(2), 192-201.

Gawronski, B., & Bodenhausen, G. V. (2014). Implicit and explicit evaluation: A brief review of the Associative-Propositional Evaluation Model. *Social and Personality Psychology Compass*, *8*, 448-462.

Gisquet-Verrier, P., and Riccio, D. C. (2018). Memory integration: an alternative to the consolidation/reconsolidation hypothesis. *Progress in Neurobiology, 171*, 15–31.

Goodman, N. (1955). Fact, Fiction, & Forecast. Cambridge: Harvard University Press.

Haselgrove (2016). Overcoming associative learning. *Journal of Comparative Psychology, 130*, 226–240.

Heyes, C., & Dickinson, A. (1993). The intentionality of animal action. In M. Davies, & G. W. Humphreys (Eds.). *Consciousness: Psychological and philosophical essays* (pp. 105–120). Oxford, UK: Blackwell.

Hofmann, S. G., & Smits, J. A. J. (2008). Cognitive-behavioral therapy for adult anxiety disorders: A meta-analysis of randomized placebo-controlled trials. *The Journal of Clinical Psychology, 69*, 621-632.

Holland, P. C. (1993). Cognitive aspects of classical conditioning. *Current opinion in neurobiology*, *3*, 230-236.

Huff, N. C., Hernandez, J. A., Blanding, N. Q., & LaBar, K. S. (2009). Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. *Behavioral neuroscience*, *123*(4), 834-843.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264. https://doi.org/10.1037/0033- 295X.110.2.220

James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., & Holmes, E. A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological science*, *26*(8), 1201-1215.

Jozefowiez, J. (2018). Associative versus predictive processes in Pavlovian conditioning. *Behavioural processes*, *154*, 21-26.

Khalaf, O., Resch, S., Dixsaut, L., Gorden, V., Glauser, L., & Gräff, J. (2018). Reactivation of recall-induced neurons contributes to remote fear memory attenuation. *Science, 360*, 1239-1242.

Kindt, M. (2018). The surprising subtleties of changing fear memory: a challenge for translational science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1742), 20170033.

Kindt, M. *,* Soeter, M., & Vervliet, B. (2009). Beyond extinction: Erasing human fear responses and preventing the return of fear. *Nature Neuroscience, 12,* 256-258.

Kleim, B., Wilhelm, F. H., Temp, L., Margraf, J., Wiederhold, B. K., & Rasch, B. (2014). Sleep enhances exposure therapy. *Psychological medicine*, *44*(7), 1511-1519.

LeDoux, J., & Daw, N. D. (2018). Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience, 19*, 269-282.

Lee, J. L. (2009). Reconsolidation: maintaining memory relevance. *Trends in neurosciences*, *32*(8), 413-420.

Levin, R. B., & Gross, A. M. (1985). The role of relaxation in systematic desensitization. *Behaviour research and therapy*, *23*(2), 187-196.

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560–572.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning and Memory*, *12*(4), 361–366.

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General*, *118*(1), 100–104.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(1), 19–31.

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J., & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review, 42*, 72-82.

Luyten L., Beckers T. (2017). A preregistered, direct replication attempt of the retrieval-extinction effect in cued fear conditioning in rats. *Neurobiology of Learning and Memory*, *144*, 208-215.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298.

Matsumoto, D., & Ekman, P. (2009). Basic emotions. In D. Sander & K. R. Scherer (Eds.), *Oxford companion to affective sciences* (pp. 69–72). Oxford, UK: Oxford University Press.

McLaren, I.P.L., Forrest, C.L.D., McLaren, R.P., Jones, F.W., Aitken, M.R.F., & Mackintosh, N.J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of Learning and Memory, 108*, 185-195.

Milad, M. R., Quinn, B. T., Pitman, R. K., Orr, S. P., Fischl, B., & Rauch, S. L. (2005). Thickness of ventromedial prefrontal cortex in humans is correlated with extinction memory. *PNAS: Proceedings of the National Academy of Sciences of the United of America, 102*, 10706e10711.

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: Ten years of progress. *Annual Review of Psychology, 63*, 129e151.

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry, 62*, 446e454.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological bulletin*, *117*(3), 363-383.

Mineka, S., & Zinbarg, R. (2006). A contemporary learning theory perspective on the etiology of anxiety disorders: it's not what you thought it was. *American psychologist, 61*(1), 10-26.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198.

Monfils, M. H., & Holmes, E. A. (2018). Memory boundaries: opening a window inspired by reconsolidation to treat anxiety, trauma-related, and addiction disorders. *The Lancet Psychiatry*.

Monfils, M. H., Cowansage, K. K., Klann, E., & LeDoux, J. E. (2009). Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science*, *324*(5929), 951-955.

Moors, A. (2009). Theories of emotion causation: A review. *Cognition and emotion, 23*, 625-662.

Moors, A. (2014). Examining the mapping problem in dual process models. In J. Sherman, B. Gawronski & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 20–34). New York: Guilford Press.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263-287.

Moors, A. (2017). The integrated theory of emotional behavior follows a radically goal-directed approach. *Psychological Inquiry, 28*, 68-75.

Moors, A., Boddez Y., De Houwer J. (2017). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, *9* (4), 310-318.

Moors, A., Ellsworth, P., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review, 5*, 119 - 124.

Moors, A. (2019, January 7). Towards a goal-directed account of weak-willed behavior [Blog post]. Retrieved from http://philosophyofbrains.com/2019/01/07/empirically-informed-approaches-to-weakness-of-will-a-brains-blog-roundtable.aspx

Moors, A., Fini, C., Everaert, T., Bardi, L., Bossuyt, E., Kuppens, P., & Brass, M. (2019). The role of stimulus-driven versus goal-directed processes in fight and flight tendencies measured with motor evoked potentials induced by transcranial magnetic stimulation. *PLoS ONE, 14*, e0217266, 1-22.

