British Journal of Psychology (2009), 100, 399–413

The British Psychological Society

www.bpsjournals.co.uk

Prejudiced learning: A connectionist account

J. Richard Eiser[1]*, Tom Stafford[1] and Russell H. Fazio[2]

[1]University of Sheffield, Sheffield, UK

[2]Ohio State University, Columbus, Ohio, USA

Connectionist simulation was employed to investigate processes that may underlie the relationships between prior expectancies or prejudices and the acquisition of attitudes, under conditions where learners can only discover the valence of attitude objects through directly experiencing them. We compared contexts analogous to learners holding either false negative expectancies ('prejudices') about a subclass of objects that were actually good or false positive expectancies about objects that were actually bad. We introduced expectancy-related bias either by altering the probability of approach, or by varying the rate of learning following experience with good or bad objects. Where feedback was contingent on approach, the false positive expectancies were corrected by experience, but negative prejudices resisted change, since the network avoided objects deemed to be bad, and so received less corrective feedback. These findings are discussed in relation to the effects of intergroup contact and expectancy-confirmation processes in reducing or sustaining prejudice.

From a meta-analytic review of 515 experimental and correlation studies, Pettigrew and Tropp (2006) recently concluded in favour of the longstanding hypothesis that intergroup contact reduces prejudice. The effect sizes for the contact-prejudice association averaged around − .21, but with considerable hetereogenity. This evidently leaves room for studying the influence of other variables. Some support was found for Allport's (1954) suggestion that the association would be stronger if the contact situation had features such as equal status, common goals, cooperation, and societal support, although these were not necessary conditions for prejudice reduction as he originally proposed. Pettigrew and Tropp note that much less research has sought to identify processes that make it less likely that contact will lead to a reduction in prejudice.

An intuitively plausible possibility is that prejudiced individuals may be less likely to expose themselves to, and/or believe, evidence that challenges their prejudices. In simple terms, intergroup contact provides information that may confirm or disconfirm prior beliefs to varying extents. Since prejudice is conceptualized as involving prior beliefs that are both negative in valence and incorrect in fact, the

*Correspondence should be addressed to Professor J. Richard Eiser, Department of Psychology, University of Sheffield, Sheffield S10 2TP, UK (e-mail: j.r.eiser@shef.ac.uk).

DOI:10.1348/000712608X357849

implication is that intergroup contact will typically disconfirm such beliefs and lead to their change. At the same time – to the extent that they have a choice – prejudiced individuals may be less likely to seek out contact with out-group members, and even when they do, may be resistant to change their viewpoints on the basis of a few positive experiences.

The issue of choice is considered by Pettigrew and Tropp (2006). Their concern was that the overall negative association between contact and prejudice might reflect a selection bias, with only non-prejudiced individuals seeking intergroup contact. This implies that the association should be stronger where participants had a choice over whether to engage in intergroup contact than where participants had no such choice. In fact, the reverse was the case. The effect size of the association was significantly larger (− .28) in those studies where participants had no choice. This implies that the contact-prejudice association is not an artefact of a selection bias. Nonetheless, several questions remain.

First, exactly what is it about contact that produces this effect? The assumption that contact provides evidence that disconfirms prejudiced beliefs is very difficult to test in field studies where mostly prejudice is operationalised as a general attitude and the feedback provided by intergroup contact (related to the validity of prior beliefs) is hard to control or identify. Also, it is unclear whether the important issue is whether initially prejudiced individuals have their false *beliefs* concerning out-group members disconfirmed through greater contact, or whether, more simply, they change their *evaluations* of out-group members in a more favourable direction, presumably as a function of positive interaction experiences.

Second, the nature and selectivity of the intergroup contact in which individuals engage may depend on how much choice they have, even though the overall negative association between contact and prejudice is no stronger (in fact, Pettigrew and Tropp found it to be weaker) under conditions of free choice. If, given free choice, individuals with negative attitudes engage in more restricted forms of intergroup contact, this means that they should receive less feedback (that might disconfirm their prejudices) from interacting with out-group members than those with positive attitudes.

