NeuroImage 87 (2014) 61–71

Contents lists available at ScienceDirect

NeuroImage

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

Attention, predictive learning, and the inverse base-rate effect: Evidence from event-related potentials

Andy J. Wills [a,*], Aureliu Lavric [b], Yvonne Hemmings [b], Ed Surrey [b]

[a] School of Psychology, Plymouth University, Drake Circus, Plymouth PL4 8AA, UK
[b] School of Psychology, Exeter University, Exeter EX4 4QG, UK

ARTICLE INFO

Article history:
Accepted 26 October 2013
Available online 2 November 2013

Keywords:
EEG
Visual attention
Learning
Categorization

ABSTRACT

We report the first electrophysiological investigation of the inverse base-rate effect (IBRE), a robust non-rational bias in predictive learning. In the IBRE, participants learn that one pair of symptoms (AB) predicts a frequently occurring disease, whilst an overlapping pair of symptoms (AC) predicts a rarely occurring disease. Participants subsequently infer that BC predicts the rare disease, a non-rational decision made in opposition to the underlying base rates of the two diseases. Error-driven attention theories of learning state that the IBRE occurs because C attracts more attention than B. On the basis of this account we predicted and observed the occurrence of brain potentials associated with visual attention: a posterior Selection Negativity, and a concurrent anterior Selection Positivity, for C vs. B in a post-training test phase. Error-driven attention theories further predict no Selection Negativity, Selection Positivity or IBRE, for control symptoms matched on frequency to B and C, but for which there was no shared symptom (A) during training. These predictions were also confirmed, and this confirmation discounts alternative explanations of the IBRE based on the relative novelty of B and C. Further, we observed higher response accuracy for B alone than for C alone; this dissociation of response accuracy (B > C) from attentional allocation (C > B) discounts the possibility that the observed attentional difference was caused by the difference in response accuracy.

Introduction

We seem to learn more about events for which our initial predictions were incorrect than we do about events for which our initial predictions were correct—the element of surprise seems conducive to learning (Kamin, 1969). Using an event-related potential methodology, Wills et al. (2007) provided evidence that one process underlying this phenomenon is the rapid re-direction of visual attention in response to prediction errors. Specifically, Wills et al. (2007) found a brain potential previously associated with attention to features (e.g. shape, color, spatial frequency)—the selection negativity (SN)—for a cue involved in multiple prediction errors, relative to an equally frequent control cue involved in fewer prediction errors. In the current article, we report that a comparable event-related component is observed in the inverse base-rate effect—a robust non-rational preference observed in post-category-learning decision making (Medin and Edelson, 1988).

The inverse base-rate procedure, in its canonical form, can be considered both as a category-learning phenomenon (because it involves inference from learned items to unseen items, see Pothos and Wills, 2011), and a predictive learning phenomenon (because it involves learning to predict outcomes on the basis of presented stimuli). For this reason, we use the terms 'predictive learning' and 'category learning' inter-changeably in the current article, although we accept that they are not entirely synonymous when considering the associative- and category-learning literatures in their entirety (see e.g. Bott, Hoffman and Murphy, 2007).

In the sections that follow, we describe the inverse base-rate effect, explain how the effect may be accommodated by theories of error-driven attention, and justify our prediction of the presence of a SN on the basis of these theories and related work. An experiment testing this prediction is then reported.

The inverse base-rate effect

Imagine the following fictitious scenario. You are a physician in training who has just seen a series of patients. You have noticed that all patients with the symptoms dizziness and skin rash have Jominy fever, whilst all patients with dizziness and back pain have Phipp's syndrome. You have seen three times as many cases of Jominy fever as you have of Phipp's syndrome. The next patient you see has back pain and skin rash. Is this patient more likely to have Jominy fever or Phipp's syndrome?

When posed the question in this manner, people typically answer that Jominy fever is more likely (Johansen et al., 2007). Such an answer is not unreasonable because, in the microcosm of this scenario, skin rash

* Corresponding author.
E-mail addresses: andy.wills@plymouth.ac.uk (A.J. Wills), a.lavric@exeter.ac.uk (A. Lavric), yvonne.hemmings@btinternet.com (Y. Hemmings), edward.surrey@gmail.com (E. Surrey).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.10.060

perfectly predicts Jominy fever, and back pain perfectly predicts Phipp's syndrome, but Jominy fever is more common overall. Indeed, medical students are often encouraged to heed the aphorism "when you hear hoof beats behind you, don't expect to see a zebra" (Imperato, 1979). In the presence of two perfectly predictive but conflicting symptoms, the underlying base rates of the diseases provide one basis on which to make a decision. The current article focuses on the opposite result where participants respond that a patient with back pain and skin rash is more likely to have the rare disease Phipp's syndrome. This non-rational response bias is robustly found when participants are presented with the same information sequentially as a series of cases (e.g. Juslin et al., 2001; Kruschke, 1996; Lamberts and Kent, 2007; Medin and Edelson, 1988; Sherman et al., 2009).

One class of theory of this inverse base-rate effect (IBRE) is that it is a relative-novelty effect (Binder and Estes, 1966). This theory combines the idea that novel or surprising events are particularly memorable (Rhetorica ad Herennium, 85BC; Von Restorff, 1933), with the availability heuristic (Tversky and Kahneman, 1973), which states that memorable events are judged more probable. The idea that the IBRE is driven by the relative novelty of the two diseases is disconfirmed by the fact that participants predict the common disease if presented with just the symptom common to both diseases (dizziness), a response that is consistent with the underlying base rates. Participants also predict the common disease if presented with all three symptoms (dizziness, skin rash and back pain; see e.g. Kruschke, 1996); this response is also consistent with the underlying base rates and inconsistent with a relative disease novelty account of the IBRE.

