Journal of Memory and Language 94 (2017) 149–165

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Does prediction error drive one-shot declarative learning?

Andrea Greve *, Elisa Cooper, Alexander Kaula, Michael C. Anderson, Richard Henson

MRC Cognition & Brain Sciences Unit, Cambridge, England, United Kingdom

ARTICLE INFO

Article history:
Received 19 February 2016
revision received 31 October 2016

Keywords:
Prediction error
Associative memory
Encoding
One-shot learning

ABSTRACT

The role of prediction error (PE) in driving learning is well-established in fields such as classical and instrumental conditioning, reward learning and procedural memory; however, its role in human one-shot declarative encoding is less clear. According to one recent hypothesis, PE reflects the divergence between two probability distributions: one reflecting the prior probability (from previous experiences) and the other reflecting the sensory evidence (from the current experience). Assuming unimodal probability distributions, PE can be manipulated in three ways: (1) the distance between the mode of the prior and evidence, (2) the precision of the prior, and (3) the precision of the evidence. We tested these three manipulations across five experiments, in terms of peoples' ability to encode a single presentation of a scene-item pairing as a function of previous exposures to that scene and/or item. Memory was probed by presenting the scene together with three choices for the previously paired item, in which the two foil items were from other pairings within the same condition as the target item. In Experiment 1, we manipulated the evidence to be either consistent or inconsistent with prior expectations, predicting PE to be larger, and hence memory better, when the new pairing was inconsistent. In Experiments 2a–c, we manipulated the precision of the priors, predicting better memory for a new pairing when the (inconsistent) priors were more precise. In Experiment 3, we manipulated both visual noise and prior exposure for unfamiliar faces, before pairing them with scenes, predicting better memory when the sensory evidence was more precise. In all experiments, the PE hypotheses were supported. We discuss alternative explanations of individual experiments, and conclude the Predictive Interactive Multiple Memory Signals (PIMMS) framework provides the most parsimonious account of the full pattern of results.

Introduction

Animals constantly extract regularities from past experiences to enable predictions about future events. Given that their environment is continuously changing, these predictions likewise need to adapt to novel information that may conflict with previously acquired expectations. The degree of conflict between predictions and new information is called prediction error (PE). PE plays a key role in many domains, such as reward learning, motivational control and decision making (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000; Sutton & Barto, 1998). Formal associative learning theories, for instance, state that learning is proportional to PE, where PE is the difference between expected and actual reward (Beesley & Shanks, 2012; Rescorla & Wagner, 1972). The recently proposed 'Predictive Interactive Multiple Memory Signals' (PIMMS) framework (Henson & Gagnepain, 2010) suggests that PE plays a general role throughout the human brain, in the service of both perception and multi-

* Corresponding author at: MRC Cognition & Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom. Fax: +44 1223 359 062.
E-mail address: andrea.greve@mrc-cbu.cam.ac.uk (A. Greve).

http://dx.doi.org/10.1016/j.jml.2016.11.001
0749-596X/

ple types of memory, from conditioning to perceptual priming and even declarative (e.g., episodic) memory.

Predictions are central to most computational models of learning. PE is the basis of the "delta" learning rule used in many connectionist models of human memory (Kinder & Shanks, 2001), though some have argued that such models cannot capture the one-shot learning that is characteristic of human declarative memory (Reber, 2002; although see Kinder & Shanks, 2003). More recent computational work suggests that the precision of predictions is an important determinant of the learning rate (e.g., Yu & Dayan, 2005), with the certainty of predictions possibly triggering a switch between gradual and rapid (e.g., one-shot) learning systems (Lee, O'Doherty, & Shimojo, 2015). Although PE is often taken for granted as a driver of learning, direct behavioural evidence for its role in one-shot learning of unique stimulus-stimulus associations, however, is scarce.

