The challenge of memory destabilisation: From prediction error to prior expectations and biomarkers

Amy L. Milton a,*, Ravi K. Das b, Emiliano Merlo c

a Department of Psychology, University of Cambridge, Downing Site, Cambridge, UK
b Clinical Psychopharmacology Unit, University College London, London, UK
c School of Psychology, University of Sussex, Brighton, UK

ARTICLE INFO

Keywords:
Reconsolidation
Memory modulation
Neuropsychiatric disorders

ABSTRACT

The re-ignition of memory reconsolidation research sparked by Karim Nader in the early 2000s led to great excitement that ‘reconsolidation-based’ interventions might be developed for mental health disorders such as post-traumatic stress disorder and substance use disorder. Two decades on, it is clear that reconsolidation-based interventions have been more challenging to translate to the clinic than initially thought. We argue that this challenge could be addressed with a better understanding of how prior expectations interact with information presented in a putative memory reactivation / cue reminder session, and through the identification of non-invasive biomarkers for memory destabilisation that would allow reminder sessions to be ‘tuned’ to enhance memory lability in an ad hoc manner.

1. Introduction

The chronic and relapsing nature of mental health disorders such as post-traumatic stress disorder (PTSD) and drug addiction (substance use disorder, SUD) has long been a challenge for treatment. Available therapies mostly rely on long-term treatment with drugs (pharmacotherapies) or on behavioural therapies (such as prolonged exposure therapy), but the latter are challenging for patients to undertake. Currently available therapies achieve long-term remission in around half of patients; only marginally higher than spontaneous remission rates among untreated individuals (Morina et al., 2014). Therefore, while therapies exist for these disorders and are effective for a proportion of patients, there remains a profound, unmet clinical need for treatments with greater long-term efficacy.

A major step forward was made in the early 2000s, with the re-ignition of research into the process called memory reconsolidation, sparked by Karim Nader (Nader et al., 2000). Revisiting a phenomenon that had first been documented as ‘cue-dependent amnesia’ in the 1960s (Misanin et al., 1968; Schneider and Sherman, 1968), Nader demonstrated that it was possible to disrupt 24-hours old, well-consolidated pavlovian fear memories through re-exposure to the pavlovian fear cue, combined with inhibition of protein synthesis within the basolateral amygdala, a region known to be critical for the storage of pavlovian fear memories. Re-exposure to the fear cue or inhibition of protein synthesis alone had no effect on conditioned fear. Targeting the gold-standard cellular mechanism for associative memory consolidation (protein synthesis), in a psychologically and neurobiologically well-characterised memory model (pavlovian fear conditioning), resulted in a breakthrough for memory research and the study of long-term memory persistence.

Although research into ‘cue-dependent amnesia’ (newly termed ‘reconsolidation’ by Przybyslawski and Sara, 1997) had continued to some extent between the late 1960s and the beginning of the 21st century (see Sara and Hars (2006), for review), from 2000, there has been a marked resurgence in research into reconsolidation, with the rapid realisation that this hypothetical process could potentially be exploited to disrupt the well-established maladaptive memories that contribute to mental health disorders such as PTSD and SUD (Dunbar and Taylor, 2016; Elsey and Kindt, 2017b, Kindt and van Emmerik, 2016; Milton and Everitt, 2010, 2012). Many studies in the first years of the new period of reconsolidation research focused on identifying ‘reconsolidation blockers’: amnestic agents that would be suitable for use in humans, such as less toxic protein synthesis inhibitors (e.g. rapamycin - Barak et al., 2013; Blundell et al., 2008; Gafford et al., 2011; Zubeidat and Akirav, 2017) β-adrenergic receptor antagonists such as propranolol (Dębiec and LeDoux, 2004; Fricks-Gleason and Marshall, 2008; Kindt,...
2. Mechanisms of memory destabilisation: the importance of prediction error

Understanding of memory destabilisation mechanisms has progressed markedly from the initial studies of ‘cue-dependent amnesia’ in the 1960s. It was established early on that a key determinant of whether a memory would destabilise or not was the induction of a ‘mismatch’ between what was expected and what actually occurred (Pedreira et al., 2004) – a concept that was readily related to that of ‘prediction error’.

