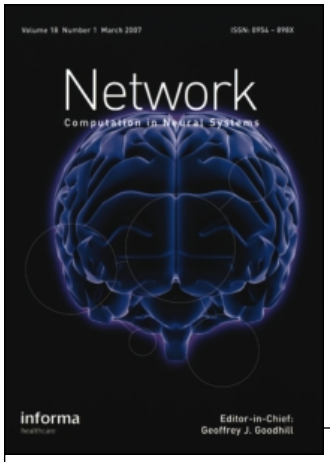

Network: Computation in Neural Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713663148

A temporal difference account of avoidance learning

Michael Moutoussis [a]; Richard P. Bentall [b]; Jonathan Williams [c]; Peter Dayan [d]
[a] Tolworth Hospital, Surbiton, England
[b] School of Psychology, University of Wales, Bangor, Gwynedd, USA
[c] Department of Child & Adolescent Psychiatry, Institute of Psychiatry, London, England
[d] Gatsby Computational Neuroscience Unit, London, England

Online Publication Date: 01 January 2008

To cite this Article: Moutoussis, Michael, Bentall, Richard P., Williams, Jonathan and Dayan, Peter (2008) 'A temporal difference account of avoidance learning', Network: Computation in Neural Systems, 19:2, 137 — 160

To link to this article: DOI: 10.1080/09548980802192784
URL: http://dx.doi.org/10.1080/09548980802192784

Network: Computation in Neural Systems
June 2008; 19(2): 137–160

informa
healthcare

A temporal difference account of avoidance learning

MICHAEL MOUTOUSSIS[1], RICHARD P. BENTALL[2], JONATHAN WILLIAMS[3], & PETER DAYAN[4]

[1]*Tolworth Hospital, Surbiton, England,* [2]*School of Psychology, University of Wales, Bangor, Gwynedd, USA,* [3]*Department of Child & Adolescent Psychiatry, Institute of Psychiatry, London, England, and* [4]*Gatsby Computational Neuroscience Unit, London, England*

(Received 21 January 2008; accepted 10 May 2008)

Abstract
Aversive processing plays a central role in human phobic fears and may also be important in some symptoms of psychosis. We developed a temporal-difference model of the conditioned avoidance response, an important experimental model for aversive learning which is also a central pharmacological model of psychosis. In the model, dopamine neurons reported outcomes that were better than the learner expected, typically coming from reaching safety states, and thus controlled the acquisition of a suitable policy. The model accounts for normal conditioned avoidance learning, the persistence of responding in extinction, and critical effects of dopamine blockade, notably that subjects experiencing shocks under dopamine blockade, and hence failing to avoid them, nevertheless develop avoidance responses when both shocks and dopamine blockade are subsequently removed. These postulated roles of dopamine in aversive learning can thus account for many of the effects of dopaminergic modulation seen in laboratory models of psychopathological processes.

Keywords: *Conditioned avoidance response, dopamine, psychosis, blockade*

Introduction

There is a wide disparity between the sophistication of our understanding of learning and choice in appetitive vs. aversive contexts. Appetitive learning has attracted theories of the acquisition and expression of habits based on temporal

Correspondence: M. Moutoussis, Tolworth Hospital, Red Lion Rd., Surbiton KT6 7Q U, England. Tel: +44 (0)2082961367. Fax: +44 (0)2083903877. E-mail: michael.moutoussis@swlstg-tr.nhs.uk

ISSN 0954-898X print/ISSN 1361-6536 online/08/020137–160
DOI: 10.1080/09548980802192784

difference (TD) learning, which somewhat seamlessly link statistical, psychological and neural ideas and data (Schultz et al. 1993; Montague et al. 1996). There is also a developing understanding at all these levels of the relationship between habits and motivationally sophisticated goal-directed actions. By contrast, despite some notable studies (Grossberg 1972; Schmajuk and Zanutto 1997; Daw et al. 2002; Johnson et al. 2002; Seymour et al. 2004), the functional bases of aversive learning remain more obscure.

We consider a paradigmatic case of aversive learning, namely the conditioned avoidance response (CAR). The CAR has particular significance for psychiatry, having inspired the development of behavioural therapy techniques (e.g. *response prevention*) by psychologists. It is also a standard test-bed for assessing *antipsychotic drugs* by psychopharmacologists (e.g. Anisman 1978; Bardin et al. 2007; Siuciak et al. 2007) It has therefore been very important to understand the CAR in detail. Despite this, efforts towards such understanding waned as it was realized that classical, operant, cognitive-expectancy and possibly other brain mechanisms were all involved. In more recent years, interest in the CAR resumed as both theoretical (e.g. Smith et al. 2005, 2006, 2007) and psychopharmacological (e.g. Wadenberg and Hicks 1999; Samaha et al. 2007) progress was made. Further, although the role of specifically aversive learning mechanisms in psychopathology is yet to be completely understood, the CAR is likely to involve key psychobiological mechanisms that may be activated to an exaggerated degree in clinically paranoid anticipation of threats (Moutoussis et al. 2007).

In this article, we explore three hypotheses. First, we suggest that a temporal-difference learning model of the CAR can capture its key qualitative, empirical psychological findings. Second, we show that the pharmaco-behavioural findings imply that dopamine is unlikely to be involved in the learning of aversive associations to stimuli, a conclusion that is in contrast to the main theories so far suggested for its role in the CAR, including those based on TD ideas (notably Smith et al. 2005, 2006, 2007). Third, we show that our model helps transcend the difficulties of the classical, qualitative psychological accounts of aversive learning even without appealing to special, additional features of the mechanisms involved. At the end, we draw out the predictions of our analysis relevant to the study of human appraisal of threat, including paranoia.

Conditioned avoidance responses

In a typical rodent version of the CAR, a subject learns that a neutral warning conditioned stimulus (CS) will be followed by an unconditioned aversive stimulus (US) – usually an electric shock (Figure 1). After the onset of the CS, the subject can avoid the US altogether by performing an experimenter-determined skeletal response within a specific CS-onset - to - US interval. This is termed the avoidance response (AR), and, for rats, usually consists of shuttling to a different part of the experimental enclosure. Generally, the AR interrupts the CS and aborts the US. Shuttling after US onset interrupts exposure to the US and is termed an escape response (ER). In the human version of the CAR, the US is often a burst of loud white noise, and the AR/ER generally involves pulling a lever rather than shuttling (Unger et al. 2003).

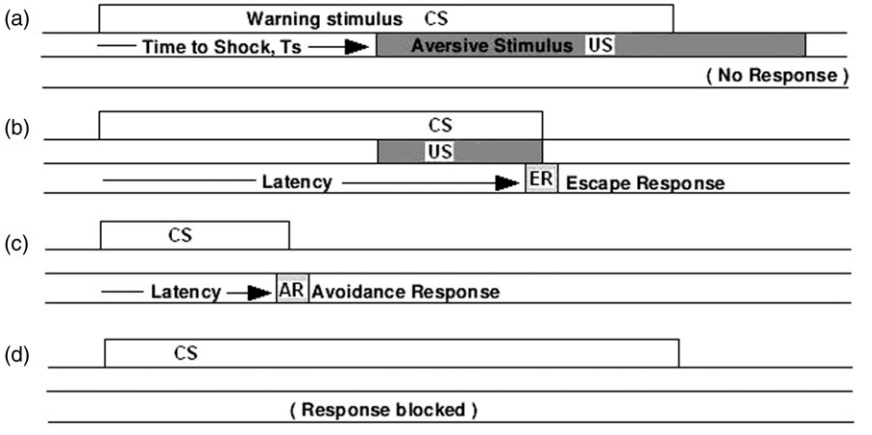

Figure 1. The onset of the Warning Stimulus signifies the start of a conditioned avoidance trial. The aversive (Shock) stimulus follows after a standard interval *Ts* (a). Performing a specific safety behaviour after US onset (ER) stops both CS and US (b). If the safety behaviour is performed with a latency $T1 < Ts$, (AR) the CS is aborted and no US is delivered (c). After acquisition, response may be blocked while no US is given (response prevention, (d)).

Under normal circumstances, animals take only a few shocks to discover that shuttling interrupts the shock (ER). From then on they quickly learn to perform the AR. Subjects achieve a high percentage of ARs in a few tens of trials [cf. Figure 2(a) (Beninger et al. 1980a)]. As learning proceeds, their latencies of responding decrease [cf. Figure 2(b) (Solomon and Wynne 1953)], and they lose any sign of overt fear to the CS. Once the AR is well-learnt, providing the shocks have been of sufficient magnitude, responding continues for many trials even if no shocks are, or would be, delivered (Figure 2b and c) (Solomon and Wynne 1953; McAllister et al. 1986). Extinction can be accelerated by physically preventing the animal from shuttling (response prevention Figure 1d). This initially leads to an increase in signs of anxiety, which successful avoidance-responding had eliminated. If, however, after a few trials the animal is allowed to shuttle, the frequency of AR is much reduced – even in the presence of residual signs of anxiety (Mineka 1979).

A variant of the CAR designed to separate a phase of Pavlovian-like aversive learning from a phase of instrumental-like learning is the escape-from-fear (EFF) paradigm (McAllister et al. 1980). Here again a warning stimulus is followed by a shock for a set number of training trials and, during these trials, the animal has no way of terminating the shocks. In the immediately subsequent trials, however, shuttling becomes available as a response. In most experiments, shocks also stop being delivered following the warning stimulus. Animals are observed to acquire a shuttling response quickly, but not immediately; its latency then decreases. Shuttling may again persist for dozens of trials before gradually extinguishing (Figure 2c).