How much, and what kind of, intergroup contact is actually engaged in by individuals with different attitudes, and how far such contact is freely chosen, are questions that fall beyond the scope of our present research. Instead, we focus on how any such contact experiences may shape, and be shaped by, processes of attitude formation and change. Previous empirical research confirms the existence of a general contact-prejudice association but does not exclusively point to a single causal explanation: contact may reduce prejudice, but prejudice may also reduce contact. This suggests we are dealing with complementary processes that interact to form a common dynamical system (Eiser, 1994).

To examine such processes, we reshape the research question in more abstract terms. Rather than consider real social groups, we use novel objects of arbitrary predefined valence, and rather than consider actual intergroup contact, we use a task in which attitude acquisition requires that these novel objects are sampled or 'approached'. The paradigm in question is a computer game, termed BeanFest, developed by Fazio, Eiser, and Shook (2004), in which players have to survive in a virtual world consisting entirely of 'beans'. Their task is to identify and eat good beans that provide energy while avoiding bad beans that are poisonous and lead to a loss of energy. In the critical conditions, the valence of these beans can only be discovered once they

are 'eaten'. Hence learning depends on approaching, or sampling, the novel objects. The central finding is a 'learning asymmetry' effect whereby bad beans tend to be better learnt than good beans, some of which tend to be misclassified, and hence avoided, as though they are bad.

In the standard version of BeanFest, players have no prior attitudes concerning any of the beans. However, in the final experiment reported by Fazio *et al.* (2004), participants received advice, supposedly from one previous player and corroborated by a second, that one of the (six) subclasses of beans presented was either good, and to be eaten, or bad, and to be avoided. Orthogonally, this information was either correct or incorrect. The main finding was that, when participants were told to avoid beans that were actually good, they tended to follow this advice and so not receive feedback to show that these negative beliefs or 'prejudices' were false. Conversely, if participants were advised to eat beans that were actually bad, they discovered the error and corrected such false positive beliefs. In other words, participants overcame invalid positive prejudices, but not invalid negative prejudices.

Applied to the issue of intergroup contact, these findings are consistent with an interpretation that contact may reduce prejudice through providing experiences that disconfirm negative attitudes or expectancies. But is the *only* part played by prejudice in this learning process one of inhibiting contact and hence exposure to disconfirming feedback? Research on expectancy-confirmation biases (e.g. Darley & Fazio, 1980; Darley & Gross, 1983), suggests that information is interpreted in the light of pre-existing hypotheses or expectancies. Prejudices, to the extent that they imply (negative) expectancies, may resist change through influencing individuals' interpretations of feedback through intergroup contact, and leading them to protect their prior beliefs by discounting the force of troublesome evidence (see e.g. Abelson, 1959; Janis & Mann, 1977).

Hence, two potentially complementary processes may be involved. The first is that prior attitudes influence *action selection*, that is whether or not a novel object is approached. This is analogous to positive attitudes leading to more intergroup contact and negative prejudices leading to intergroup avoidance. Since avoidance means less feedback, false negative beliefs are less likely to be corrected and should therefore resist change more than false positives. The second possible process is that prior attitudes, expectancies, or prejudices influence how individuals interpret the feedback they receive from sampling novel objects. In other words, *after feedback is received* (as a consequence of approach), its impact on the learner's representations of the objects will be constrained by prior expectancies. This is analogous to prejudiced individuals, even when they experience intergroup contact, interpreting such experience differently from non-prejudiced individuals so as to leave their prejudice intact.

