NeuroImage 132 (2016) 11–23

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Reverse inference of memory retrieval processes underlying metacognitive monitoring of learning using multivariate pattern analysis

Peter Stiers [a],*, Luciana Falbo [a], Alexandros Goulas [a], Tamara van Gog [b], Anique de Bruin [c]

[a] *Department of Neuropsychology and Psychopharmacology, Maastricht University, Maastricht, The Netherlands*
[b] *Department of Educational Psychology, Erasmus University Rotterdam, The Netherlands*
[c] *Department of Educational Research & Development, Maastricht University, The Netherlands*

ARTICLE INFO

Article history:
Received 15 March 2015
Accepted 7 February 2016
Available online 13 February 2016

Keywords:
Working memory
Judgement of learning
Metacognition
Functional MRI
Educational psychology

ABSTRACT

Monitoring of learning is only accurate at some time after learning. It is thought that immediate monitoring is based on working memory, whereas later monitoring requires re-activation of stored items, yielding accurate judgements. Such interpretations are difficult to test because they require reverse inference, which presupposes specificity of brain activity for the hidden cognitive processes. We investigated whether multivariate pattern classification can provide this specificity. We used a word recall task to create single trial examples of immediate and long term retrieval and trained a learning algorithm to discriminate them. Next, participants performed a similar task involving monitoring instead of recall. The recall-trained classifier recognized the retrieval patterns underlying immediate and long term monitoring and classified delayed monitoring examples as long-term retrieval. This result demonstrates the feasibility of decoding cognitive processes, instead of their content.

Introduction

A key aspiration of psychology is to understand complex human behaviour in terms of its constituent psychological processes. The metacognitive ability to monitor one's own state of learning, for example, is thought to be an essential aspect of academic learning (Koriat and Goldsmith, 1996; Thiede and Dunlosky, 1999). It is assumed to involve, a try-out re-activation of the learned material, informing the student about the acquisition status (Nelson and Dunlosky, 1991; Rhodes and Tauber, 2011; Thiede and Anderson, 2003). This assumption is supported, on the one hand, by a loss in relative accuracy of monitoring when performed immediately after the learning (Nelson and Dunlosky, 1991; Rhodes and Tauber, 2011; Thiede et al., 2003; Thiede et al., 2005). In that case, the judgement is thought to be based on information still active from the encoding, independent of the quality of storage in long term memory. On the other hand, restricting the opportunity for retrieval eliminates the higher relative accuracy of delayed monitoring (Dunlosky and Nelson, 1992).

While these findings lend support to the re-activation theory of the delayed judgement of learning effect (Rhodes and Tauber, 2011) the hypothesized processes are only indirectly open to empirical observation, through behavioural measures. Neuroimaging, however, holds the promise of making the hidden processes more directly identifiable through their neurophysiological markers. Long term memory retrieval, for instance, has been associated with increased activity in the hippocampus and in the lateral and medial parietal cortex (Cabeza et al., 2012; Daselaar et al., 2009; Huijbers et al., 2012; Kirwan and Stark, 2004; Okada et al., 2012; Vannini et al., 2011). Likewise, working memory retrieval has been associated with stronger activity in ventral lateral prefrontal cortex, anterior superior frontal gyrus and lateral temporal cortex (Nee and Jonides, 2011, 2013; Oztekin et al., 2009). Such an activation pattern may be used as a biological marker for the underlying process, allowing to conclude that, for instance, memory retrieval took place every time the characteristic pattern is observed. This type of deduction, known as “reverse inference” (Aguirre, 2003), holds the promise of an alternative route to the dissection of complex behaviour.

A first challenge for reverse inference of long-term memory retrieval during monitoring of learning is the superimposed processes related to metacognitive monitoring and the judging response, which are likely to distort the overall brain activation pattern. Metacognition has, for instance, been shown to elicit specific activation in subregions of medial and orbital prefrontal cortex (Chua et al., 2009; Do Lam et al., 2012; Kao et al., 2005). Hence, an effective reverse inference procedure must be able to distinguish activity specific of the target processes from everything else.

This brings us to a second and more fundamental challenge. Several authors have criticized reverse inference on more principle grounds. This critique focuses on the notion of process-specific activity (Aguirre, 2003; Christoff and Owen, 2006; D'Esposito et al., 1998; Poldrack, 2006, 2011; Fox and Friston, 2012). They argue that a

* Corresponding author at: Maastricht University, Faculty of Psychology and Neuroscience, Department of Neuropsychology and Psychopharmacology, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
E-mail address: peter.stiers@maastrichtuniversity.nl (P. Stiers).

http://dx.doi.org/10.1016/j.neuroimage.2016.02.008
1053-8119/© 2016 Elsevier Inc. All rights reserved.

consistent activity increase in particular brain structures during execution of a particular cognitive process does not tell us whether the structures are indicative of or selectively engaged by the processes under study. As an illustration of this, the well-established attribution of long term memory storage to the hippocampus was criticized recently by demonstrations of activation in this structure also during working memory retrieval (Nee et al., 2008; Nee and Jonides, 2011; Postle, 2006; Ranganath and Blumenfeld, 2005). Likewise, the traditional attribution of working memory processes to the ventral–lateral prefrontal cortex (e.g., Fuster, 1989; Fuster and Alexander, 1971; Goldman-Rakic, 1987; Miller et al., 1996; Ptito et al., 1995), has been challenged by studies showing that activity in this part of prefrontal cortex reflects attention control mechanisms that are not specific to working memory (D'Esposito et al., 1998; D'Esposito and Postle, 1999; Nee et al., 2008; Passingham et al., 2000; Postle, 2006). It should be clear from this that lacking knowledge of the specificity of neural structures or brain activation patterns for particular cognitive processes imposes limits to reverse inference. Practically, the selectivity and specificity of activation patterns for a particular process can be investigated using available large-scale brain activation databases (e.g., BrainMap.org, NeuroSynth.org). Such databases allow to estimate the probability of activation given the execution of tasks thought to activate the process and the execution of tasks that should not activate the process (e.g., Chang et al., 2013; Poldrack, 2006).

