Neuroscience and Biobehavioral Reviews 80 (2017) 223–239

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Review article

A conditioning model of delusion

Douglas John Roy

School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Keywords:
Frustration theory
Partial reinforcement extinction effect
Counter-conditioning
Latent inhibition
Learned irrelevance
Conditioned attention theory
Psychoticism

ABSTRACT

"Delusions" are beliefs that are false and persistent. It is suggested here that these characteristics can emerge from interplays between two fundamental learning processes: (1) the allocation of attentional resources among stimuli; and (2) the effects of feedback on learning. The former of these has been operationalized in the *learned irrelevance* and *latent inhibition* paradigms; the latter in studies of the effects of *persistence-training*. Normally, the attentional process functions to constrain persistence-training effects so that only valid associations acquire persistence. But when persistence-training is less influenced in this way, its mechanisms can interact with a noisy environment to gradually insulate maladaptive associations from disconfirming feedback. When unchecked, these dynamics likely lead to a systematic distortion of beliefs that can become increasingly persistent regardless of their validity. Delusions are therefore predicted to tend to arise whenever the balance of (1) is weakened in favour of (2), whether by experimental manipulation, trait-related factors, cultural causes or evolutionary history. Existing evidence is consistent with the model and further implications are discussed.

1. Aims and overview

Delusions are a well-known characteristic of many psychopathologies. Although there are many variations in definition, in popular parlance delusions are beliefs with the properties of being both (1) *persistent* and (2) *unusual*, irrational, or otherwise incorrect. This paper aims to explain how beliefs with both of these characteristics can be acquired; how delusions may arise in an initially more or less delusion-free mind, and how they gradually evolve to become entrenched.

The basic model consists of two main cognitive processes. Firstly, the persistence aspect of delusion is accounted for here by the psychological effects of randomly occurring events. In this aspect the thesis draws on principles from Amsel's theory of Learned Persistence (Amsel, 1992a; pp 72–75; Amsel, 1992b, 1994) and related literature on inconsistent reward and punishment learning schedules. Here, as elsewhere (Amsel, 1992a; p. 54), "persistence" is defined as *whenever some association continues to guide behaviour or thought in the face of non-reinforcement, punishment, obstacles, or other sources of interfering feedback*. Interestingly, a noisy environment, with various recurring events, can superficially resemble the very sorts of learning environments that are optimal for training persistence. Normally a second set of learning processes biases the animal's learning so only the most valid relationships between stimuli are learned (Lubow, 1989). These other learning processes govern the allocation of attention among stimuli, whereby one learns to ignore irrelevant stimuli (Gray and Snowden, 2005). The present thesis is predicated on the possibility of this attentional mechanism failing to constrain the effects of randomly varying reinforcement—that is, failing to ensure that only stimuli that have had a track record of being relevant can become more persistent. It is suggested here that this occurs when there is an imbalance in the strength of the processes that reduce attention toward irrelevant stimuli *relative* to the strength of processes that insulate learning from corrective feedback.

Both irrelevance- and persistence-learning are described in more detail in the sections that follow. Then, in introducing the model, we consider the likely effects of persistence-training processes when they are less constrained by irrelevance learning. The dynamics between these learning processes are then further considered to suggest how anomalies in attention and persistence can be mutually amplifying. In this way, a mind with initially few false beliefs may come to develop snowballing misattributions of causation. How aspects of delusions may be made better understood in light of the model, and how its predictions can be tested, are then discussed.

2. Premises

Associative conditioning is an important factor in how humans learn causal structures in the environment (Dickinson, 1980). It follows from this assumption that if these processes become systematically distorted they can mislead one into believing there are causal relationships between stimuli that do not in fact exist. Beliefs and actions based on inappropriate associations can then give rise to pathological behaviour

E-mail address: droy8625@uni.sydney.edu.au.

http://dx.doi.org/10.1016/j.neubiorev.2017.05.024
Received 7 December 2016; Received in revised form 14 May 2017; Accepted 29 May 2017
Available online 07 June 2017