Another variant of the relative-novelty explanation of the IBRE focuses on the relative novelty of the symptoms. The symptom back pain is relatively novel in this scenario compared to skin rash, which makes it more memorable, and hence its associated disease (Phipp's syndrome) is judged more probable. However, this version of a relative-novelty account is disconfirmed by the observation that the IBRE is only observed if there is a shared cue during training (Kruschke, 2001a; Medin and Edelson, 1988; Medin and Robbins, 1971). The shared cue in the above example is dizziness, which occurs in all presented cases. If the shared cue is replaced by further perfectly predictive cues, base-rate following is observed. For example, if dizziness and skin rash predict the common disease Jominy fever, but ear ache and back pain predict the rare disease Phipp's syndrome, then participants' modal response to the symptom combination skin rash and back pain is now Jominy fever, in agreement with the underlying base rates. Under a relative novelty account, the IBRE should still be observed, because back pain is more novel than skin rash. In summary, the shared-cue effect disconfirms the relative-novelty account of the IBRE.

The shared-cue effect also disconfirms the eliminative-inference account suggested by Juslin et al. (2001). For an extended discussion of this point see Kruschke (2001a) but, in essence, the eliminative-inference account proposes that participants are more likely to remember what skin rash predicts than what back pain predicts because they see skin rash more often. Faced with novel symptom combination skin rash and back pain, participants may therefore forget what back pain predicts (the rare disease) but remember what skin rash predicts (the common disease). However, skin rash plus back pain is a novel symptom combination and participants are assumed (under eliminative inference theory) to respond to this novel combination with a novel response. Specifically, they respond that skin rash and back pain predict the rare disease, because this is a novel response (responding "common disease" would be the familiar response because it is brought to mind by the more frequent symptom skin rash). Such a theory applies equally in the presence or absence of a shared cue, yet the IBRE effect depends on the presence of a shared cue. Hence, the shared-cue effect disconfirms the eliminative-inference account of the IBRE.

Although the above examples of the IBRE involve verbal descriptions of symptoms within a fictitious medical scenario, the IBRE has also been observed with abstract pictorial stimuli, and in non-medical scenarios (Binder and Estes, 1966; Johansen et al., 2010; Kalish, 2001; Lamberts and Kent, 2007; Sherman et al., 2009). We therefore subsequently discuss the IBRE and the shared-cue effect in terms of their abstract structure, which is summarized in Table 1. In Table 1, A is the shared cue, B and D are perfect predictors of the common disease *com*, C and E are perfect predictors of the rare disease *rare*, and F and G are further perfect predictors whose main role is to replace the shared cue. The result that the rare outcome is more likely to be predicted than the common outcome in response to a particular cue combination can be represented as: *rare* > *com*. Thus, the three key results of the IBRE and shared-cue effect, expressed in terms of the abstract design of Table 1 are (1) *com* < *rare* for BC, (2) *com* > *rare* for DE, and (3) *com* > *rare* for A. In interpreting Table 1, it is important to note that compounds (e.g. AB) are presented simultaneously — in other words, the two component cues (e.g. A and B in AB) appear on the screen at the same time. It is also important to note that trial order is randomized, and thus the order of the rows in Table 1 is arbitrary.

Error-driven attention

Certain error-driven attention theories of learning (e.g. Kruschke, 2001b) can accommodate both the IBRE and the shared-cue effect. These theories are expressed in mathematical terms but, for current purposes, a natural-language approximation (Wills and Pothos, 2012) will suffice. The central concept behind these theories of error-driven attention is that people re-direct their attention to particular components of a presented stimulus in order to minimize future prediction errors. In the context of the IBRE, one has to make the additional assumption that participants learn more quickly about what predicts the common outcome than about what predicts the rare outcome. Such an assumption is not unreasonable given that participants see the common disease more often, and it is supported by previous studies of the IBRE (e.g. Kruschke, 1996, Fig. 1).

In approximate terms, the explanation provided by error-driven attention theory on the basis of these premises is as follows. Relatively early in the case series, participants learn AB → *com*. This leads them to initially predict AC → *com*, because of the similarity of AC to AB. The participant's prediction turns out to be wrong, because AC → *rare*. The participant concludes that it was cue A that led to this erroneous prediction (nothing has been learned about C yet). Error-driven attention theory states that people act to reduce the likelihood of a subsequent error in predicting the outcome of AC by reducing the attention paid to A and increasing the attention paid to C. The cue B does not see a corresponding increase in attention, because the participant has already learned AB → *com*. When AB was originally learned, the participant knew nothing about A or B, so any initial errors would not lead to B being differentially attended relative to A.

When subsequently asked about the cue combination BC, these error-driven changes in attention are assumed to persist, and thus C attracts more attention than B. This difference in attention is presumably sufficiently large that C (which is associated with the rare disease) dominates the decision. Note that this explanation of the IBRE, like the IBRE itself, depends on the presence of the shared cue A. In the absence of A, base-rate following is expected because there is no shared cue to cause the re-direction of attention, and the participant has had more opportunity to learn about D than E, because D occurs more often. For similar

Table 1
Abstract design.

Frequency	Symptoms → Disease
2	AB → com
1	AC → rare
2	FD → com
1	GE → rare