Prediction errors in changing fear

Stemerding, L.E.

Publication date
2023

Link to publication

Citation for published version (APA):
Stemerding, L. E. (2023). *Prediction errors in changing fear*. [Thesis, fully internal, Universiteit van Amsterdam].

CHAPTER 1

General introduction

Emotional memories and psychopathology

When asked to consider the impact of emotional memories, most people will be inclined to think about how easily and vividly they can remember the loss of a loved one, an embarrassing event, or on the positive side, their graduation, or a unique holiday. Indeed, every person knows from experience that emotional events are deeply ingrained in our memory and readily retrieved. Yet the power of these memories is not limited to conscious recollection. In day-to-day life, emotional memories shape our thoughts, actions, and behaviour in various ways – often without realising – and therefore play a critical role in our physical and mental well-being. The impact of emotional memories becomes especially clear when considering mental disorders such as fear and anxiety disorders. Is it generally believed that the behaviour that characterises these disorders, for example an intense fear of needles, is rooted in aversive emotional memories that have formed at some point in one's life (Kindt, 2014; Mineka & Zinbarg, 2006). While the responses that patients exhibit may be initially functional, these fear responses tend to generalise to situations where there is no actual danger, resulting in behaviour that is maladaptive. Moreover, this behaviour is extremely resistant to change, as is often observed when attempting to treat fear and anxiety disorders. Current treatments can be very effective, but a relatively large proportion of patients does either not respond to treatment at all or experiences relapse after an initial reduction of symptoms (Levy et al., 2021; Lorimer et al., 2021). With the rising prevalence of mental disorders (ten Have et al., 2022), improving the effectivity of treatments is more important than ever, and key to this goal is a better understanding of how the impact of emotional memories can be changed.

Changing the impact of aversive memories

Assuming that fear and anxiety disorders result from the persistence of maladaptive emotional memories that drive involuntary and unwanted behaviour, there are two potential pathways to treatment. Put simply, either the original fear memory can be changed to reduce its effect over behaviour, or a new, more adaptive memory can be created that inhibits the impact of the fear memory. In humans, emotional memory is a latent construct, and we can only infer its strength indirectly from behavioural observations. Based on these behavioural read-outs, it is impossible to decisively determine whether original memories have changed or new memories have been formed, yet there are some critical differences. Theoretically speaking, updating original memories should be most effective in instigating durable changes to behaviour, as it implies that (in the absence of a new aversive experience) there is little risk that the fear memory regains strength. In contrast, when a new memory is created, symptoms may at first reduce but the fear memory can eventually come to control behaviour again.

Typically, it is assumed that the fact that symptoms can return after initially successful treatments demonstrates that most of the current CBT-based interventions like exposure therapy depend on the formation of new memories (Bouton, 2002; Brewin, 2006; Craske et al., 2008). During exposure therapy, patients undergo repeated confrontations with feared objects or situations, usually under the guidance of a therapist, with the goal of reducing the unwanted behavioural responses to these objects or situations. Throughout the past decades, various theories have been proposed to explain the effects of exposure therapy (e.g., Foa & Kozak, 1986; Wolpe, 1968), but currently the most dominant hypothesis is that exposure effects are driven by learning from the absence of reinforcement, known as extinction learning[1] (Craske et al., 2014; Craske, Treanor, et al., 2022). For example, a patient with needle phobia may fear disastrous consequences of receiving an injection. During exposure therapy, this patient eventually experiences that if they do receive an injection, nothing disastrous happens, resulting in learning that needles are basically safe. If successful, extinction learning is characterised by reductions in fear responding when presented with the initially feared cue (the needle). While the exact neural mechanisms of these changes are only beginning to be understood (Bouton et al., 2021; Craske, Sandman, et al., 2022; Delamater & Westbrook, 2014), extinction learning is in principle believed to result in the formation of a new memory. This extinction memory then competes with, or inhibits, the fear memory, subsequently resulting in behavioural change. The return of fear behaviour is then the consequence of a shift in the balance between the strengths of the extinction memory and the fear memory. Changes in this balance can result from various factors including a change of context, the passage of time, and re-exposure to aversive events (Bouton, 1993, 2002; Brewin, 2006).