While databases of published data provide a practical by-pass for our limited knowledge, they also point towards a third problem of reverse inference, recently addressed by Hutzler (2014). Ideally, the functional signature for a particular cognitive process should allow inferring the involvement of the process in any context. In reality, however, its validity is restricted to the contexts used to establish the characteristic signature, and validity beyond these contexts needs to be established empirically. Hutzler (2014) showed that this limitation can be turned into an advantage. By explicitly taking the context of the task under study into account, justified inferences can be made about processes taking place within this specific context. This was investigated for the (left) fusiform face area, which is known to activate during both face recognition and reading tasks. However, when the studied task involves visually presented words (e.g., a reading task without pictures of faces) it is safe to infer that activity in this area marks processing of word images. Consequently, a quantitative data-base driven meta-analysis of experiments using visual-verbal tasks can be sufficient to yield the voxels that are uniquely associated with the process in this type of tasks. The advantage here is that specific reverse inference questions about processes underlying particular task paradigms can get a quantitative answer, without the requirement to answer the most general question of the unique functional signature of the processes under all possible contexts.

In the present paper we follow this line of restriction to address the problem of the hypothesized long-term retrieval process underlying the typical delayed judgement of learning task paradigm. However, we do not rely on meta-analysis of already existing data to delineate voxel activation patterns with significant predictive power of reverse inference. Instead, we make use of the typical task paradigm to collect new data and use multivariate pattern classification as our method to find the indicative activation pattern. It was Poldrack (2011) who suggested that multivariate pattern classification could provide a formal means to implement reverse inference, because these methods quantitatively estimate the degree to which a pattern of brain activation is predictive of the engagement of a specific cognitive process. These methods use brain activation maps derived under two (or more) prototypical conditions as examples to train a statistical machine learning algorithm to find the optimal pattern to distinguish the example classes. The trained classifier is subsequently used to make predictions about the activation patterns in a new set of similar examples (O'Toole et al., 2007; Pereira et al., 2009). These techniques have been used to predict which stimulus classes participants were viewing, or imagining (Haxby et al., 2001; Kamitani and Tong, 2005; Lewis-Peacock and Postle, 2008; Lewis-Peacock et al., 2012). The techniques were also successful in predicting more cognitive aspects of behaviour, such as the intention to perform one or the other task (Haynes et al., 2007), the stimulus–response mapping rules in a task (Woolgar et al., 2011), and which of a set of tasks was being performed (Poldrack et al., 2009; Stiers et al., 2010).

To support reverse inference of cognitive processes, however, the training examples need to reflect as closely as possible the theoretically relevant difference: cognitive processes, rather than cognitive contents (stimulus classes, response classes, task rules, etc.). Under these circumstances, the training set constitutes an ostensive definition of the brain functioning patterns that distinguish the two cognitive processes. The multivariate pattern analysis translates this defining set into a high-dimensional statistical pattern, which can be applied to brain activation examples generated during tasks where the underlying processes are unknown. Thus, in the multivariate pattern classification approach, instead of relying on large-scale data bases (Poldrack, 2006, 2011) or task-specific meta-analyses (Hutzler, 2014), new data are collected that are specific to the process and paradigm of interest and the critical alternative processes, and the predictive patterns are generated from these data. The selectivity and specificity of the pattern for inferring the process of interest can then be computed from the classification accuracies in the reference task.

The first aim of this study was to investigate the feasibility of multivariate pattern classification for reverse inferring. We adopted a single participant approach (Formisano et al., 2008), because people may use different strategies to learn and evaluate their state of learning, and consequently manifest idiosyncratic activation patterns. Randomization statistics were applied to establish above chance classification performance. To make sure that results are not subject dependent we repeated the analysis independently in five different individual data sets. We tested the feasibility of reverse inference in two steps. In the first experiment the decoding of immediate and long term retrieval was validated using an overt cued recall task for word pairs. The aim was to show that the classification procedure had sufficient selectivity and specificity to correctly infer the known retrieval processes underlying overt recall. In the second experiment reverse inference was critical put to test by having participants perform judgements of learning of word pairs, instead of overt recalling them. Reverse inference would be established if judgements of items stored in long term memory prior to the task are recognized as long term memory retrievals, while identical judgements made immediately after encoding of the content are not.

The second aim of our study was to investigate the long term memory retrieval interpretation of the delayed judgement of learning effect. Conditional to the confirmation that multivariate pattern analysis allows reverse inference of long-term memory retrieval underlying monitoring of long term learning, the re-activation hypothesis would predict that the classifier, trained on immediate and long term overt recall activation patterns, would also recognize in the delayed monitoring trials the long term memory retrieval pattern.

Materials and methods

Participants

Six healthy right-handed volunteers (2 males, mean age 26.94 (3.69) years) took part in the study after giving their written informed consent. The study was approved by the local Ethical Committee. Participants were recruited from the university community and screened for psychological and medical problems, right-handedness and absence of contra-indications for exposure to magnetic field. Due to technical failure data recorded from one participant during the monitoring task were lost. Hence, for this task data from only 5 subjects were available for analysis.