Critically, the CAR is suppressed by blocking Dopamine receptor type 2 (D2) function with antipsychotic drugs (Figure 2a). Although at high doses, motor output itself is compromised, at lower doses, the escape response is unaffected, suggesting that antipsychotics affect acquisition during training (Smith et al. 2007).

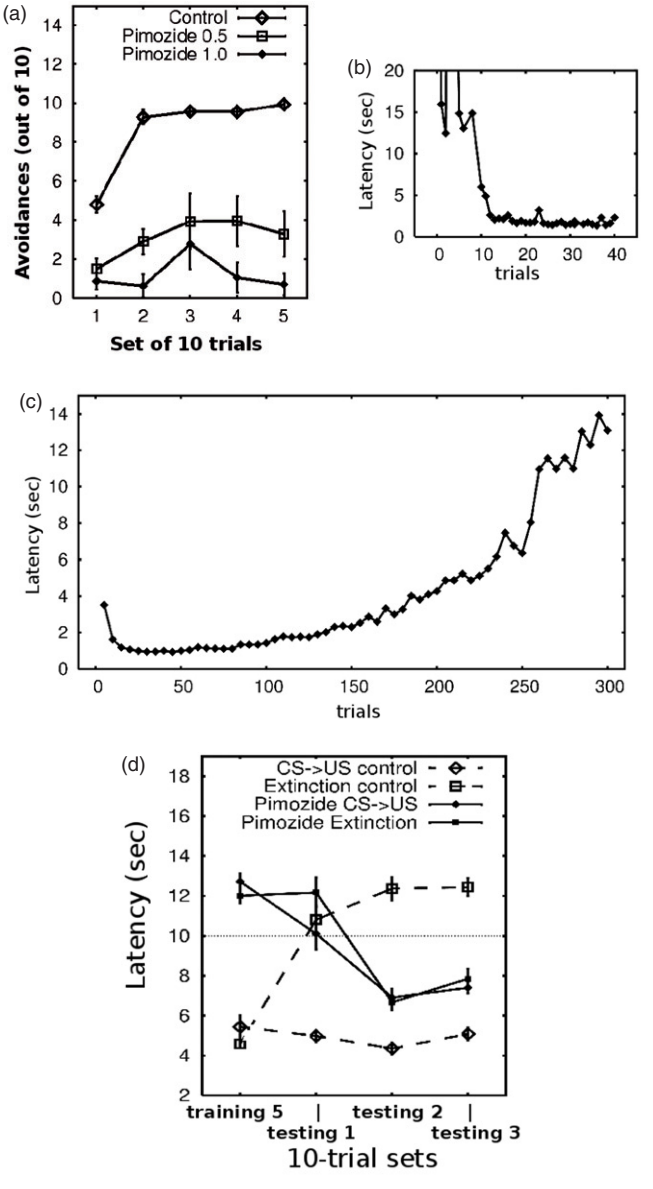

Figure 2. (a) Pooled avoidance probability data from rats. In normal CAR learning, near-perfect avoidance is achieved within a few trials (from (Beninger et al. 1980b)). DA block shows a powerful dose-dependent effect (Pimozide doses in mg/kg). (b) Sample of data from a dog (Solomon and Wynne 1953) where response latencies decrease for many trials after achievement of 100% avoidance. (c) Pooled latency data from animals subject to Escape-From-Fear training. Following the CS they were first given inescapable shocks; then they were given the opportunity to shuttle with the shocks turned off. Performance increases (i.e. latency decreases) for about 20 trials, remains stable for another 50 trials, then declines slowly first, then faster. From (McAllister and McAllister 1991). (d) Latency data from rats (similar to a.). Unfilled diamonds: unmedicated rats performing the standard CAR in training and testing. Unfilled squares: shocks turned off (normal extinction) during testing. 10 s is indicated, being the delay with which the US was given in training trials. Filled symbols are data from rats that received Pimozide during training only. From Beninger et al. (1980b).

Equally, DA blockers reduce expression of the AR, in a way that resembles the extinction-like effect that administering such drugs has on reward-motivated behaviour (Wise et al. 1978).

One might think that dopamine is involved in pathways reporting aversive events in a way analogous to its role in reporting better-than-expected outcomes in reward learning (Schultz et al. 1993). Indeed, microdialysis and other studies showed that dopamine is released in response to aversive stimuli (Horvitz 2000). In addition, imaging studies showed activation in response to aversive stimuli in areas innervated by the monoaminergic systems (Jensen et al. 2003; Menon et al. 2007).

There are, however, reasons to believe that dopamine is not directly involved in reporting negative outcomes. First, the ventral-tegmental neurons that are excited (at least under some conditions) by aversive stimuli are likely not dopaminergic, as originally thought, (Ungless et al. 2004). Second, dopaminergic neurons that do predict outcomes only code better-than-expected ones (Bayer and Glimcher 2005) with any substantial fast fidelity. Third, in a recent human imaging study (Menon et al. 2007) it was noted that dopaminergic enhancement or blockade did not affect the subjectively reported anxiety experienced in response to a conditioned stimulus predicting pain; neither did dopaminergic manipulations affect the subjects' ability to learn which conditioned stimulus predicted the painful one.

However, perhaps the most significant challenge to existing dopaminergic (Smith et al. 2005; Smith et al. 2006, 2007) and indeed some non-dopaminergic (Schmajuk and Zanutto 1997) models of the CAR is from a condition that turns out to resemble escape from fear. In this, the animals are trained with a blockade of dopamine D2 receptors, and thus fail to acquire the AR. However, if the dopamine blockade *and the shocks* are removed after the subjects have experienced a few shocks, subjects actually acquire the AR (Beninger et al. 1980b), following a learning curve that resembles that obtained in the EFF paradigm (Figure 2c). The obvious interpretation of this is that dopamine is more likely to be involved in learning to respond based on the Pavlovian CS-US association, rather than in the formation of the association itself.

In the present study, we built a temporal-difference learning (Sutton and Barto 1998) model of the CAR and investigated the consequences of manipulating dopamine in the model. The model successfully reproduced a broad range of CAR phenomena. We argue that this bolsters the hypothesis that DA is involved in boosting predictions and actions when outcomes are better than expected. Other systems, putatively (though controversially) serotonin (Daw et al. 2002) could play a similar role for aversive prediction learning when outcomes are worse than expected, but appear not to be able to affect action learning by itself, at least in contexts like CAR. The model does not seek to account for slow (tonic-like) timescale dopamine effects, which are important for understanding some results of pharmacological manipulations and some microdialysis findings. We shall refer to these separately.

The model

In the TD variant of reinforcement learning (RL), subjects come to expect particular gains or losses (collectively value) to accrue from each situation or state they encounter. The *change* in these expectations should stochastically match the

immediate gains and losses they experience; if it does not, then there is a prediction error (PE) that can be used to improve the estimates of the returns. Based on the substantial evidence about the phasic activity of dopamine cells (Schultz et al. 1993; Bayer and Glimcher 2005), we modelled the appetitive portion of the PE as being dopaminergic. In the CAR context, appetitive PEs arise when the subject performs the avoidance response, and so changes from being in a state of fear, anticipating the shock, to being in a state of safety, when the shock has been averted. It is this transition that is reported by phasic dopamine. On the other hand, given our TD architecture, the post-dopamine-blockade avoidance acquisition data (Beninger et al. 1980b) suggest that aversive value-learning can proceed even in the presence of dopamine blockade. This conclusion is also consistent with theoretical suggestions about appetitive-aversive opponency (Solomon and Corbit 1974; Daw et al. 2002).

Values are only one part of RL; and are normally acquired in the service of learning policies, which are a systematic (though possibly stochastic) ways of assigning actions to states. In several variants of TD, the aversive values and PEs can be directly used to learn actions that minimize the values. This can be seen as a form of Mowrer's two-factor theory (Mowrer 1947), with the conditioned fear (or anxiety (Gray and McNaughton 1996)) arising from the value predictions; and with the (dopaminergically reported) reduction in conditioned fear acting like an appetitive reinforcer, boosting the subsequent selection of the associated action. There are several ways that policies may be represented, notably indirect methods, in which they are derived from predictions of long run values, and direct methods, in which they have their own parameters. For appetitive learning, about which rather more is known in this respect, there may even be functional and indeed structural transitions between different forms of policy over the course of learning (Belin and Everitt 2008).

More precisely, the class of models that the present work belongs to is termed actor–critic (Sutton and Barto 1998) models. They have two key components:

- A critic which learns affective expectations. This is the part of the model which associates with each distinct state that the animal perceives the affective value which summarizes how good or bad this state is (a measure of the total return to be expected to follow this state).
- An actor which learns to make appropriate decisions in the light of these expectations. The (usually probabilistic) rules of taking these decisions is what the actor learns and is termed the 'behavioural policy'. In our case, the optimal policy is the one that minimizes long-term costs.

Long-term costs, and thus values themselves, will therefore depend in turn on the actions that the animal takes (e.g., whether it avoids the shock or not). Note that the output of the critic embodies the expectations and predictions that appeared so puzzling to some in the purely behaviourist era (Lovibond 2006).