The present paper seeks to examine these processes through connectionist computer simulation. This choice of methodology arises from two considerations. Firstly, the processes identified above may be difficult to distinguish in field or experimental studies, whereas they can be operationalized separately within an artificial learning system. Secondly, precisely because we are dealing with an artificial system, we are required to attempt to develop an account of the dynamics of prejudiced learning without a premature appeal to the involvement of higher mental processes or even conscious awareness. Hence, if an effect conceptually resembling prejudice can be simulated on a computer, that is an 'automaton', this adds weight to the idea that processes underlying prejudice in humans may operate to a great extent 'automatically' (cf. Bargh & Ferguson, 2000; Fazio, 2001).

Method

Overview

The simulations to be described employ a form of connectionist neural network (Ellis & Humphreys, 1999; Read & Miller, 1998; Smith, 1996; Van Overwalle & Siebler, 2005). A neural network, or net, consists of interconnected nodes or units. Activation spreads between the units as a function of the strength, or weight, of the connection between them. These connection weights are modified through training. Commonly this involves comparing the output produced by the net in response to some input with a predefined target value or 'correct answer'. The connection weights are then modified so that, on the next presentation of the input, the discrepancy (Δ) between the output and target is reduced, by means of a defined learning algorithm, We here use an adaptation of the algorithm described in detail by Eiser, Fazio, Stafford, and Prescott (2003) (see also, Eiser, Stafford, & Fazio, 2008), who achieved a successful simulation of the BeanFest game and its central finding of a 'learning asymmetry', whereby good objects ('beans') are less well learnt than bad objects.

The main innovation introduced by Eiser *et al.* (2003) was to compare a 'full feedback' condition, where connection weights are modified on every trial, with a 'contingent feedback' condition, where weight modification (analogous to learning in the human experiments) only occurs when the net selects an action representative of approach. Essentially this transforms the training procedure from one of 'supervised' to reinforcement learning (Sutton & Barto, 1998). We here report results obtained under contingent feedback only,[1] but extend this method by separately manipulating an *action bias* and a *learning bias*. The first of these refers to a bias towards approaching ('eating') or avoiding a specific subclass of objects, and is analogous to a bias towards greater or lesser intergroup contact. The second is analogous to individuals placing less reliance (and so exhibiting slower learning) when confronted with feedback that is inconsistent with their prior beliefs.

The inputs presented to the network comprised 36 different patterns forming six clusters in different regions of a two-dimensional array (see Figure 1). These are equivalent formally to the 'beans' presented to participants in Fazio *et al.* (2004), with one axis representing the bean's shape (circular to oval) and the other, the number of speckles. Particularly relevant here are two regions: region 1, consisting of good beans and region 6, consisting of bad beans. Our manipulation of (action selection and/or learning) bias were designed so as either to make the net treat the beans in region 1, all of which are good, as though they were bad (*negative bias*), or the beans in region 6, all of which are bad, as though they were good (*positive bias*). We report the performance of the network with respect to identification of beans in these two regions[2] at the end of 5,000 epochs of training (one epoch involving a single presentation of each of the 36 input patterns). Three sets of simulations are described, involving combinations of full versus contingent feedback and negative versus positive bias, operationalized through action selection and/or learning (the rate of weight modification following feedback).

[1] *In fact, simulations were also run under full feedback. These yielded perfect or near perfect identification of both good and bad objects.*

[2] *Data are not reported for identification of the 'beans' in the other four regions, since these were not the focus of the manipulations. However, these replicated the 'learning asymmetry' effect observed in human data by Fazio* et al. *(2004) and simulations by Eiser* et al. *(2003).*