In contrast to interventions that aim to change behaviour by facilitating the formation of new, more adaptive memories, there is considerable treatment potential in understanding how we can change the original memory itself. No matter how effective the treatment is, if the original memory is not changed there is always a chance that symptoms return. Although it was long believed that memories were more or less permanent entities in the brain, early studies in animals found that a brief reactivation can render memories into a destabilised form, after which they are susceptible to manipulations (Misanin et al., 1968). The subsequent blocking of protein synthesis prevented the restabilization of the memories, resulting in a strong reduction of associated behaviour (Nader et al., 2000; Przybyslawski & Sara, 1997). While blocking protein synthesis is not feasible in humans, the beta-blocker propranolol has

[1] The term extinction learning refers to both the process and the procedure. The process extinction learning describes learning from the absence of reinforcement (Pavlov, 1927). The procedure, on the other hand, describes the manipulation that is often used in the lab to model the process extinction learning. During this manipulation, the conditioned stimulus is presented multiple times without reinforcement.

also been shown to intervene with reconsolidation, providing potential clinical utility (Debiec & LeDoux, 2004). Indeed, administration of propranolol before or after memory reactivation in humans led to abrupt and long-lasting reductions in fear behaviour in the lab (Kindt et al., 2009; Soeter & Kindt, 2010, 2011, 2012a). Yet while some reconsolidation-based treatments of PTSD and phobias in humans resulted in a strong reduction of symptoms (Brunet et al., 2018; Soeter & Kindt, 2015), other clinical studies failed to find similar results (Elsey et al., 2020; Wood et al., 2015). Moreover, recent attempts to replicate the initially highly effective experimental effects of blocking memory reconsolidation have also been unsuccessful (Bos et al., 2014; Chalkia et al., 2020; Schroyens et al., 2017). These mixed findings are puzzling, but may to a certain extent result from a relatively poor understanding of what it takes to effectively destabilise memories (Rotondo et al., 2022). While it is now widely accepted that one critical condition is the occurrence of an unexpected event or match-mismatch (Agustina López et al., 2016; Forcato et al., 2009; Pedreira et al., 2004; Pedreira & Maldonado, 2003; Sevenster et al., 2012b, 2013, 2014), the exact conditions that are necessary for memory destabilization remain to be identified.

While extinction learning and memory reconsolidation are different processes to alter the impact of emotional memories, a more recent computational model of learning has suggested how the transition between original memory change and new learning takes place. The latent cause model proposes that during learning, individuals infer unobservable latent causes that determine the relationship between all stimuli in the environment (Gershman, 2015; Gershman & Niv, 2012). For example, if a cue is known to predict an aversive outcome, an individual may infer that a latent cause is currently active in which a cue predicts an outcome. The next time the individual observes this cue, it estimates which latent cause is most likely to be active, and adapts its behaviour accordingly (i.e., prepare for the occurrence of the aversive outcome or not). The latent cause that is most likely to be active is determined by comparing the current characteristics of the environment to known latent causes. Critically, the cue → outcome relationship is a part of the environment, and thus if the outcome does not occur, the animal may infer the existence of a new latent cause. From a neurobiological perspective, the creation of a new latent cause could be equated to the creation of a new memory (Gershman et al., 2017). It has subsequently been suggested that original memories can be updated by "unlearning" the cue → outcome relationship within the original cause, so that the safety information is incorporated into the original memory (Gershman et al., 2013; Gershman & Hartley, 2015; Shiban et al., 2015). Extinction learning may thus not per definition have to result in the formation of a new "safe" memory. Instead, under the right circumstances, safety information could be incorporated into the original memory through reconsolidation-like mechanisms.