Advantage learning (Dayan and Balleine 2002) is a form of the actor–critic in which action choice depends on a particular aspect of the value of an action. The advantage $m(a, s)$ of action a in state s quantifies how much better this action is compared to the policy followed on average, i.e. it is defined as the difference between the value of the particular action $Q(a, s)$ and the value of the state $V(s)$. A major spur to our use of it is that O'Doherty and co-workers

(O'Doherty et al. 2004) showed that advantage learning provided a good model for the BOLD signal in the dorsal striatum during the acquisition of a simple (appetitive) instrumental task.

The experimental data strongly constrained the architecture of our model. First, a model that used only one set of values (i.e. action values) to learn could be discounted in favour of one that had both state-values and action-related (e.g. advantage) values. This is because the action-value models under dopamine blockade or EFF would only have these action-values to remember, and hence would have no basis for preferring the avoidance response once that becomes available (as per Figure 2). Once we adopted an actor–critic architecture, we were forced to interpret the PE differently in the critic (value) vs. actor (policy) limbs. Worse-than-expected PEs could not depend on dopamine in the critic part, as aversive value learning survives dopamine blockade as discussed above. However, in the actor limb learning, signals should depend on dopamine, as avoidance action learning does not survive dopamine blockade (Beninger et al. 1980b). An additional experimental constraint determined our choice of the advantage-learning variant of the actor–critic method. This is that under dopaminergic blockade, the asymptotic frequency of avoidance responding appears to change in a quantitative, dose-dependent manner (Smith et al. 2007). In other common variants of the actor–critic formalism (e.g. Sutton and Barto 1998, Chapter 6) the rate of policy learning would change, but not the asymptotic policy preference.

We now provide an algebraic description of our advantage-learning model.

Given a particular environment, the animal encounters a sequence of states: $s(first) \ldots s(k),\ s(k+1), \ldots, s(last)$. The value of a state, $V[s(k)]$ reflects the expected return for the whole sequence following $s(k)$. In TD learning, rewards anticipated in a particular state of the animal are compared to rewards actually received during the immediately subsequent state. The simplest formulation of this difference, the PE, is

$$\delta_V(k) = (R(k+1) + V[s(k+1)]) - V[s(k)] \tag{1}$$

where $R(k+1)$ is the return (reward or cost) experienced in going from $s(k)$ to $s(k+1)$. The difference between expectation and return, δV, can be used by the learner to improve its estimate of what value to attach to the original state:

$$V[s(k)]_{\text{new}} = V[s(k)]_{\text{old}} + \alpha * \delta_V(k) \tag{2}$$

where α is a learning rate parameter. For trials where dopamine modulation was simulated, the term δV in Equation 2 was first scaled by an additional factor *mDA* – but only if Equation 1 showed a better-than-expected outcome.

In tasks such as CAR, subjects need to *associate actions with returns*. If the subject took a particular action a when it was in $s(k)$ (and ended up in $s(k+1)$), then a positive δV means that the action led to a more rewarding outcome than expected. It would be beneficial for this action to be associated with this state, i.e. taken with a higher probability $P[a \mid s(k)]$ in this state.

$$P[a|s]_{\text{new}} = f(P[a|s]_{\text{old}},\ \delta_v(k)) \tag{3a}$$

where f is an increasing function of δV(k). We used an important example of a parametric expression for $P[a \mid s(k)]$, the Gibbs function:

$$P[a_i|s] = \frac{e^{m(a_i,s)/T}}{\sum e^{m(a_k,s)/T}} \tag{3b}$$

where T is a parameter that determines how sharply a difference between policy parameters $m(a, s)$ is translated into probabilities to select these actions. In the advantage-learning variant of TD, the learner uses Equations 1, 2 and 3, and estimates policy parameters as follows:

$$m(a', s)_{\text{new}} = m(a', s)_{\text{old}} + \varepsilon * (\delta_V - m(a', s)) \tag{4}$$

where ε is a policy-learning rate. It can be shown that this is equivalent to the learner coming to basing its policy $m(a, s)$ on the advantage estimated for action a in state s relative to the outcome expected from state s overall. To simulate dopamine modulation, the term δ_V in Equation 4 was first scaled by the factor *mDA* above – this time irrespective of whether the outcome was better-than or worse-than-expected. In order to keep to the convention used in this article, where greater values are more aversive but greater policy parameters denote greater tendency to action, the sign of δ_V also has to be reversed before its use in Equation 4. Note that for a suboptimal policy the action *amax* that leads to the optimal reward from state s will not always be chosen. Therefore the $\delta_V(k)$ calculated in the trials where *amax* is chosen will not be zero according to Equation 1. In contrast, Equation 4 will converge for non-optimal policies to $m(a,s)$ equalling the average (non-zero) δ_V for the particular action a. This is important for the model to be consistent with the tendency of animals to choose actions with probabilities depending on their return, rather than simply maximizing by choosing simply the action associated with highest return.

Parameters and implementation

The basic structure used for the simulations is shown in Figure 3. Six identical, perfectly generalizing 'stay' actions and one 'shuttle' action were available from each state. 'Stay' actions had cost zero, while the cost of shuttling was 0.2. The cost of the

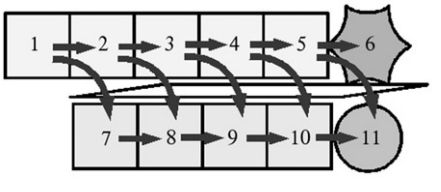

Figure 3. Schematized state and action diagram for the standard CAR. The CS starts at state labelled 1. Action 'stay' results in horizontal arrows and has no intrinsic cost. Action 'shuttle' moves from the unsafe side (top row) to the safe side of the apparatus (bottom row) and incurs a motoric cost of jumping the barrier. There are five time steps between CS onset and the end of each trial. The transition $5 \rightarrow 6$ results in shock. Once in state 6 or 11 the trial ends. States 7–11 could have been condensed to one safety state – they are shown separately just for added clarity of timing.

shock was set at or above 4.0. This 20-fold or greater ratio was meant to simulate the traumatic nature of the traditional animal CAR (and, possibly, the dire anticipations of persecutory ideation).

In our advantage-learning models, exploration and asymptotic behaviour depend on a 'brittleness' (to use the term adopted by Williams & Dayan 2005; or 'inverse temperature') parameter $1/T$. The more brittle the model, the more a difference in costs (or rewards) between actions translates to a difference in probability of their selection. $1/T$ can therefore be thought of as an index of how important for action selection the typical costs and benefits involved in the experiment are. Decreasing $1/T$ leads to more exploration of alternatives before the final policy probabilities are reached. This affects, in turn, whether the learner has a propensity to modify the policy followed once a way of avoiding shock (by shuttling from a particular state) has been discovered. $1/T$ should be large enough to allow exploration around the cost of the non-traumatic 'shuttle' action but not to dampen the effect of the high cost of shock. In the simulations shown here the temperature/brittleness parameter T was 0.2.

The learning rate for state values was 0.5, while for policies it was 0.075. As is typical in actor–critic schemes, the learning rate of the critic must be substantially greater than that of the actor for performance to be appropriate. For Figure 6, *mDA* was set at 0.2 (see description of Equation 2 and 4 above), while for Figure 7 *mDA* was 0.15 but only during the 'shocked' trials (trials 101–130). Each run of trials included 100 unshocked trials initially, so that by the time shocks started the 'stay' vs. 'shuttle' policies were determined by the relative costs rather than any initial values.

Results

Simulation of escape-from-fear learning

The inescapable-shock phase of the EFF paradigm is conceptually simpler than the escapable-shock CAR. This phase also serves as a useful comparison for the shocked phase of the CAR that takes place under dopamine block. Figure 4 shows a simulation of EFF. Before the onset of shocks, the subject explores the available actions and occasionally shuttles. The values of states 1 to 5 stay around zero, as the only slightly aversive outcome is the small motoric cost of the occasional shuttling. When shocks start the 'safe' states (S7–S11) become unavailable. The states temporally near to the shock (e.g. S5, diamonds in Figure 4a) first acquire aversive values. These they gradually feed back to earlier predictive states (cf. S3, triangles, and S1, crosses, in Figure 4a).

Shuttling is then allowed. Most learners soon try shuttling again, and hence experience a large positive PE – from high value (S1–S5) to zero (S7–S11). This gradually reduces the values of S1 to S5, but at the same time teaches that shuttling is quite advantageous. The probability of avoidance rises rapidly and persists at an elevated level for tens of trials (Figure 4b; cf. Figure 2c). Features of this latter, unshocked phase of the EFF, which resembles closely the corresponding phase of the standard CAR, will be presented in the context of the latter. The underlying mechanics of learning are presented in the Appendix – Figures A1 and A2).