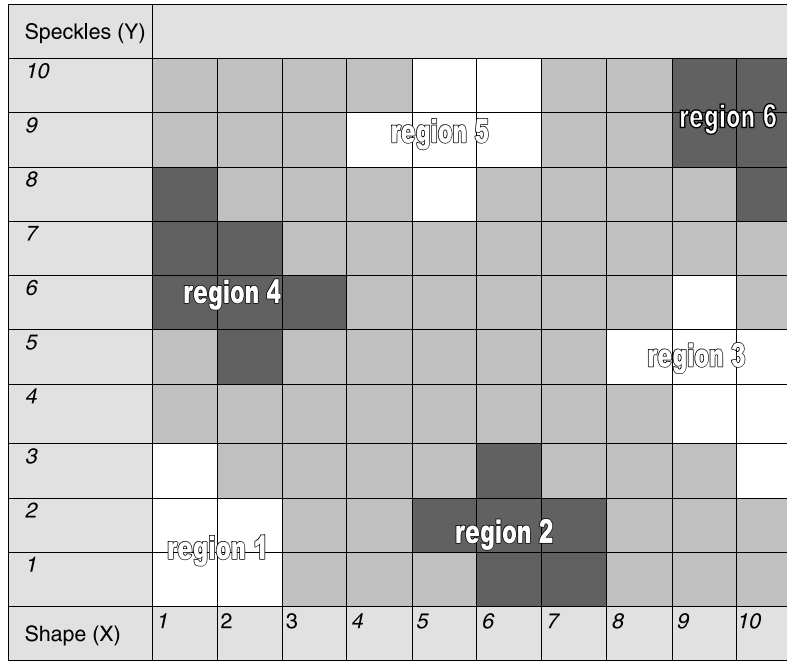

Figure 1. The BeanFest matrix. Clear squares (regions 1, 3, and 5) represent 'good beans' and dark grey squares (regions 2, 4, and 6) represent 'bad beans'. Region 1 was the target of negative bias, region 6 of positive bias.

Network architecture

The network architecture is as described in Eiser *et al.* (2003; Study 2) and comprises a separate learning system and action selection mechanism. The *learning system* is a fully connected, three-layer, feed-forward network. The first layer comprises 22 *input* units, of which 11 are used to encode one dimension (e.g. shape) and the remaining 11 the other dimension (speckles). These input units take values between 0 and 1, with each level of an attribute being represented by activation (>0) on up to 6 of the 11 units. For example, one speckle would be encoded by the vector [1,1,0.5,0.25,0,0,0,0,0,0,0], 4 speckles as [0,0.25,0.5,1,1,0.5,0.25,0,0,0,0], through to 10 speckles as [0,0,0,0,0,0,0,0.25,0.5,1,1]. This method of encoding was chosen so that adjacent (similar) levels of an attribute are represented by overlapping (similar) patterns of activation. As a result, the net not only identifies which stimuli are which, but also recognizes how similar individual stimuli are to one other.

The second layer comprises three *hidden* units. Inclusion of a hidden layer is required because the categories of stimuli to be discriminated, as in Figure 1, are not linearly separable: in other words, one cannot separate the good and bad regions by drawing a single straight line across the matrix. These hidden units produce condensed representations of the input patterns. The way this happens is that, whenever a stimulus is presented to the network, the activations of all the 22 input units are fed forward separately to each of the hidden units, after multiplication by the weights on the respective (66) connections (This may be thought of as roughly analogous to applying *B*-coefficients to predictor variables in a multiple regression, treating each input unit as a predictor and each hidden unit in turn as a dependent variable). The resulting

activations of the three hidden units (analogous to the values of dependent variables after applying a regression equation to a set of predictors) are then fed forward, through weighted connections, to a third layer consisting of a single *output* unit (Again, one can think of this as regressing a single dependent variable on to three predictors). The resulting activation of this output unit represents the learning system's 'evaluation' of each stimulus, varying continuously between 0 (bad) and 1 (good).