The conditions that determine change

As discussed above, current developments in interventions for fear and anxiety disorders focused on how to directly update memories (reconsolidation-based interventions) or how to strengthen extinction learning (exposure-based interventions). Both interventions depend to some extent on associative learning processes and therefore a better understanding of the conditions that govern this type of learning is essential to comprehend the respective mechanisms of change. In the past decade, the construct prediction error as critical condition for learning is gaining ground in clinical science. Defined as the discrepancy between an expected outcome and an actually experienced outcome, prediction errors are a core feature of most associative learning models (Pearce & Hall, 1980; Rescorla & Wagner, 1972; Sutton & Barto, 1987). The occurrence and magnitude of a prediction error determines when learning takes place and how much can be learned. For example, if an outcome is fully predicted by the preceding cue, no more learning about this cue occurs, and behaviour is thus not affected. In contrast, if an outcome is suddenly omitted (as is the case during extinction), a prediction error occurs and the predictive value of the cue is updated, typically resulting in a decrease in fear behaviour. While prediction errors have been central in studies of reinforcement learning and decision making for decades, the field of clinical psychology has recently embraced this concept as a potentially important condition for treatment success. The occurrence of a prediction error is deemed to be a critical component of both reconsolidation and exposure-based treatments. In particular, the magnitude or frequency of prediction error occurrence has been suggested to determine the effectiveness of interventions. The manner in which prediction errors are believed to affect interventions, however, differs depending on the learning process that is believed to take place.

The view that prediction errors are essential to the effectiveness of exposure treatments is formalised in the inhibitory retrieval model. This model proposes that extinction learning can be strengthened by enhancing prediction errors, which are often operationalised as violations of outcome expectations (Craske et al., 2014; Craske, Treanor, et al., 2022). Stronger extinction memories should be better in limiting the impact of aversive memories and therefore result in longer-lasting symptom reductions. Based on this proposition, a wide range of potential expectation-based strategies to enhance extinction learning have been investigated both in the lab and clinical practice (Craske et al., 2014, 2018; Dunsmoor et al., 2015; Lipp et al., 2020). While some of these strategies have proven to effectively strengthen long-term intervention effects (Coelho et al., 2015; Lucas et al., 2018; Thompson et al., 2018), many of them do not necessarily outperform control groups (Buchholz et al., 2022; Kircanski et al., 2012; Lancaster et al., 2020; Meulders et al., 2016; Scheveneels, Boddez, van Daele, et al., 2019; Scheveneels, Boddez, Vervliet, et al., 2019). Furthermore, while these strategies are based on the notion that expectancy violations should be maximised, the underlying mechanisms of this hypothesis are not always well defined. A more detailed understanding of the exact role of prediction errors on extinction learning is thus needed.

In memory reconsolidation interventions, prediction errors are suggested to play a critical role in the destabilization of the memory. The necessity of an unexpected outcome for the occurrence of memory destabilization has been formalised in terms of prediction error (Exton-McGuinness et al., 2015; Fernández et al., 2016; Krawczyk et al., 2017; Lee, 2009). More specifically, it was shown that the destabilization of memory is conditional on the number of prediction errors that occur during a reactivation session (Merlo et al., 2014; Sevenster et al., 2014). While too little prediction errors may not prompt any change, too many prediction errors can already result in new learning, leaving the original memory intact and thus immune to any manipulations. The exact number of prediction errors that is necessary for successful destabilization depends, among other factors, on what was learned during conditioning (see e.g., Sevenster et al., 2013, 2014). For example, following a relatively weak conditioning phase, a single prediction error may induce destabilization whereas one or two more prediction errors may already instigate the process of new learning, rendering destabilization unsuccessful. This delicate balance between no change, memory destabilization, and new learning is a highly challenging aspect of memory reconsolidation, and potentially the reason why some fail to find any effects. Importantly, this balance also plays an important role in latent cause theory (Gershman & Niv, 2012), which proposes that prediction errors are among some of the most critical factors that determine whether the original memory is updated, or a new memory formed. Specifically, strong and frequent prediction errors can signal that the environment has fundamentally changed and that thus a new latent cause (memory) should be formed, whereas weaker or less frequent prediction errors can trigger updating of the current latent cause due to a smaller discrepancy. However, a better understanding of the exact role of prediction errors in memory updating versus new learning is essential to ensure the effectivity of reconsolidation interventions as well as extinction effects.