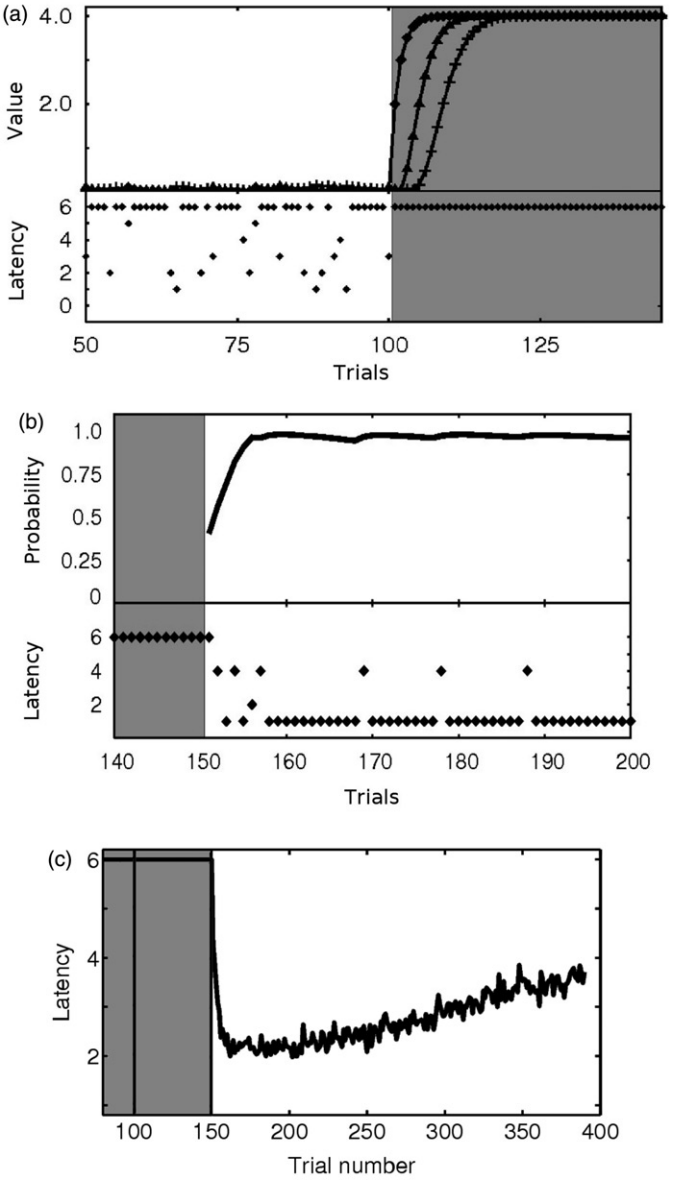

Figure 4. (a) First part of Escape-From-Fear trial for one subject. In this and subsequent figures, shocks are turned on at trial 101. Lower plot: response latency. Each black diamond corresponds to the last state of the 'unsafe' side visited during the trial (e.g. a value of 1 means shuttling from state 1 while 6 means no shuttling during that trial). Top panel – Values of states 1 (crosses), 3 (triangles) and 5 (diamonds), showing how they converge in turn to the cost-of-shock when the latter is inescapable. (b) Second part of EFF. Bottom panel – same as a. Top – overall probability of avoidance calculated from model variables (based on Equation 4; See Appendix for further details). Avoidance probability increases quickly; then decays slowly, but is boosted when later states are visited. (c) Pooled Latency for 100 simulated identical subjects. Note qualitative similarity to Figure 2(c) – and also to 2(d) 'Pimozide Extinction'.

Simulation of normal CAR learning

Figure 5 shows a typical example of a learning curve from our model. First, as is true of control subjects, after receiving a small number of shocks, the model learns to favour the avoidance action, with a probability approaching one. If adequate exploration occurs, responses tend to move to earlier states quickly, reducing response latency. The steep learning curve is followed by consistent avoidance, which is quite persistent (though not wholly immune from being extinguished) even after programmed shocks have stopped. This is similar to the animal data shown in Figure 2(b) and (c) (Solomon and Wynne 1953; McAllister et al. 1986).

If the learner is forced to follow action 'stay' rather than 'shuttle' during extinction, then the model learns not-to-avoid more rapidly. It is thus sensitive to response-prevention. At the onset of this phase, the values of early states show a transient increase (Figure 5c, first 5-10 trials of RP). If we assume that visiting a high-value state is psychologically alarming to the animal, we have a situation analogous to the alarm initially experienced by animals or humans subject to response-prevention (this is similar to transient increases during response persistence – cf. Figure A1). At the same time, values of later states decay to zero. With more learning, all values decay to near-zero levels.

Simulation of dopaminergic manipulations in the CAR

DA manipulations can be initiated and removed at various times during learning. Most straightforwardly, if D2 blockade is affected after acquisition, simulations show that persistent responding in the absence of shocks is more likely to spontaneously extinguish earlier (data not shown).

If D2 receptors are blocked during initial learning, then the model is much slower to acquire the avoidance response (Figure 6). This is similar to the animal finding that D2 blockade results in a dramatic decrease in AR learning. In Figure 2(d), for example, rats treated with pimozide during training showed an average latency of responding of 12 s, as opposed to 5 s for the control rats (Beninger et al. 1980b). In our models, peak responding (and shortest latency) takes much longer to achieve, and furthermore, the peak probability of response is also significantly reduced.

Empirically, if DA blockade is removed when shocks are also turned off, avoidance dramatically strengthens in extinction (as in Figure 2(b), 'Pimozide Extinction'). In our model, removing the blockade affects both the probability of response and the latency for responding (Figure 7a and b). Note that ARs are even more persistent, and average latencies even shorter, than in the normal case shown in Figure 4. This is largely due to these learners having been exposed to more shocks during acquisition of the aversive state-value structure, on which the subsequent response acquisition and persistence depends.

Finally, we simulated the effects of boosting rather than suppressing dopamine. This increases the rate of learning and subsequently the persistence of behaviour in the absence of further shocks, but does not result in true 'resistance to extinction', i.e. an impairment of the effect of response prevention. The experimental literature on this topic is limited but suggests that low doses of DA agonists *enhance* the effect of response prevention (Cooper et al. 1974; Christy and Reid 1975), consistent with our models (data not shown).

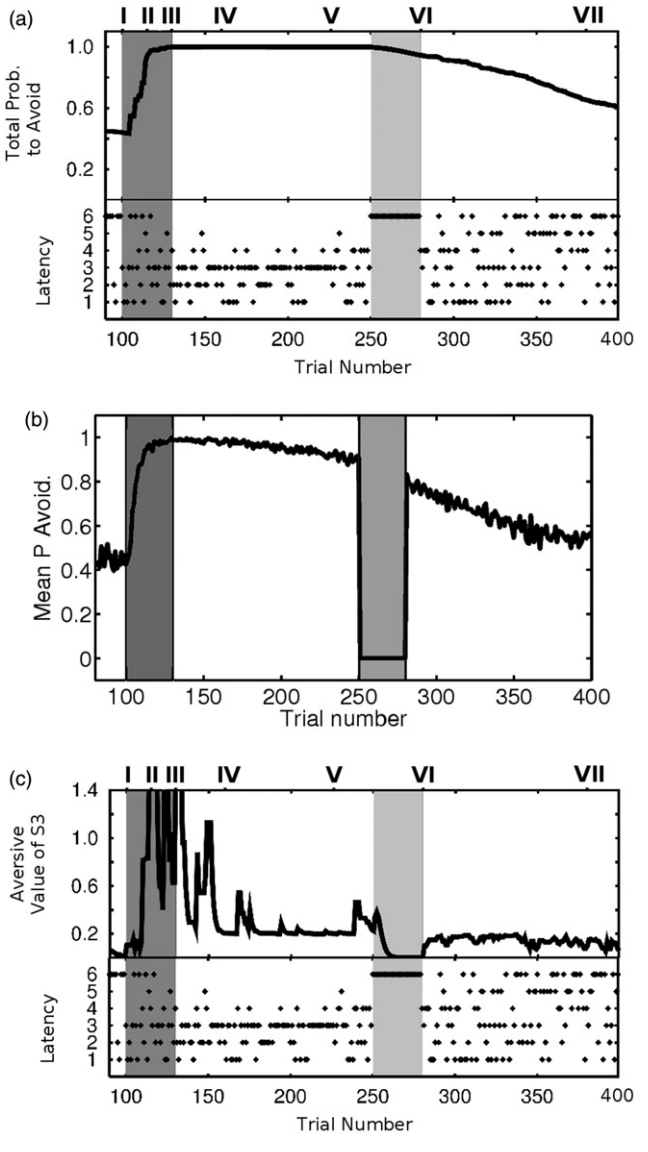

Figure 5. (a) Unmodulated CAR simulation. Shocks are turned on between trials 101 and 150 (dark grey), but they are escapable. Response prevention (forced staying) occurs between trials 250 and 280 (light grey). Top plot – overall probability of avoidance calculated from model variables (based on Equation 4). Probabilities are shown on the left ordinate. Roman numerals mark the trials where latency statistics in Figure 5(b) are calculated. Bottom ('Latency'): as in Figure 4. (b) Avoidance probabilities averaged over 300 'subjects' $\times 380$ trials each, protocol as per Figure 5(a). Before training (I) most learners reach state 6, where shock is delivered. There is relatively little degradation in performance 30 trials after shocks stop (IV), but some degradation is evident 65 trials later (V). Avoidance is reduced faster after response prevention (VI). 100 trials after, responding is much like pre-training (VII). (c) Values of state 3, from which much shuttling takes place, of example of Figure 5(a). Note that during the late parts of shocked-learning, but also the early parts of 'persistence' trials, choosing action 'stay' results in a dramatic increase of the value of state 3. Trials 250–280 show details of response-prevention. Note the transient increase in state 3 value.