This 'evaluation' is then fed forward to a separate action selection mechanism where a number of mathematical transformations are performed to decide whether the net will 'eat' (approach) or 'avoid' a given bean. The first of these is a logistic function that makes the evaluations more polarized as the move away from neutral (0.5) (This is a familiar feature of many connectionist simulations designed to make selected outputs more easily distinguished and hence more stable). The next transformation is a 'hunger function'. The purpose of this is to make the net less risk-averse, that is prepared to eat more beans even when doubtful of their valence, when it is 'hungry', that is low on 'energy' (This happens if the net has eaten too many bad beans and/or failed to eat enough good beans). We here use throughout the version of the function described by Eiser *et al.* (2003) as 'neutral hunger'. Since we are not manipulating this function, we are not examining any specific hypotheses about how it might impact on prejudice. However, one might speculate that, if individuals are failing to achieve particular goals that depend on social interaction, they may be less risk-averse in engaging in contact or relationships with out-group members. Finally, a stochastic (chance) aspect is introduced by comparing the resulting output with a randomly generated variable ('noise') between 0 and 1. If the output exceeds this random value, the net will select an action equivalent to 'eating' the bean, otherwise not. The effect of this is that, although positively evaluated beans will tend to be eaten most of the time, and negatively evaluated beans mostly avoided, this relationship is not deterministic. Occasionally, apparently good beans will be avoided and apparently bad beans will be eaten.

Training procedure

This follows the procedure described in Eiser *et al.* (2003), involving variations of the standard backpropagation of error ('backprop') algorithm (Rumelhart, Hinton, & Williams, 1986) to modify the connection weights (Parameter settings were 0.02 for the learning rate and 0.06 for momentum). Within each epoch, the network is presented, in a random order, with the 36 input patterns corresponding to the 18 good and 18 bad 'beans' shown in Figure 1.The learning system generates an evaluation between 0 and 1, with scores >0.5 representing a positive evaluation and <0.5, a negative evaluation of any given input pattern. This evaluation is then compared with a 'training signal' or 'target value' appropriate to that input pattern. This is here defined as 1 for a 'good' bean and 0 for a 'bad' bean. The discrepancy (target minus evaluation) defines the error (Δ), and the algorithm modifies the connection weights so as to reduce Δ on subsequent trials (Extending the previous analogy with regression analysis, this is like continuously updating a regression equation to improve prediction).

Contingent feedback

Under this procedure, Δ was calculated and connection weights were modified only when the net 'chose to eat' (i.e. when the action selection output equalled or

exceeded the 0.5 threshold). No learning (modification of weights) took place on any trial where the output was below threshold. This represents a situation where someone has to eat a bean to discover whether it is good or bad. If a bean is avoided, no information is provided about its true value and so no learning takes place on that trial.

Biasing action selection

This was achieved by transforming the input received by the action selection mechanism from the learning system. To reflect a *negative* action bias, all the evaluations (*after* conversion by the logistic function but *before* the addition of hunger and noise) *of the beans in region 1 only* (all of which are actually good) are multiplied by a scaling coefficient, S, where $S < 1$. This has the effect of reducing the probability of the selection of an 'eat' response at each level of judgment. In fact, scaled in this way, no judgment, no matter how positive, evokes a 100% chance of an 'eat' response. Conversely, a *positive* action bias is produced by multiplying the evaluations *of the beans in region 6 only* (all of which are actually bad) by the same scaling coefficient S but also adding a constant, $1 - S$. This means that no evaluation, no matter how negative, ensures a 100% chance of producing an 'avoid' response. Under *neutral* bias, no such scaling coefficient was applied. Following piloting, S was set at 0.5. The effect of this was that, under strong negative bias, even a maximally positive evaluation (1) of a good bean evokes only a 50% chance of an 'eat' response. Conversely, under strong positive bias, even the most negative possible evaluation of a bad bean (0) will still tend to lead to 'eating' 50% of the time. Following piloting, it was found appropriate to apply this bias at full strength for the first 300 epochs of training only. Thereafter the scaling coefficient was linearly attenuated through to epoch 900, after which time the standard (control) version of the action selection algorithm was applied. The sense of this attenuation is to allow for the fact that initial biases, possibly based on indirect experience such as information gained from others (as in Fazio *et al.*, 2004, Experiment 5), may be more highly weighted in the early stages of learning, before individuals gain direct experience for themselves.