In sum, prediction errors are deemed to play a critical role in a range of learning-based interventions that target emotional memories, including memory reconsolidation and extinction learning. While the necessity of a prediction error for learning is unquestioned, detailed knowledge on the relationship between prediction error magnitude and learning (and ultimately treatment success) is lacking. To potentially utilise and manipulate prediction errors to improve psychological treatments, a more thorough understanding of their exact effects on fundamental learning processes is necessary.

The challenges of studying prediction errors

Ensuring the occurrence of a prediction error is relatively straightforward. As long as an outcome is not 100% predictable, a prediction error will occur. However, quantifying their magnitude is notoriously more difficult. Currently, a common behavioural quantification of prediction errors in humans is the consequence of their occurrence: a change in behaviour. Yet this poses two major problems. First, the behavioural change can only be observed upon a new presentation of the cue, as it is the cue that triggers the behaviour. Showing the cue again, however, also constitutes a new learning experience, and therefore a new prediction error (in the absence of reinforcement). This can be a major challenge when translating the process of reconsolidation to interventions, as the destabilization of memories depends on subtle conditions. For example, in the case where the occurrence of a single prediction error is both necessary and sufficient for memory destabilization, the occurrence of a new prediction error could already result in the transition to new learning (e.g., Sevenster et al., 2014). Another concern of using conditioned behaviour to infer prediction errors is that behaviour may not always update after the experience of a prediction error, and can therefore not serve as an independent operationalisation. The absence of a behavioural change cannot be used as evidence that “no prediction error must have occurred” as there are potential other reasons why no learning took place. Further, a behavioural change does not necessarily result from a prediction error, and may also be driven by other (non-associative) learning processes. Hence, developing an independent measure of prediction errors is critical to advance insights into their effect on behaviour.

Animal work has identified a few neural substrates of aversive prediction error signaling (Delgado et al., 2008; Iordanova et al., 2021; McHugh et al., 2014), which have to some extent been translated to humans (Roy et al., 2014; Thiele et al., 2021). Yet in humans these signals are not sufficiently accurate to infer prediction error occurrence at a single timepoint. In contrast, a frequently employed read-out in clinical science concerns the violation of conscious expectations of an outcome. For example, if a patient indicates that they are 80% sure that they will faint when seeing a needle, and they do not faint, the prediction error is 80 (on an arbitrary scale from 0 to 100). However, verbal expectations may not always reflect the affective value of predictions and may therefore not capture the full extent of prediction errors. It has alternatively been investigated whether psychophysiological measurements can be used to assess prediction error magnitude, which could be a useful read-out in experimental studies (Spoormaker et al., 2012; Willems & Vervliet, 2021). While initial results are promising, the relatively large signal-to-noise ratio of physiological measurements complicates measurements at a single timepoint. One more practical method to investigate the effect of prediction errors in experimental paradigms is to manipulate the probability that an outcome occurs, thereby assuming that this will result in different magnitudes of prediction error. The important advantage of using fear-conditioning paradigms (further explained below) is that researchers have full control over the learning phase and can thus manipulate the extent to which an outcome is expected. Such manipulations are often employed in both reconsolidation and extinction studies investigating prediction errors. However, it is not a given that prediction errors linearly relate to the probability of an outcome. While this is thus by no means the most optimal measure of prediction error, throughout this thesis we have aimed to manipulate their magnitude by creating different outcome probabilities.