The literature on prediction error, and its relationship to midbrain dopaminergic signalling, is extensive and reviewed elsewhere (Schultz, 2013, 2017; Schultz et al., 1997; Waelti et al., 2001). A key point relevant to memory reconsolidation is that prediction error is necessary for destabilisation to occur (Gershman et al., 2017; Pedreira et al., 2004; Sevenster et al., 2014), which is consistent with its hypothesised function as a means for memory updating (Lee, 2009; Lee et al., 2017; Tronson and Taylor, 2007). However, the relationship between prediction error and destabilisation does not appear to be linear. Without prediction error, the memories seem to only be retrieved rather than destabilised (Ben Mamou et al., 2006; Milton et al., 2013; Santoyo-Zedillo et al., 2014; Sevenster et al., 2012, 2014). Once the putative prediction error ‘boundary’ has been crossed for memory destabilisation, there is an extent of prediction error that supports memory destabilisation (i.e. the opening of the reconsolidation window) but as prediction error increases further, the reconsolidation window appears to ‘close’ and the memory to enter a state in which it becomes once again impervious to interference. This ‘limbo’ state or ‘null point’ is dissociable at the molecular level from both memory reconsolidation and the formation of a new extinction memory, which occurs with higher levels of prediction error (Cassini et al., 2017; Merlo et al., 2018, 2014; Sevenster et al., 2014). The formation of a new extinction memory appears not to ‘undo’ neural changes that supported the original memory (though see Delamater and Westbrook, 2014) but rather involves the formation of a new, usually more contextually specific, memory that competes with the original memory for behavioural expression (Bouton, 2002). This is thought to underlie the moderate efficacy of extant extinction-based therapies (i.e. prolonged exposure). Notably, the ‘extent’ of prediction error described above is typically operationalised as the number of reminder cue presentations in the absence of reinforcement. It is therefore largely collinear with the number of presented reminder cues and temporal ‘length’ of reminder procedures; which may also play a role in determining which of the above processes is targeted.

Although it is well-established that memory destabilisation relies on prediction error (Cahill et al., 2019; Fernández et al., 2016; Pedreira et al., 2004; Sinclair and Barense, 2018) and this in turn appears to rely upon dopaminergic signalling (Merlo et al., 2015), the relationship is not simple. Dopaminergic signalling is not sufficient to enhance memory destabilisation (Flavell and Lee, 2019) and memory destabilisation also depends upon a host of other signalling mechanisms that appear to converge upon dynamic patterns of intracellular calcium signalling that bias towards protein degradation (Jarmoe et al., 2011; Kaang et al., 2009; Lee, 2008; Lee et al., 2008, 2019) – for example, a reliance on signalling at the GluN2B-subtype of NMDA receptor (Ben Mamou et al., 2006; Ferrer Monti et al., 2016; Jarmoe et al., 2011; Milton et al., 2013) or at L-type voltage-gated calcium channels (De Oliveira Alvares et al., 2013; Flavell et al., 2011). Furthermore, although there has been much focus in reconsolidation research on the contribution of prediction error to destabilisation, there has been relatively little investigation of the generation of the expectation to be violated.