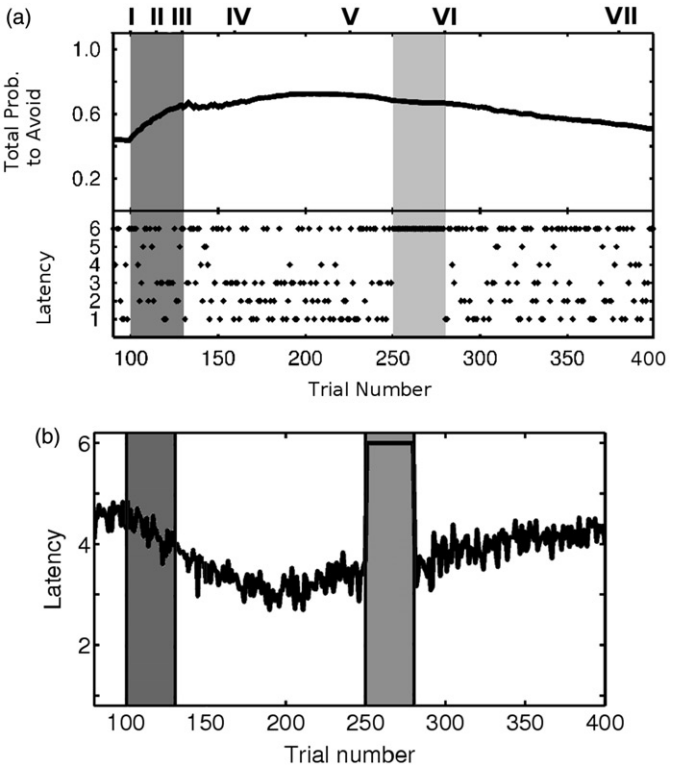

Figure 6. Example of simulation of 80% DA blockade affecting positive corrections to values and also all corrections to policy parameters. Otherwise the trials are conducted as in Figure 5. Grey-scale coding also as per Figure 5. (a) Example of avoidance probability and latency. A much lower level of responding is established (cf. Figure 2a 'Pimozide'). Response prevention has little effect. (b) Latencies averaged over 100 'subjects'. DA blockade has caused much less avoidance of state 6, where shocks were delivered in training. Some reduction in latency took place during the 'persistence' trials (IV–V).

Discussion

The temporal-difference model of avoidance successfully describes a large range of qualitative experimental results. It suggests specific computational roles for the elements of two-factor theory, linking it directly to dopaminergic mechanisms. The model resolves the apparent paradox that dopamine receptor antagonists dramatically suppress avoidance responding, and yet appear not to be involved in reporting worse-than-expected, aversive outcomes. It also sheds light on other findings that appear puzzling or counterintuitive from the point of view of qualitative two-factor theory. These include the relatively persistent, efficient shuttling in the absence of fear, which has been subject to much debate. In our models, this naturally follows from the fact that it is the cumulative experience of differences between outcomes, not the outcomes themselves that result in policy learning.

There have been other models of the CAR. In the category of two-factor models (like ours), we note the work of Grossberg (1972), who has carefully analysed

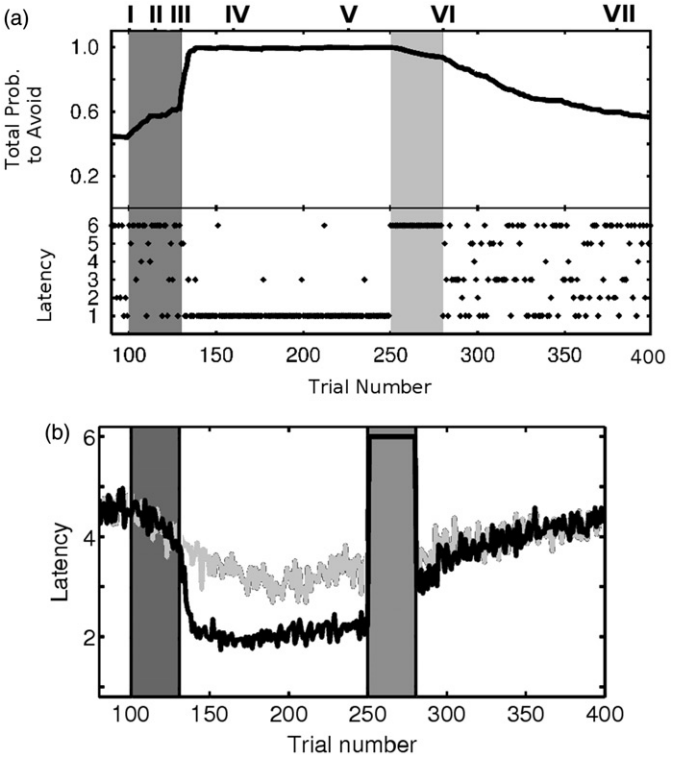

Figure 7. Advantage-learning model that has been DA-blocked during learning is then freed both from DA block and from shocks from trial 131 on. (a) Example of avoidance probability and latency. Spontaneous learning occurs to a high level, with a dramatic increase upon release from DA blockade. (b) Averaged latency plot, showing the dramatic response to withdrawal of DA blockade (black curve). Light grey curve is the same as Figure 6(b), persisting DA blockade, for comparison. Note effect of response prevention.

opponency in nonlinear dynamical settings, and of Schmajuk and Zanutto (1997) and Johnson and co-workers (2002). Despite their many attractive features, these models are more psychological in nature. They pay less attention to the substantial data on the involvement of dopamine in appetitive learning, together with its apparent opponent-based role in the CAR. They therefore do not set out to address the key experimental data that constrains our model. Note that in our model the opponent dopamine signals (due to 'relief' that the aversive stimulus has been avoided) are on the same footing with all others, and that signals due to different psychological sources (CS, AR) sum linearly. It was not necessary to appeal to any special properties of the 'safety signals' [for instance as extinction-resistant conditioned inhibitors of fear (Gray 1987)] that accompany the avoidance response.

A second category of model of avoidance involves versions of Expectancy theories, the prototype of which is that of Seligman and Johnston (1973). Expectancy theories emphasize that, subsequent to the initial associative fear-conditioning, animals explicitly learn which events are likely to follow each action in each state. Subjects can then decide which action to take by comparing the ultimate

expected outcomes of sequences of actions. The learner is said to use a 'forward model' (Wolpert and Miall 1996), or, in a conditioning context, goal-directed actions (Dickinson and Balleine 2001; Daw et al. 2005). Two of Smith and co-workers' influential models (Smith et al. 2005, 2006) are perhaps most correctly seen in expectancy terms. Dopamine plays two roles in these particular models. One, putatively ascribed to tonic levels of this neuromodulator, controls the course of inference through the forward model. The second role, identified with phasic DA, is to report a product of *surprise* and *significance* which is used to learn transition coefficients connecting the states comprising the forward model. These models have trouble accounting for the gradual acquisition seen in Beninger's blockade study (Beninger et al. 1980b), since the expectancy of fear that is clearly established does not immediately lead to appropriate responses even when the blockade is removed. Their contact with existing data on DA's involvement in appetitive learning is also somewhat distant.

In the context of appetitive conditioning it has been suggested (following Coutureau and Killcross 2003) that model free, cached, RL methods would coexist with model-based, goal-directed, RL (Daw et al. 2005). Expectancy theories involve just the sort of forward models that are implicated in the goal-directed system; while our two-factor model is a form of cached controller. Thus it would be natural to suppose that both sorts of model may coexist in this aversive case too; a possibility that opens up various lines of experimental enquiry. Daw and co-workers (2005) suggested that arbitration between the two different controllers should depend on their relative uncertainties; a suggestion that would also be suitable for this aversive case.

Rather more *sui generis* is the third model of CAR suggested by Smith and colleagues (2007). In this work, affective values translate directly to probabilities of action and hence it is again difficult to account for the development of responding in extinction (Beninger et al. 1980a). Compared with this, we placed particular emphasis on the aversive framing of the CAR, and the apparent opponent interaction between dopamine and its putative aversive opponent.

Neurobiological substrate

A rich nexus of neural areas appears to be involved in the habitual system's instantiation of the CAR (noting that different paradigms of aversive learning recruit different such groupings, e.g. Reis et al. 2004). First, the basolateral amygdala (BLA) appears necessary for the acquisition of instrumental avoidance (Poremba and Gabriel 1995) whereas different amygdalar regions appear to be involved in other types of fear learning, such as conditioned suppression (Killcross et al. 1997). The human amygdala may be involved in the evaluation of unexpected costs (Yacubian et al. 2006) although some studies did not detect such activity (Seymour et al. 2004). The amygdala appears only to be involved in the initial stages of aversive learning and even the passage of time alone (not learning trials) reduces its involvement (Poremba and Gabriel 1999). This may contribute to the heterogeneity of imaging findings (Jensen et al. 2003). Basolateral amygdala projection neurons receive glutamatergic projections from sensory association cortices. They also receive inhibitory input from prefrontal cortex, via inhibitory interneurons. It is here that mesolimbic dopaminergic

inputs intervene. Both D1 and D2 receptors on BLA neurons serve to suppress medial prefrontal inhibition and enhance responses to sensory inputs (Rosenkranz and Grace 2001, 2002).