Investigating aversive memories in the lab

To investigate the development, maintenance, and updating of aversive memories in this thesis, we employed the Pavlovian fear-conditioning paradigm. This paradigm is considered an excellent method to study aversive memories in the lab, and is typically used to address hypotheses concerning the development and extinction of fear responses. Fear-conditioning experiments that are used to assess the effectiveness of various learning and memory-based interventions typically consist of three phases (Lonsdorf et al., 2017). During the conditioning phase, an initially neutral stimulus, the conditioned stimulus (CS+), is paired with an aversive outcome, the unconditioned stimulus (US). In our experiments, we used a mild, yet uncomfortable, electrical stimulus to the wrist as a US. Multiple pairings of the CS with this US should result in a conditioned fear response to presentations of the CS. In humans, there exist a variety of physiological measures to index the strength of this conditioned response, including fear-potentiated startle (FPS) responses, skin conductance responses (SCRs), and pupil dilation (PD) responses (Leuchs et al., 2019; Ojala & Bach, 2020). In order to assess baselines of these physiological responses, most human fear-conditioning paradigms include a second CS (the CS-) that is never followed by the US. The differential response between the CS+ and the CS- is considered a valid measure of fear responding in the lab (Lonsdorf et al., 2017). The extinction phase consists of presenting the CS+ multiple times without any reinforcement so that participants are able to learn that the CS no longer signals threat, which is usually characterised by a reduction of conditioned responses. To truly investigate the effect of extinction learning on aversive memories, extinction should take place at least one day after the conditioning phase, to allow for consolidation of the original fear memory. Notably, to investigate the effectiveness of reconsolidation-based interventions, the extinction phase is replaced with a session in which participants are only briefly presented with the CS+ (e.g., only one presentation). This brief reactivation should result in the destabilization of the original memory rather than extinction learning. Lastly, most experiments include a test phase to assess the extent to which the conditioned response returns. The return of fear can be probed with different manipulations (Bouton, 2002; Lonsdorf et al., 2017), including the delivery of unsignaled USs (reinstatement), the presentation of the CS+ in a new context (renewal), or the passage of time (spontaneous recovery). Because more effective interventions should reduce or entirely eliminate the return of fear, the extent to which fear responses return after these manipulations compared to a control group or stimulus is considered an index of the effectiveness of the intervention.

Although physiological measures of conditioned responding are typically the main outcome variables in fear-conditioning studies, it remains elusive what exact processes these measures reflect. All the measures mentioned above tend to increase to CS+ presentations relative to CS- presentations, yet they may index differences in learning about the CS+. Both skin conductance and pupil dilation responses are driven by the autonomic nervous system and reflect emotional arousal (Bradley et al., 2008). While SCRs have previously been suggested to align with probabilistic learning, they do not mirror US expectancy ratings exactly on a trial-by-trial basis, although this may also be explained by habituation of SCRs (Blechert et al., 2008). Computational modelling of SCRs during fear learning suggested that responses reflect a mix between US probability and uncertainty (Ojala & Bach, 2020; Tzovara et al., 2018). Changes in pupil dilation under constant lumination have been found to reflect noradrenergic activity in the locus coeruleus (LC), which is known to respond to environmental stressors (Joshi et al., 2016; Morris et al., 2020). While pupil dilation has on the one hand been suggested to reflect outcome probability, pupil responses are strongly related to uncertainty in non-fear learning tasks (Nassar et al., 2012; Tzovara et al., 2018; Vincent et al., 2019). Critically, the few studies that have compared SCR and pupil responses to CSs with high and low US probabilities within subjects found that SCRs increase more to consistently reinforced CSs, whereas pupil responses increase more to uncertain CSs (Koenig et al., 2017; Leuchs et al., 2017). Thus, both SCRs and pupil measurements can reliably index conditioned responses, but there is no clear consensus what these responses precisely reflect. It appears likely that both measures are sensitive to uncertainty and to outcome probability, and that this balance is determined by the exact circumstances under which learning takes place. In contrast to arousal-based measurements, fear-potentiated startle responses are suggested to index the valence or affective value of the conditioned stimulus, and to reflect more involuntary amygdala-based fear responses (Bradley et al., 2018; Ojala & Bach, 2020; Sege et al., 2014). For example, FPS responses do not respond well to instructions (i.e., telling the participants that the US no longer follows; Sevenster et al., 2012a, but see Mertens & de Houwer, 2016 for opposed effects). Furthermore, in reconsolidation paradigms, FPS responses are swiftly reduced one day later whereas expectations about the US are unaffected, showing a dissociation between FPS responses and cognitive expectations (Soeter & Kindt, 2010). In sum, FPS responses are suggested to reflect the affective value of emotional memories rather than more cognitive or probabilistic learning and have therefore more convincing translational utility than some other physiological measures.