3. Mechanisms of memory destabilisation: the importance of prior expectations

Prediction error can only be generated in the context of an expectation, or prior belief, that can subsequently be violated by the current sensory experience presented by the environment. In mathematical terms, this can be best conceptualised as a form of Bayesian inference, which is central to hierarchical ‘predictive coding’ models of neural processing (Knill and Pouget, 2004). A full review of these models is beyond the scope of the current manuscript (see Courville et al., 2006; Fernández et al., 2017; Soltani and Izquierdo, 2019, for more extensive discussion) but our understanding of memory destabilisation would be markedly enhanced if the contribution of prior experience to the boundary conditions of reactivation (i.e. the interaction between prior learning and the reminder) was studied. This would have both theoretical and practical importance: from a theoretical perspective, it can be hypothesised that differences in prior experience may alter the boundary conditions limiting memory destabilisation (Fig. 1). It is well-established anecdotally in the memory reconsolidation literature that successful memory destabilisation occurs when new sensory evidence is sufficiently similar to what has been learned before, but with some differences (to engage violation of expectations). However, if the sensory input is too different from what has been previously experienced, this appears to engage new learning and does not update the original memory. Thus, it can be conceptualised that the interaction between prior experience, including the precision of the prior (i.e. how specific or noisy the expectations are) may lead the same sensory input to support the formation of a new memory or updating of an old memory, depending on whether it can be accommodated within the existing prior expectations of an individual. This would yield the testable hypothesis that manipulations thought to relax precision weighting of neural priors (e.g. serotonergic psychedelics) would also enhance memory destabilisation. Further research on pre-reactivation manipulations (pharmacological or behavioural) that enhance the potential for destabilisation (e.g. Graff et al., 2014) may therefore be particularly fruitful.

4. Confirming that memory destabilisation has occurred

A key issue for preclinical studies of memory reconsolidation, with clear clinical relevance, is the determination of whether memory destabilisation has occurred following a putative memory reactivation session or reminder procedure. As has been noted before (Elsey et al., 2018), there is a challenge with the interpretation of apparent failures to observe memory reconsolidation interference effects. Any study that shows an amnestic effect following a mnemonic disruption technique – pharmacological or behavioural – given in conjunction with memory reactivation potentially supports memory reconsolidation theory, but any lack of effect can potentially be attributed to a failure to engage memory destabilisation, bringing the reconsolidation theory perilously close to unfalsifiability. To avoid this philosophical and practical problem, attempts have been made to identify memory destabilisation biomarkers that would indicate whether or not reconsolidation has
occurred or not, independently of the results of pharmacological or behavioural interventions.

It is well-established that memory destabilisation requires cellular protein degradation machinery, including the ubiquitin proteasome system (UPS; Jarome et al., 2016; Jarome and Helmstetter, 2013; Lee, 2008; Sol Fustiñana et al., 2014). Reassuringly, receptors that have been identified as necessary for memory destabilisation, including GluN2B-containing NMDA receptors and L-type voltage-gated calcium channels, affect intracellular changes in calcium concentration with appropriate dynamics to engage both protein phosphatases and the proteasome system (Ferreira et al., 2021, 2015). More recently, specific proteins that are targeted for degradation during the destabilisation process have been identified, including the post-synaptic density scaffolding protein Shank (Jarome and Helmstetter, 2013; Lee et al., 2008; Rotondo et al., 2022). Memory lability is associated with reduced Shank expression for 1–2 h after memory reactivation, and UPS inhibition by clasto-Lactacystin-β-lactone prevents this reduction in Shank and the induction of memory lability (Kaang et al., 2009; Lee, 2008).

Although Shank is a useful marker of memory destabilisation, it can only be measured post mortem. This is not a problem for preclinical research, but renders it unusable within the clinical setting and makes it impossible to relate neural expression to corresponding long-term behavioural effects within animals. What would be needed within the clinical environment is a destabilisation marker determined by a non-invasive method applied to awake, behaving individuals. This biomarker would also need to be specific for memory destabilisation, as compared to other mnemonic processes such as memory retrieval or extinction.

To date, no such biomarker has been identified. Clinical studies have used proxies such as self-report measurements of whether expectations have been violated (Elsey and Kindt, 2017a). These have been sufficiently sensitive to provide context for apparent failures to replicate amnestic effects: for example, in one study using a reconsolidation-based approach to weaken memories underlying alcohol-seeking, participants were told that they would be required to drink (unknowingly to them, non-alcoholic) beer during the reactivation session. The reactivation session differed between groups, such that it should engage prediction error (i.e. participants were instructed at the last moment that they should not drink the beer) or not (i.e. participants were allowed to drink the beer). When assessed at the group level, the amnestic agent nitrous oxide (an NMDA receptor antagonist) was ineffective at weakening cue-alcohol memories in this study (Das et al., 2018b). However, when the participants’ self-reported expectations of whether they would be allowed to drink the beer were factored into the group assignments, those for whom their expectations had been violated did show a subsequent reduction in beer-drinking when they had received nitrous oxide at reactivation compared to both placebo-treated participants, and those who had received nitrous oxide but had not had their expectations violated at reactivation (Das et al., 2018b). Thus, self-reported expectations of outcomes can be useful in interpreting reconsolidation-based interventions, but they still suffer from the limitation that they can only be applied retrospectively (i.e. after reactivation and an amnestic intervention) and therefore do not offer a means to tune reactivation procedures to increase the likelihood of destabilisation in an ad hoc manner.