Biasing learning as a function of expectancies

The purpose of this manipulation was to simulate the situation where participants might put less credence in experience that contradicted prior 'advice' concerning the valence of particular beans in a specified region. Except in the control condition where no 'advice' was given, any feedback received by the net from approaching a bean in the specified region would contradict expectancies. Thus, if the net approached a bean in region 1 (all of which are good), it would receive positive feedback that contradicted its negative 'expectancy'. Conversely, if it approached a bean in region 6 (all of which are bad), its positive expectancy would be contradicted. This was achieved by incorporating a 'learning bias' parameter (G) into the operation of the backprop algorithm. Recall that learning in this kind of net is driven by Δ, or the discrepancy between the output generated by the net and the target value for each input pattern. Our approach involved dividing Δ by G (where $G \geq 1$) on all trials where feedback was received concerning the valence of the beans in the relevant targeted region (1 or 6). Under *negative learning bias*, this applied only if the net received feedback relating to any of the (good) beans in region 1. Under *positive learning bias*, this applied only

following experience with any of the (bad) beans in region 6.This rescaling was *not* applied to feedback relating to beans in the remaining five regions.[3] Following piloting, G was set at 1.5.The effect of this was to slow down the rate at which connections weights would be modified following feedback that contradicted expectancies. A *neutral* condition ($G = 1$) was also included, meaning that no rescaling was applied.

STUDY 1: ACTION BIAS

This first study was designed to compare how evaluations of the beans in the two critical regions (1 and 6) might be affected by biasing the action selection mechanism early in training so that the net would be either (a) less likely to approach good beans in region 1 under negative bias or (b) more likely to approach bad beans in region 6 under positive bias. The design was a 3 × 2 (action selection bias × region: 1 vs. 6) factorial with repeated measures on the last factor. The learning bias parameter (G) was ignored, that is set to 1 (neutral) for all conditions. Note that the neutral conditions in this study are formally equivalent to all other neutral conditions in other studies reported in this paper, any differences in results being attributable to the randomized setting of initial weights and the stochastic element in the action selection mechanism. This study (like both subsequent ones) included 10 completed replications (i.e. runs with different randomly initialized weights) in each cell, with training continuing for 5,000 epochs. In two replications (out of 102), the net 'died' – that is failed to learn before running out of energy – so the runs were restarted with new random initial weights.

Results

The dependent variable of interest in the mean proportion of 'beans' in regions 1 and 6 correctly identified within each condition by the end of training. The bad beans in region 6 are perfectly identified throughout, regardless of the bias. Since we are dealing with the net's performance at the *end* of learning, any procedure that encourages the net to explore a particular region of the input space should lead to improved knowledge of the actual valence of stimuli within that region. Thus, bad beans in region 6 will still be avoided even after training with the positive action selection bias, since the net will have learnt to recognize these as bad. By contrast, identification of the good beans in region 1 is significantly suppressed to just 32% correct under negative bias. Since negative bias reduces the chance of such beans being sampled during training, the net has less chance of receiving feedback to correct its inital 'prejudice'. In the neutral bias (control) condition the figure in 90%, a significant difference ($t = 3.22, p < .005$). In the positive bias condition (i.e. where this bias is applied to presentations of bad beans in region 6 but not to any of the good beans from region 1), correct identification is again 90%. When the data for all conditions are submitted to a repeated measures analysis of variance, the predicted action bias × region interaction was significant, $F(2, 27) = 7.93$, $p < .002$. Thus, an action bias that encouraged contact eventually yielded an appreciation of the actual positive value of the beans in region 1, whereas an

[3] *Eiser, Stafford, and Fazio (2008) report simulations that vary the strength of the* G *parameter when applied to feedback relating to all 36 beans. The main finding is that learning of good beans is differentially impaired by negative expectancies about the valence of the set of beans as a whole.*

action bias that discouraged contact was more likely to promote the mistaken beliefs that these beans were negative.