Aim and outline of the dissertation

The work presented in this dissertation aspires to improve our understanding of the role of prediction errors in governing the effectiveness of memory reconsolidation and extinction learning in instigating long-term behavioural change. Yet as explained above, to truly advance our understanding of prediction errors on memory processes, an independent read-out of their occurrence is essential. In **Chapter 2** we therefore aimed to further develop a physiological measure of prediction error occurrence in associative learning. We reasoned that no prediction errors occur when outcomes are entirely predictable, meaning that any difference in responding to outcomes that are unpredictable versus entirely predictable can be interpreted as an effect of prediction error occurrence. In three fear-conditioning experiments, we compared skin conductance and pupil dilation responses to outcomes (US presentations and omissions) that were only 50% predicted with responses to outcomes that were 100% predicted. We further investigated whether outcome responses related to changes in conditioned responding on a future presentation of the CS, as would be expected from prediction-error based learning.

The other chapters in this thesis investigate the extent to which manipulations of prediction error magnitude or frequency affect the learning processes that are deemed critical for psychological treatments. In **Chapter 3** we report the results of a replication study on the effect of prediction errors in memory reconsolidation. Before further investigating the role of prediction errors in memory destabilization, we intended, as a proof of concept, to first replicate the critical effect of prediction error frequency on memory reconsolidation that was previously established in our lab (Sevenster et al., 2014). In line with the original study, we tested whether increasing the number of prediction errors during the memory reactivation session would critically determine the effectiveness of the reconsolidation intervention.

In **Chapter 4** we aspired to investigate the effect of enhancing expectancy violations[2] on extinction learning. The dominant inhibitory learning view of exposure posits that “the more the expectancy can be violated by experience, the greater the inhibitory learning” (Craske et

[2] The diligent reader may note that in this chapter we have departed from the term *prediction error* and instead use *expectancy violation*. This is in part to ensure consistency with the inhibitory retrieval model, in which the term expectancy violation is used, and in part because in our experiments we used US expectancy ratings to quantify our manipulation. We thus explicitly operationalised prediction errors as expectancy violations in this chapter.

al., 2014, p.12). Greater inhibitory learning should subsequently result in more durable reductions in fear responses, and therefore in better treatment outcomes. This idea has already been quite influential in recommendations for clinical practice (e.g., van Emmerik & Greeven, 2020), yet superior treatment effects as well as a detailed understanding of the fundamental processes that drive these effects remain to be established. We therefore investigated in two separate fear-conditioning experiments whether strengthening expectancy violations or fostering awareness of expectancy violations during extinction could improve extinction retention. In the first experiment our intervention consisted of manipulating the expectation of the US during extinction between two groups (100% versus 50%), and comparing the groups on return of fear one day later. In the second experiment we aimed to foster awareness of violations by asking participants in the experimental group whether "the outcome they expected had actually occurred" after every trial. While both manipulations intended to enhance the experience of expectancy violations, the underlying mechanisms of these potential effects fundamentally differ. We specifically investigated these different manipulations because on the one hand, the inhibitory retrieval model is theoretically grounded in the idea that larger expectancy violations strengthen learning (Experiment 1), whereas on the other hand, practical implementations of this model are often centred around enhancing awareness of expectancy violations (Experiment 2).

In **Chapter 5,** we re-evaluate the relevance of prediction errors in psychological treatments from a computational and neurobiological perspective, and attempt to come to a comprehensive understanding of the role of prediction errors in psychotherapy. Throughout the process of designing the experiments on prediction error, we realised that the general notion that prediction errors drive learning is often used to explain the effectiveness of psychological interventions, without a detailed description or understanding of how this process is supposed to work. The use of a term like prediction error or expectancy violation that is relatively self-explanatory and descriptive may help patients and clinicians to identify what needs to be learned during treatment. Yet to really understand the value of prediction error for learning outcomes, a more detailed and comprehensive understanding of its role in learning is necessary.

Finally, in **Chapter 6**, the theoretical and clinical implications of the results from the studies in this dissertation will be discussed, and we will further reflect on the challenges and limitations that we have encountered in our investigations. All in all, we hope that this work will contribute to a clearer understanding of the effect of prediction errors on treatment outcomes, but especially on what it means to manipulate prediction errors and how this can and cannot be achieved.