The orbitofrontal cortex also plays an important role in avoidance, at least in human subjects. There is evidence that this region encodes not-incurring the 'cost' or 'loss' associated with an aversive stimulus during avoidance tasks as if it were a reward value (Kim et al. 2006). Such a 'reward' value could serve to reinforce safety behaviours.

Imaging studies also implicate two other areas in encoding the discrepancy between how aversive a person estimates a situation to be, and how aversive this situation actually turns out to be (what we termed the PE). The areas involved are the insular cortex and the corpus striatum (Seymour et al. 2004, 2007). The corpus striatum appears to contain different functional subregions, which are preferentially activated by rewarding PE (near the nucleus accumbens) and aversive PE (slightly more posterior, in the putamen).

Finally, the most critical question for our model is the neural substrate for the representation of the aversive PE to control the learning of the aversive values. Based on various lines of evidence, it has been suggested that serotonin may play a critical role, perhaps as an opponent (Grossberg 1972; Solomon and Corbit 1974) to dopamine (Deakin and Graeff 1991; Daw et al. 2002). Indeed, serotonin has been shown to play an important role in learning in the CAR (Titov et al. 1983; Ma and Yu 1993; Wadenberg and Hicks 1999; Wadenberg et al. 2001). However, there are known synergies between DA and 5-HT as well as this opponency (for instance, 5-HT2A receptors in the striatum appear to boost the effects of dopamine), making this prediction complicated. Further, systemic opioids and ACh-muscarinic agonists reportedly have less pronounced neuroleptic-like effects on the CAR (Shannon et al. 1999; Aguilar et al. 2004).

As currently constituted, our model incorporates an important asymmetry between dopamine and its putative opponent. (a) Value learning depends on PEs from both dopamine and its opponent. However, (b) action learning and extinction depend exclusively on dopaminergic activity being greater than and less than its baseline, respectively. In the latter case, note that a PE whose net affective value is aversive is represented by a net negative dopamine signal. This is partly a placeholder for what must surely be a more complicated relationship underpinning other phenomena too such as the nature (Tobler et al 2003) and non-extinction of conditioned inhibitors. We postulate this asymmetry because of the effect of dopamine blockade, with aversive values being well learned when DA is blocked (a), unlike avoidance (b). Unlike current theories (Smith et al. 2006), a forward-model-based account that would respect the DA-blockade data would also have to differentiate between the roles of DA in appetitive vs. aversive learning and in forward-model structure vs. action-preference learning.

One possible rationale for this asymmetry is that, at least in this particular context, there is much more information in the one avoidance action than in the many actions that do not prevent the oncoming shock, a fact that the subjects could learn whilst flailing. However, it may be instead that a more fundamental role for tonic levels of dopamine in controlling action vigour (Niv 2007; Niv et al. 2007) could include an effect of completely blocking the impetus to learn about effortful actions such as shuttling. Testing this possibility would require dissociating tonic and phasic

dopamine signalling, something of active interest in the literature (Grace 2000; Cagniard et al. 2006; Goto et al. 2007).

Our model only treats part of the involvement of dopamine in CAR, leaving out effects of tonic or sustained concentrations or release, and also involvement in freezing. The release of DA in aversive situations has been amply demonstrated using microdialysis, and is known not to be simply due to the offset of punishment, either through avoidance or otherwise (Young 2004). One interpretation of this again comes back to tonic dopamine signalling, arguing that this release is associated with an expectation that an effortful, vigorous, avoidance or escape action will be required (Niv et al. 2007). The main effect of this in the model would be to enhance the sloth of acquisition and the speed of extinction, the latter simply by reducing exposure to the consequences of the 'shuttle' action (data not shown). Freezing behaviour in response to shock is usually thought of as a Pavlovian, simpler paradigm than the CAR. Pavlovian responses to aversive events are, in fact, modulated in complicated and controversial ways by different dopaminergic drugs (Miyamoto et al. 2004; Reis et al. 2004; de Oliveira et al. 2006) and there are interesting avenues for investigation as to the best interpretation of this with respect to PEs (Iordanova et al. 2006).

Relevance for human psychopathology and future research

There are good psychopathological grounds for focussing on aversive processing as an essential component of psychosis. The literature has traditionally concentrated on neurotic disorders (Forsyth et al. 2006), but recent research highlights the importance of the aversive learning system in conditions where dopamine plays an important role. People suffering from paranoid psychosis often have a background of victimization (Mirowsky and Ross 1983; Janssen et al. 2003) and thus exposure to aversive learning. They tend to make exaggerated predictions of aversive events, even considering this background (Kaney et al. 1997; Corcoran et al. 2006; Bentall et al. in press). Their use of safety-behaviours (strategies for avoiding situations associated with perceived threat) seems to perpetuate their problems (Freeman et al. 2001). We have argued (Moutoussis et al. 2007) that efforts to avoid both external threats and, crucially, threatening private experiences (Bach and Hayes 2002) contribute to the aetiology of persecutory syndromes. The current study provides a detailed basis for linking the psychology of anticipation of threat and the biology of psychosis. It permits, for example, calculation of regression coefficients relevant to avoidance-learning, so as to help locate corresponding anatomical areas by functional imaging (as per Seymour et al. 2004). It also suggests a set of functional mechanisms whose activity may be exaggerated in persecutory syndromes.

Our model makes some predictions that would be important to put to experimental test. First, the basic patterns of avoidance behaviour considered here in the CAR, as well as their modulation by dopaminergic drugs, should correspond to analogous patterns in healthy humans. However, unfortunately, we have been able to find little evidence on the effect of psychotropic drugs on human aversive learning. Second, we might expect unmedicated, psychosis-prone individuals to show biased *learning* in CAR-like situations. Third, our model suggests that the 'better than expected' signal that follows perceived avoidance of an aversive outcome is reported by dopamine, both within value-learning and

within action-learning circuitries. This is consistent with a large body of psychological research stressing the reinforcing property of reaching 'safety states' and of some instances of CS offset (e.g. Gray 1987). It is also consistent with recent fMRI evidence (Kim et al. 2006). We note that within our framework it is not the offset of an aversive stimulus *per se* that would be associated with positive dopamine activity, but its better-than-expected significance. Experimentally, how closely dopamine is involved and whether it has similar roles in value vs. action learning remains to be tested directly. Fourth, if the dopaminergic-opponent process relies on serotonin, we would expect 5HT blockade significantly to delay acquisition but not extinction. Finally we would expect much less 'rebound' action-learning on removal of 5HT blockade, as compared to DA. The activity of the goal-directed system could complicate outcomes and therefore results would be clearer in experiments where the goal-directed system is inhibited (Blundell et al. 2003; Killcross and Coutureau 2003).

In considering the psychobiological mechanisms involved in the CAR a clinically important corollary immediately follows which lies outside the scope of the present study, but is highly relevant for future research. The psychological processes examined here, especially vigour, detection of threat and the anticipation of rewards of actions, are very important in manic disorders. These are characterized by pathological anticipation of *positive* returns. Temporal difference modelling may help clarify the neurobiological mechanisms involved and link them to emerging psychological research, for example to the precipitation of manic episodes by positive goal attainment events (Johnson et al. 2000).

Finally, the present models help set the scene for an investigation of the cooperative and competitive interactions of the habitual, forward-model and Pavlovian controllers in aversive processing. That the goal-directed controller favoured by expectancy theory does not appear to control performance on at least some of the existing collection of animal CAR studies, does not imply that it will not trump our habitual controller under any circumstance. Further, there also appears to be a third, Pavlovian, controller, which directly couples predictions of aversion to stereotypical defensive actions (Gray and McNaughton 1996; Dayan et al. 2006) in a way that can be synergistic or opponent to the instrumental choices of either the habitual or the goal-directed controller. Indeed, perhaps the most important future direction for this work is towards understanding the integration of these three mechanisms in terms of learning and performance, and thus, most optimistically, the potential combination of pharmacological and learning-based interventions in the treatment of psychotic psychopathology.

Acknowledgements

We are very grateful to Nathaniel Daw, Quentin Huys, Yael Niv and Ben Seymour for discussions. Funding was from the Gatsby Charitable Foundation (PD).

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Aguilar MA, Minarro J, Simon VM. 2004. Morphine potentiates the impairing effects of neuroleptics on two-way active conditioned avoidance response in male mice. Progress in Neuropsychopharmacology & Biological Psychiatry 28:225–237.

Anisman H. 1978. Psychopharmacology of aversively motivated behavior. New York: Plenum Press. pp 1–62.

Bach P, Hayes SC. 2002. The use of acceptance and commitment therapy to prevent the rehospitalization of psychotic patients: a randomized controlled trial. Journal of Consulting and Clinical Psychology 70:1129–1139.

Bardin L, Auclair A, Kleven MS, Prinssen EP, Koek W, Newman-Tancredi A, Depoortere R. 2007. Pharmacological profiles in rats of novel antipsychotics with combined dopamine D2/serotonin 5-HT1A activity: Comparison with typical and atypical conventional antipsychotics. Behavioural Pharmacology 18:103–118.

Bayer HM, Glimcher PW. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141.

Belin D, Everitt BJ. 2008. Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum. Neuron 57:432–441.

Beninger R, Mason S, Phillips A, Fibiger H. 1980a. The use of conditioned suppression to evaluate the nature of neuroleptic-induced avoidance deficits. Journal of Pharmacology and Experimental Therapeutics 213:623–627.