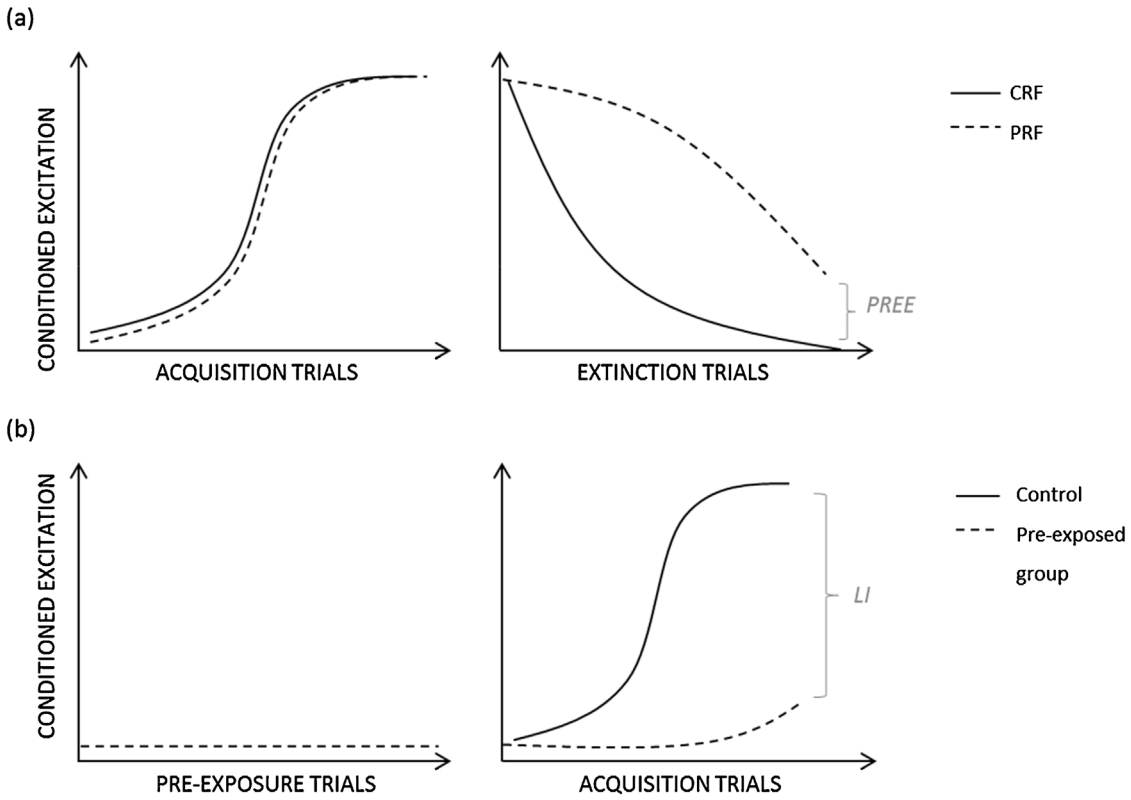

Fig. 1. Partial Reinforcement Extinction and Latent Inhibition Effects. Idealized curves depicting (a) acquisition and extinction of partially-reinforced (PRF) versus continuously-reinforced (CRF) animals (after Fig. 9 .13 in Bouton, 2007); and (b) pre-exposure and acquisition for subjects in the latent inhibition (pre-exposure) group versus control animals.

(Eysenck and Rachman, 1965). The view taken here is that if such false associations support beliefs that become invulnerable to conflicting evidence, then they have developed into "delusions". The aim of this article is to account for how such associations may arise and become entrenched by appealing to aberrations in certain basic learning mechanisms.

2.1. Defining delusion

A major characteristic of delusions and other psychopathologies is their *persistence*. This cannot be emphasized enough here for it is the persistence aspect of delusions that is in most need of explanation. Many discussions purporting to explain "delusion" merely suggest ways in which false or unusual beliefs are acquired, neglecting to demonstrate why false beliefs should persist.

For instance, Howes and Kapur (2009) suggest psychotic delusions are beliefs based on aberrant experiences, such as hallucinations. This style of explanation holds that, because individuals are confronted with such odd perceptions, they develop explanations to make sense of them. When a concocted explanation appears to make sense, it produces "insight relief" that reinforces the belief. The individual may then seek out further confirmatory evidence in line with the delusion. Cognitive biases also feature prominently in many other theories of delusion (for a review, see Bell et al., 2006), such as being prone to make errors in attributing events and experiences to intentions (Frith, 1992), and a tendency to jump to conclusions (Blackwood et al., 2001, pp. 529–530).

As important as such explanations might be for how certain kinds of ideas might occur to an individual, these approaches cannot sufficiently account for delusion more generally. Nor can any explanation that does not consider why the normal processes by which false beliefs are weeded out fail to operate in the particular cases where delusions develop. It is all too easy to explain how one might come to hold a false opinion: one can make missteps in chains of logical reasoning, fall prey to appealing fallacies, become deceived by a conspiracy, be fooled by a magician, hallucinate under conditions of sensory-deprivation, and countless other ways; and powerful reinforcements may indeed play an important role in explaining delusion. But what is also required in a theory of delusion is that it should account for why beliefs continue to be held in the face of costly feedback or in spite of disagreeing with other beliefs one might hold (Corlett et al., 2009; Dennett, 1987, pp.13–35; Bell et al., 2006). For instance, a person might feel a powerfully reinforcing "relief" from anxiety by vividly imagining that they are Superman. But so long as there is gravity this notion is unlikely to have an enduring and pervasive influence on their behaviour. It is an inadequate explanation to note that sincerely believing one is Superman may give some psychic benefits, if for no other reason than it ignores associated costs. In short, we need to explain how these beliefs become persistent in the face of contrary evidence, especially where the costs of believing them far surpass any likely benefits.

Thus "delusion" is defined here both by its unusualness *and* its persistence, as distinguished from simply false beliefs or opinions, which are termed hereinafter mere "illusions". Note this definition is neutral as to the content of delusion. Stricter definitions may be given for certain categories of delusion in diagnostic manuals and elsewhere. Here we are concerned with how delusions *in general* are acquired, as distinguished from the "delusions" clinically defined. It is hoped that an explanation of these more narrowly defined types of delusion will be subsumed by our more general account.

2.2. Inconsistent reinforcement and persistence learning

The paradigmatic demonstration of how persistence can be acquired is the Partial Reinforcement Extinction Effect (PREE). This refers to the observed differences in persistence between animals trained with only inconsistent or "partial" reinforcement compared to those reinforced on every trial during training (that is, by "continuous" reinforcement). As depicted in Fig. 1a, when such partially-reinforced animals are exposed to extinction — that is, when the response is no longer reinforced at all