The most relevant work is by Le Pelley and colleagues (Griffiths & Mitchell, 2008; Le Pelley, 2010; Le Pelley, Beesley, & Suret, 2007; Le Pelley & McLaren, 2003), in particular their studies examining the impact of trained predictive and non-predictive cues on learning (Le Pelley, 2010). Across several experiments, these authors demonstrated more rapid discrimination learning when cues were previously established to be predictive. However, the content of the predictions was unrelated to the content of the information subsequently learned, leaving open the question of the role PE plays in learning. Moreover, most of these studies were based on animal learning paradigms, which differ in several ways from the paired-associate tradition often used to measure human declarative memory. Firstly, the animal learning paradigms normally involve multiple learning trials, rather than the single-trial learning. Secondly, these paradigms generally pair each stimulus with one of two outcomes (e.g., A+, B+, C−, D−, where + and − signify presence or absence of reward), rather than pairing two unique stimuli, as here (e.g., A−B, C−D, etc.). Perhaps most importantly, the animal paradigms tend to involve multiple stimuli (cues) associated with an outcome (e.g., AB+, AC−). With multiple cues, factors like selective attention become more important, in that participants can devote relatively more attention to those cues that are more predictive (Le Pelley, 2010; see also Kruschke, Kappenman, & Hetrick, 2005).

Other related work has investigated the effect of temporal context predictions on human memory, using repeated sequences or sub-sequences. These findings include better encoding into short-term memory for items that are less well predicted (Farrell & Lewandowsky, 2002) and better memory for items whose temporal context is repeated, even if the items themselves are not (Smith, Hasinski, & Sederberg, 2013), provided the prediction is strong (else the memory can actually be weakened, Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014). However, none of this work has systematically tested long-term memory for multiple, unique stimulus-stimulus associations, nor manipulated PE by independently varying the precision and the accuracy of predictions.

PIMMS is a general framework for understanding how prior knowledge influences the perception and acquisition of new information. The brain is assumed to contain hierarchical representations of the world, where representations active at one level of the hierarchy predict the activity of representations in lower levels. The difference between those predictions and the sensory evidence from lower levels comprises the PE, and this PE is assumed to drive synaptic change (learning) between levels, so as to improve predictions and ultimately minimise PE in future (Friston, 2005; Mumford, 1992). PIMMS does not specify the precise mechanism by which PE drives learning, e.g., whether that be "local" PE in a delta-learning rule that updates individual weights, "global" PE that modulates overall learning rate across all weights, or even whether PE only triggers increased attention, and it is attention that actually mediates memory encoding. What PIMMS offers is a Bayesian framework for considering how PE might vary in the world, and therefore be manipulated experimentally in the laboratory. As an example, PIMMS assumes that walking into a familiar room activates a representation of that room, which in turn predicts what objects are expected inside the room. Objects that are present and predicted do not produce a PE, and no learning results; an unexpected (but familiar) object, however, produces a PE, which causes learning to update the predictions so that the object becomes more expected in that room in future. Walking into a completely novel room, on the other hand, does not generate predictions, so a novel object in a novel room will not be encoded because there is no PE (despite the maximal novelty of the situation). A familiar object in a novel room, on the other hand, does produce a PE and its association with that room will be learned faster than a novel object.

In the present study, we set out to provide evidence for this core assumption of PIMMS that PE drives one-shot learning within a more typical human associative memory (paired associate) laboratory paradigm. We conducted five behavioural experiments that measured memory for a single pairing of two visual stimuli, as a function of the prior history of those stimuli. According to PIMMS, PE can be defined as the divergence between an expected outcome (prior) and an observed outcome (evidence); also called the "Bayesian Surprise" (Friston, 2010). Assuming unimodal probability distributions, Fig. 1 illustrates three ways in which this PE can be manipulated. Firstly, one can vary the accuracy of the prediction, i.e., the difference between the modes of the prior and evidence distributions, with a larger difference producing greater PE (Fig. 1a). This is akin to a room predicting a familiar object that is different from the one encountered there. This is the approach we took in Experiment 1, by establishing predictions (priors) for the valence of a word given a category of scenes in a Training Phase, and then testing associative memory for new scene-word pairings presented in a second "Study" phase, where the new word (evidence) was either consistent (low PE) or inconsistent (high PE) with the valence predicted by the trained scene category. Importantly, we tested memory with three-alternative forced choice (3AFC), in which all the response options came from the same Study phase, in order to prevent proactive interference from the Training phase.

A second way to manipulate PE is to vary the precision of the prediction (i.e., sharpness of the prior distribution;