Beninger R, Mason S, Phillips A, Fibiger H. 1980b. The use of extinction to investigate the nature of neuroleptic-induced avoidance deficits. Psychopharmacology (Berlin) 69:11–18.

Bentall RP, Kinderman P, Howard R, Blackwood N, Cummins S, Rowse G, et al. In press. Paranoid delusions in schizophrenia and depression: The transdiagnostic role of negative self-esteem. Journal of Nervous Mental Disease (in press).

Blundell P, Hall G, Killcross S. 2003. Preserved sensitivity to outcome value after lesions of the basolateral amygdala. Journal of Neuroscience 23:7702–7709.

Cagniard B, Beeler JA, Britt JP, McGehee DS, Marinelli M, Zhuang X. 2006. Dopamine scales performance in the absence of new learning. Neuron 51:541–547.

Christy D, Reid L. 1975. Methods of deconditioning persisting avoidance: Amphetamine and amobarbital as adjuncts to response prevention. Bulletin of the Psychonomic Society 5:175–177.

Cooper S, Coon K, Mejta C, Reid L. 1974. Methods of deconditioning persisting avoidance: Amphetamine chlorpromazine and chlordiazepoxide as adjuncts to response prevention. Physiological Psychology 2:519–522.

Corcoran R, Cummins S, Rowse G, Moore R, Blackwood N, Howard R, Kinderman P, Bentall R. 2006. Reasoning under uncertainty: Heuristic judgements in patients with persecutory delusions or depression. Psychological Medicine 36:1109–1118.

Coutureau E, Killcross S. 2003. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. Behavioural Brain Research 146:167–174.

Daw N, Kakade S, Dayan P. 2002. Opponent interactions between serotonin and dopamine. Neural Networks 15:603–616.

Daw N, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience 8:1704–1711.

Dayan P, Balleine BW. 2002. Reward motivation and reinforcement learning. Neuron 36:285–298.

de Oliveira AR, Reimer AE, Brandao ML. 2006. Dopamine D2 receptor mechanisms in the expression of conditioned fear. Pharmacology, Biochemistry & Behavior 84:102–111.

Deakin J, Graeff F. 1991. 5-HT and mechanisms of defence. Journal of Psychopharmacology 5:305–315.

Forsyth J, Eifert G, Barrios V. 2006. In: Craske MG, Hermans D, Vansteenwegen D, editors. Fear and Learning: From basic processes to clinical implications. Washington, DC: American Psychological Association. pp 133–153.

Freeman D, Garety PA, Kuipers E. 2001. Persecutory delusions: Developing the understanding of belief maintenance and emotional distress. Psychological Medicine 31:1293–1306.

Goto Y, Otani S, Grace AA. 2007. The Yin and Yang of dopamine release: A new perspective. Neuropharmacology 53:583–587.

Grace AA. 2000. The tonic/phasic model of dopamine system regulation and its implications for understanding alcohol and psychostimulant craving. Addiction 95 (Supplement 2) S119–S128.

Gray J. 1987. The psychology of fear and stress. New York, NY: McGraw-Hill.

Gray JA, McNaughton N. 1996. The neuropsychology of anxiety: Reprise. Nebraska Symposia on Motivation 43:61–134.

Grossberg S. 1972. A neural theory of punishment and avoidance. Mathematical Biosciences 15:39–68.

Horvitz JC. 2000. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience 96:651–656.

Iordanova MD, Westbrook RF, Killcross AS. 2006. Dopamine activity in the nucleus accumbens modulates blocking in fear conditioning. European Journal of Neuroscience 24:3265–3270.

Janssen I, Hanssen M, Bak M, Bijl RV, De Graaf R, Vollenberg W, et al. 2003. Discrimination and delusional ideation. British Journal of Psychiatry 182:71–76.

Jensen A, Sauerberg P, Jeppesen L, Sheardown M, Swedberg M. 1999. Muscarinic receptor agonists like dopamine receptor antagonists antipsychot inhibit conditioned avoidance response in rats. Journal of Pharmacology and Experimental Therapeutics 290(2):901–907.

Jensen J, McIntosh AR, Crawley AP, Mikulis DJ, Remington G, Kapur S. 2003. Direct activation of the ventral striatum in anticipation of aversive stimuli. Neuron 40:1251–1257.

Johnson J, Li W, Li J, Klopf A. 2002. A computational model of learned avoidance behavior in a one-way avoidance experiment. Adaptive Behavior 9:91–104.

Johnson S, Sandrow D, Meyer B, Winters R, Miller I, Solomon D, Keitner G. 2000. Increases in manic symptoms after life events involving goal attainment. Journal of Abnormal Psychology 109:721–727.

Kaney S, Bowen-Jones K, Dewey ME, Bentall RP. 1997. Frequency and consensus judgements of paranoid paranoid-depressed and depressed psychiatric patients: Subjective estimates for positive negative and neutral events. British Journal of Clinical Psychology 36:349–364.

Killcross S, Coutureau E. 2003. Coordination of actions and habits in the medial prefrontal cortex of rats. Cerebral Cortex 13:400–408.

Killcross S, Robbins TW, Everitt BJ. 1997. Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. Nature 388:377–380.

Kim H, Shimojo S, O'Doherty JP. 2006. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biology 4:e233.

Lovibond P. 2006. Fear and Avoidance: An integrated expectancy model. In: Craske MG, Hermans D, Vansteenwegen D, editors. Fear and Learning: From basic processes to clinical implications. Washington, DC: American Psychological Association. pp 117–132.

Ma T, Yu Q. 1993. Effect of 20(S)-ginsenoside-Rg2 and cyproheptadine on two-way active avoidance learning and memory in rats. Arzneimittelforschung 43:1049–1052.

McAllister D, McAllister W, Hampton S, Scoles T. 1980. Escape-from-fear performance as affected by handling method and an additional CS-shock treatment. Animal Learning Behavior 8:417–423.

McAllister WR, McAllister DE, Scoles MT, Hampton SR. 1986. Persistence of fear-reducing behavior: Relevance for the conditioning theory of neurosis. Journal of Abnormal Psychology 95:365–372.

Menon M, Jensen J, Vitcu I, Graff-Guerrero A, Crawley A, Smith MA, Kapur S. 2007. Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: Effects of dopaminergic modulation. Biological Psychiatry 62:765–772.

Mineka S. 1979. The role of fear in theories of avoidance learning flooding and extinction. Psychological Bulletin 86:985–1010.

Miyamoto J, Tsuji M, Takeda H, Ohzeki M, Nawa H, Matsumiya T. 2004. Characterization of the anxiolytic-like effects of fluvoxamine milnacipran and risperidone in mice using the conditioned fear stress paradigm. European Journal of Pharmacology 504:97–103.

Montague R, Dayan P, Sejnowski T. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of Neuroscience 76:1936–1947.

Moutoussis M, Williams J, Dayan P, Bentall RP. 2007. Persecutory delusions and the conditioned avoidance paradigm: Towards an integration of the psychology and biology of paranoia. Cognitive Neuropsychiatry 12:495–510.

Mowrer O. 1947. On the dual nature of learning: A reinterpretation of conditioning and problem solving. Harvard Educational Review 17:102–148.

Niv Y. 2007. Cost benefit tonic phasic: What do response rates tell us about dopamine and motivation?. Annals of the New York Academy of Sciences 1104:357–376.

Niv Y, Daw ND, Joel D, Dayan P. 2007. Tonic dopamine: Opportunity costs and the control of response vigor. Psychopharmacology (Berlin) 191:507–520.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan R. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.

Poremba A, Gabriel M. 1995. The amygdala is necessary for the initial acquisition but not for maintenance of discriminative avoidance behavior in rabbits. Society of Neuroscience Abstracts 21:1930.

Poremba A, Gabriel M. 1999. Amygdala neurons mediate acquisition but not maintenance of instrumental avoidance behavior in rabbits. Journal of Neuroscience 19:9635–9641.

Reis FL, Masson S, de Oliveira AR, Brandao ML. 2004. Dopaminergic mechanisms in the conditioned and unconditioned fear as assessed by the two-way avoidance and light switch-off tests. Pharmacology, Biochemistry & Behavior 79:359–365.

Rosenkranz JA, Grace AA. 2001. Dopamine attenuates prefrontal cortical suppression of sensory inputs to the basolateral amygdala of rats. Journal of Neuroscience 21:4090–4103.

Rosenkranz JA, Grace AA. 2002. Cellular mechanisms of infralimbic and prelimbic prefrontal cortical inhibition and dopaminergic modulation of basolateral amygdala neurons in vivo. Journal of Neuroscience 22:324–337.

Samaha AN, Seeman P, Stewart J, Rajabi H, Kapur S. 2007. ''Breakthrough'' dopamine supersensitivity during ongoing antipsychotic treatment leads to treatment failure over time. Journal of Neuroscience 27:2979–2986.

Schmajuk N, Zanutto B. 1997. Escape avoidance and imitation: A neural network approach. Adaptive Behavior 6:63–129.

Schultz W, Apicella P, Ljungberg T. 1993. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience 13:900–913.

Seligman M, Johnston J. 1973. Contemporary approaches to conditioning and learning Washington, DC: Winston. pp 69–110.

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frackowiak RS. 2004. Temporal difference models describe higher-order learning in humans. Nature 429:664–667.