5. Hopes for the future: a destabilisation marker in awake, behaving individuals

Considering the limitations of the currently available destabilisation biomarkers (at least from a clinical, if not empirical, perspective), we would like to encourage research into the identification of a destabilisation biomarker that could be used in awake, intact, behaving individuals. This could be used in a clinical setting to facilitate that destabilisation is occurring in an online manner such that an amnestic intervention could be administered at an appropriate timepoint.

Although this will need to be a topic of future research, we would suggest the following as criteria for a suitable destabilisation biomarker:

1. Should be observable in awake and behaving individuals, requiring any measurements to be non-invasive;
2. Should be reliably distinguishable from other memory processes (e.g. memory retrieval or extinction) to identify memory destabilisation specifically;
3. Should be sufficiently reliable that the same algorithm can be used to identify a destabilised memory despite differences in prior experience or training history.

Given recent developments in brain-computer interfaces and inferring covert neural states, it is our opinion that the most likely approach to satisfy these criteria will come from a combination of non-invasive neuronal activity recordings (e.g. electroencephalography or magnetoencephalography) and machine learning to identify specific patterns of brain activity associated with a destabilising memory. This would need
to be tested in both human and non-human animals (since the majority of direct pharmacological reconsolidation-blockade studies have been conducted in animals), and would need to be tested across different types of memory; for example, aversive vs. appetitive, and pavlovian vs. instrumental. The different behavioural tasks used to assess these types of memory would also allow for investigation into the contribution of differences in prior learning history on memory destabilisation. For example, aversive tasks are often learned within a single session of pavlovian conditioning, while appetitive tasks are often learned across multiple sessions. It may be the case that a single destabilisation biomarker does not exist across these different memory types, but considering the similarities in neural mechanisms of reconsolidation that have already been observed, it would be parsimonious to take this as a starting point. This will require concerted effort from preclinical and clinical reconsolidation researchers, but may hold the key to realising the potential of reconsolidation-based therapies.

6. Conclusions

Nader’s transformative rediscovery of memory reconsolidation has profoundly changed the types of treatments being developed for chronic and relapsing mental health disorders. Much progress has been made in the characterisation of mechanisms of memory reconsolidation and the identification of amnestic treatments (pharmacological and behavioural), and our understanding of the mechanisms of memory destabilisation is developing. The investigation of electrophysiological correlates occurring during memory destabilisation could provide not only a bridge between the molecular and behavioural data that have been collected to date, but could also provide a means for identifying appropriate intervention points during therapy for patients. A fitting tribute and development in this second decade of ‘new’ reconsolidation research would be the agreement of a set of criteria for a destabilisation biomarker, and collaboration across labs researching different memory types, in different species, to unravel the core mechanisms by which memories become unbound despite differences in expectations.

CRediT authorship contribution statement

Amy L Milton: Conceptualisation, Writing — original draft, Writing — review & editing. Ravi K Das: Conceptualisation, Writing — review & editing. Emilio Merlo: Conceptualisation, Writing — review & editing.

Data Availability

No data was used for the research described in the article.

Acknowledgements

ALM’s research is supported by funding from UKRI (MR/N02530X/1 and BB/W001195/1). RD’s research is supported by funding from the AMS and UKRI (MR/R004919/1). EM is supported by the University of Sussex School of Psychology start-up funds. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.


Lee, J.L.C., Everitt, B.J., 2008. Appetitive memory reconsolidation depends upon NMDA receptors in the destabilization and restabilization of a reconsolidating memory. eNeuro 2, 1