Moors, A., & Fischer M. (2019). Demystifying the role of emotion in behaviour: toward a goal-directed account. *Cognition & Emotion*, *33* (1), 94-100.

Niv, Y., & Chan, S. (2011). On the value of information and other rewards. *Nature neuroscience*, *14*(9), 1095-1097.

Orederu, T., & Schiller, D. (2018). Fast and slow extinction pathways in defensive survival circuits. *Current Opinion in Behavioral Sciences, 24*, 96-103.

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

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532–552. doi: 10.1037/0033-295X.87.6.532

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

Pittig, A., Treanor, M., LeBeau, R. & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience and Biobehavioral Reviews, 88*, 117-140.

Pittig, A., van den Berg, L., & Vervliet, B. (2016). The key role of extinction learning in anxiety disorders: behavioral strategies to enhance exposure-based treatments. *Current opinion in psychiatry*, *29*(1), 39-47.

Rachman, S. (1968). The role of muscular relaxation in desensitization therapy. *Behaviour Research and Therapy, 6*(2), 159-166.

Raes, A. K., De Houwer, J., Verschuere, B., & De Raedt, R. (2011). Return of fear after retrospective inferences about the absence of an unconditioned stimulus during extinction. *Behaviour Research and Therapy, 49*, 212-218.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist, 43*, 151-160.

Rescorla, R. A. (2001). Are associative changes in acquisition and extinction negatively accelerated? *Journal of Experimental Psychology: Animal Behavior Processes, 27*, 307-315.

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 (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Scherer, K.R., Moors A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology*, *70*, 719-745.

Scheveneels, S., Boddez, Y., Vervliet, B., & Hermans, D. (2016). The validity of laboratory-based treatment research: Bridging the gap between fear extinction and exposure treatment. *Behaviour Research and Therapy*, *86*, 87-94.

Scheveneels, S., Boddez, Y., De Ceulaer, T., & Hermans, D. (2019). Ruining the surprise: the effect of safety information before extinction on return of fear. *Journal of behavior therapy and experimental psychiatry*, *63*, 73-78.

Scheveneels S., Boddez Y., Hermans D. (2019). Learning mechanisms in fear and anxiety: It is still not what you think it is. In: Olatunji B. (Eds.), *The Cambridge Handbook of Anxiety and Related Disorders* (13-40). Cambridge: Cambridge University Press.

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*, 49-53.

Schmidt, J. R., De Houwer, J., & Rothermund, K. (2016). The Parallel Episodic Processing (PEP) model 2.0: A single computational model of stimulus-response binding, contingency learning, power curves, and mixing costs. *Cognitive Psychology*, *91*, 82-108.

Schroyens N., Beckers T., Kindt M. (2017). In search for boundary conditions of reconsolidation: A failure of fear memory interference. *Frontiers in Behavioral Neuroscience*, *11*, Art.No. 65.

Shaw, J., & Porter, S. (2015). Constructing rich false memories of committing crime. *Psychological Science*, *26*(3), 291–301.

Siegel, P., & Weinberger, J. (2012). Less is more: The effects of very brief versus clearly visible exposure. *Emotion, 12*, 394-402.

Skinner, B. F. (1977). Why I am not a cognitive psychologist. *Behaviorism, 5,* 1-10

Soeter, M., & Kindt, M. (2010). Dissociating response systems: erasing fear from memory. *Neurobiology of learning and memory*, *94*(1), 30-41.

Soeter, M., & Kindt, M. (2015). An abrupt transformation of phobic behavior after a post-retrieval amnesic agent. *Biological psychiatry*, *78*(12), 880-886.

Squire, L.R. (1987). *Memory and Brain*. Oxford University Press, Oxford.

Stroebe, W. (2018). The task of social psychology is to explain behavior not just to observe it. *Social Psychological Bulletin*, *13*, e26131.

Van Dessel, P., & De Houwer, J. (2019). Hypnotic suggestions can induce rapid change in implicit attitudes. *Psychological science, 30*, 1362-1370.

Vervliet, B., Baeyens, F., Van den Bergh, O., & Hermans, D. (2013). Extinction, generalization and return of fear: A critical review of renewal research in humans. *Biological Psychology, 92*(1), 51–58.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: state of the art. *Annual review of clinical psychology*, 9, 215-248.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). Hillsdale, NJ: Erlbaum.

Weisman, J. S., & Rodebaugh, T. L., (2018). Exposure therapy augmentation: A review and extension of techniques informed by an inhibitory learning approach. *Clinical Psychology Review, 59*, 41-51.

Wellman, L. L., Fitzpatrick, M. E., Machida, M., & Sanford, L. D. (2014). The basolateral amygdala determines the effects of fear memory on sleep in an animal model of PTSD. *Experimental brain research*, *232*(5), 1555-1565.

Williams, S. L., Kinney, P. J., & Falbo, J. (1989). Generalization of therapeutic changes in agoraphobia: The role of perceived self-efficacy. *Journal of Consulting and Clinical Psychology*, *57*, 436-442.

Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., ... & Fratiglioni, L. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European neuropsychopharmacology, 21*, 655-679.

Wolpe, J. (1958) Psychotherapy by reciprocal inhibition. Stanford University Press, Stanford.

Zenses A-K., Beckers T., Boddez Y. (2018). A novel fear-conditioning procedure to model intrusive thinking. Presented at the European Meeting on Human Fear Conditioning, Cardiff, Wales, 16 Apr 2018-18 Apr 2018.