Seymour B, Singer T, Dolan R. 2007. The neurobiology of punishment. Nature Reviews Neuroscience 8:300–311.

Shannon H, Hart J, Bymaster F, Calligaro D, DeLapp N, Mitch C, Ward J, Fink-Jensen A, Sauerberg P, Jeppesen L, et al. 1999. Muscarinic receptor agonists like dopamine receptor antagonist antipsychotics inhibit conditioned avoidance response in rats. Journal of Pharmacology and Experimental Therapeutics 290:901–907.

Siuciak JA, Chapin DS, McCarthy SA, Guanowsky V, Brown J, Chiang P, Marala R, Patterson T, Seymour PA, Swick A, et al. 2007. CP-809101 a selective 5-HT2C agonist shows activity in animal models of antipsychotic activity. Neuropharmacology 52:279–290.

Smith A, Becker S, Kapur S. 2005. A computational model for the functional role of ventral-striatal d2 receptor in the expression of previously acquired behaviors. Neural Computation 17:361–395.

Smith A, Li M, Becker S, Kapur S. 2006. Dopamine prediction error and associative learning: A model-based account. Network: Computation in Neural Systems 17:61–84.

Smith A, Li M, Becker S, Kapur S. 2007. Linking animal models of psychosis to computational models of dopamine function. Neuropsychopharmacology 32:54–66.

Solomon RL, Corbit JD. 1974. An opponent-process theory of motivation: Temporal dynamics of affect. Psychological Review 81:119–145.

Solomon R, Wynne L. 1953. Traumatic avoidance learning: Acquisition in normal dogs. Psychological Monographs 67:19.

Sutton R, Barto A. 1998. Reinforcement learning: An introduction Cambridge, Massachusetts: MIT Press.

Titov SA, Shamakina IIu Ashmarin IP. 1983. Effect of lysyl vasopressin and vasotocin on a disorder of the conditioned avoidance reaction by a serotonin receptor blockader. Biulleten Eksperimentalnoi Biologii I Meditsiny 95:31–33.

Tobler PN, Dickinson A, Schultz W. 2003. Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. Journal of Neuroscience 23:10402–10410.

Unger W, Evans I, Rourke P, Levis D. 2003. The s-s construct of expectancy versus the s-r construct of fear: Which motivates the acquisition of avoidance behaviour ? The Journal of General Psychology 130:131–147.

Ungless M. 2004. Dopamine: The salient issue. Trends in the Neurosciences 27:702–706.

Wadenberg M, Browning J, Young K, Hicks P. 2001. Antagonism at 5-HT(2A) receptors potentiates the effect of haloperidol in a conditioned avoidance response task in rats. Pharmacology, Biochemistry & Behavior 68:363–370.

Wadenberg M, Hicks P. 1999. The conditioned avoidance response test re-evaluated: Is it a sensitive test for the detection of potentially atypical antipsychotics ? Neuroscience and Biobehavioral Reviews 23:851–862.

Williams JOH, Dayan P. 2005. Dopamine, learning and impulsivity: A biological account of ADHD. Journal of Child and Adolescent Psychopathology 15:160–179.

Wise R, Spindler J, deWit H, Gerberg G. 1978. Neuroleptic-induced ''anhedonia'' in rats: Pimozide blocks reward quality of food. Science 201:262–264.

Wolpert D, Miall R. 1996. Forward models for physiological motor control. Neural Networks 9:1265–1279.

Yacubian J, Glascher J, Schroeder K, Sommer T, Braus DF, Buchel C. 2006. Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. Journal of Neuroscience 26:9530–9537.

Young AM. 2004. Increased extracellular dopamine in nucleus accumbens in response to unconditioned and conditioned aversive stimuli: Studies using 1 min microdialysis in rats. Journal of Neuroscience Methods 138:57–63.

Appendix: Details of simulation results

During the first phase of the EFF, from trials 100 on in Figure 4, subjects have learnt that all states 1–5 predict shock for all actions. Within about 20 trials, all states have values about equal to the return of the shock. During the second phase 'shuttle' actions lead to safety states of value zero. This takes place from trial 150 onwards in the example of Figure 4(b) (same example in Figure A1 and A2). Therefore when these shuttle actions are taken, the values of the originating states reduce according to Equation 1. In our example, the value of state 4 decays from 4 (= return of shock) towards 0.2 (= return of shuttle) each time shuttling occurs from that state, as happens in trials 152, 154, etc. The reduction in value of state 5 reduces for 'stay' too, as there is now no shock (return = 0). If we take the first non-shocked trial, trial 151, as an example:

$$\begin{aligned}\delta_V(5) &= R_{non\text{-}shocked}\,(5 \to 6) + V(6)_{\text{old}} - V(5)_{\text{old}} \\ &= 0 + 0 - 4 = -4 \quad \text{(from Equation 1)}\end{aligned}$$

This gives

$$\begin{aligned}V(5)_{\text{new}} &= V(5)_{\text{old}} + 0.5 \times \delta_V(5) = 4 + 0.5^*(-4) \\ &= 2, \quad \text{(Equation 2, as per Figure A1)}\end{aligned}$$

Note the unusual dynamics for the value of state 1. When shuttling occurs, V(S1) reduces as expected. However, since S2 is little visited, its value (V(S2)) is not greatly reduced (in fact, mostly what was inherited from S4), Thus, when 'stay' is chosen in S1, and so S2 is visited, the value of S1 usually increases sharply (trials 153, 155, 169 etc.) because of ongoing learning.

These phenomena have a very interesting effect on policy, shown in Figure A2. The advantages for the action 'shuttle' (upper curve) and 'stay' (lower curve). While V(S1) is high, each 'shuttle' leads to learning that this action is more advantageous, according to Equation 4. Let us take trial 153 as an example. Here shuttling occurs for the first time from state 1 (to the 'safe' state 7). We now have, as above:

$$\delta_V(1) = R_{shuttle}(1 \to 7) + V(7) - V(1)_{\text{old}} = 0.2 + 0 - 4 = -3.8$$

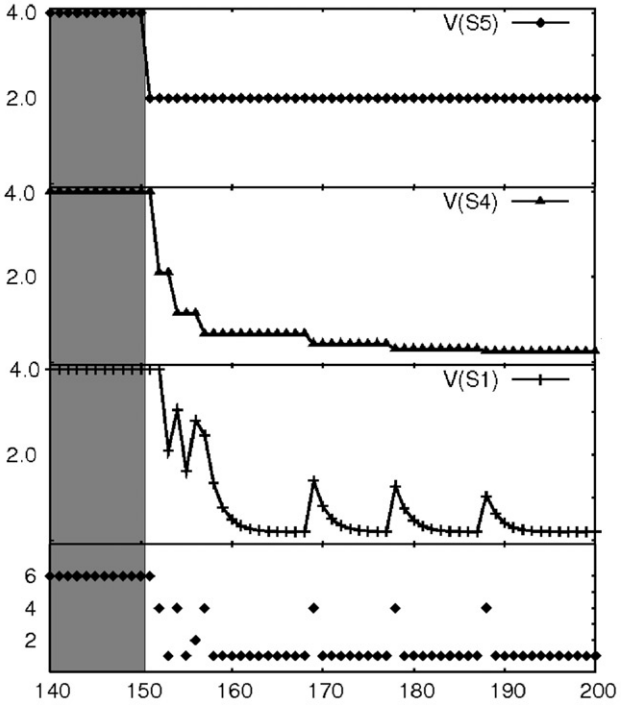

Figure A1.

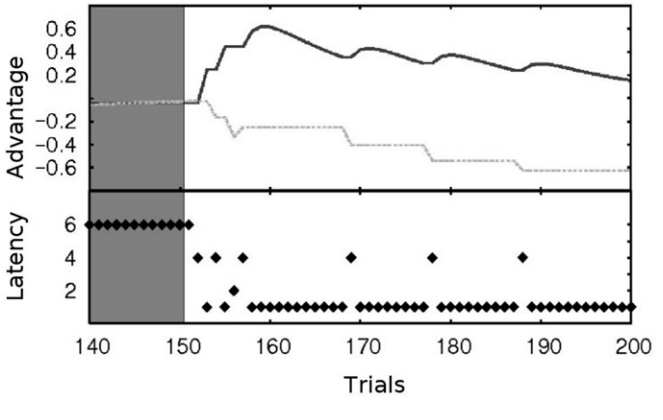

Figure A2.

and, substituting in Equation 4 while reversing the sign of the value of δV to respect the convention of the model description above,

$$m(\mathrm{a} = \mathrm{shuttle}, \mathrm{s} = 1)_{\mathrm{new}} = m(\mathrm{shuttle}, 1)_{\mathrm{old}} + 0.075^{*}; (\delta_V - m(\mathrm{shuttle}, 1)\mathrm{old})$$
$$= 0 + 0.075^{*}(3.8 + 0) = 0.285 \quad \text{(as per Figure A2)}$$

In addition, the occasional 'stay' actions in S1 lead to visiting the unextinguished S2, and hence teach that 'stay' is disadvantageous. This happens, in trials 154, 169 etc. in our example. The policy of shuttling from this early state is therefore boosted (Equation 3). This is similar to the psychological explanation that has been given for the tendency of the CAR to persist. Conversely, delaying the AR exposes the animal to later, still unextinguished, parts of the CS, increasing momentary fear, and thus delaying extinction.