STUDY 2: LEARNING BIAS

The design of this study was again a 3 × 2 (learning bias × region) factorial with repeated measures on the last factor, corresponding closely to that of study 1, except that the action selection bias was shut off and bias was manipulated instead via the learning bias parameter *G*. Since this bias only applies to beans where actual feedback disconfirms prior expectancies or 'prejudices', its effect is to slow down the rate at which the net modifies connection weights to reduce the discrepancy Δ between the judgment of a bean in the identified region and its correct 'target' or 'training' value. This is because the prior expectancies are *always* inconsistent with the actual feedback relating to the targeted region. Thus, in the negative bias condition, if the net approaches a bean in region 1, it will experience positive feedback (that the bean is good) and this should lead to connection weight modification ('learning') so that subsequent evaluations of the same bean are somewhat more positive. However, because of the negative learning bias operating in this condition, less 'reliance' will be placed on this corrective feedback and the rate at which the connection weights are modified (to produce a more positive evaluation) will be slowed down. Likewise, in the positive bias condition, if the net approaches a bad bean in region 6, the feedback should lead to subsequent evaluations of the bean becoming somewhat more negative. However, the positive learning bias means that this change takes place at a slower rate, with the predicted consequence that bad beans should continue to be approached for somewhat longer than if the positive learning bias was inoperative.

Results

The 100 runs of the simulation were completed without the need for any restarts due to the net prematurely running out of energy. Identification of bad beans in region 6 was again perfect by the end of training under all conditions. Identification of the good beans in region 1 was suppressed to 50% under negative learning bias but was perfect under neutral bias ($t = 3.00$, $p < .01$) and also when positive bias was applied to the opposite region of good beans. The critical learning bias × region interaction is significant ($F(2, 27) = 5.56$, $p < .01$. Thus, even if the net was (positively) biased so as to be slower to recognize region 6 as bad, it nonetheless came round to forming negative presentations of these beans on the basis of the feedback it received from approaching them. However, if the net was (negatively) biased so as to be slower to recognize region 1 as good, this was sufficient to inhibit approach (and hence opportunity for corrective feedback) to the point that these good beans often continued to be misclassified as bad.

STUDY 3: COMBINED ACTION AND LEARNING BIASES

Although our first two studies were designed to consider two distinct biases (through differential contact and resistance to accepting disconfirming evidence), these processes could be complementary rather than mutually exclusive in real-world

contexts such as intergroup contact. Our third study was therefore designed to look at what might happen if both sources of bias were combined, that is if both action and learning were biased in either the positive or negative direction, as compared with a neutral condition in which neither action nor learning was biased. To this end, the design was again a 3 × 2 (bias × region) factorial with repeated measures on the last factor, involving comparisons between negative, neutral, and positive bias. Thus, $S = 0.5$ and $G = 1.5$, except in the neutral conditions, where neither parameter was applied (i.e. $S = G = 1$). We anticipated that the combined effect of these two sources of bias would be at least as strong as that observed in either of the first two studies when the biases were manipulated separately.

Results

Only one restart was required because of an early loss of energy by the net. Identification of bad beans in region 6 is near perfect (90% under positive bias, 100% in the other two conditions). As in studies 1 and 2, bad beans are still avoided under positive bias, since such bias, by increasing the chance of these beans being sampled during training, leads to the net learning that they are bad. More striking is the failure of the net to 'unlearn' the negative prejudice attached the good beans in region 1.Even after 5,000 epochs of training, the hit-rate was only 10%, compared with 90% under neutral bias ($t = 5.66$, $p < .001$) and 100% when positive bias was applied to region 6.The bias × region interaction was significant both at 5,000 epochs, $F(2, 27) = 28.00$, $p < .001$.

To summarize these main findings, Figure 2 presents the means for bad beans (region 6) under positive bias and good beans (region 1) under negative bias in each of the above

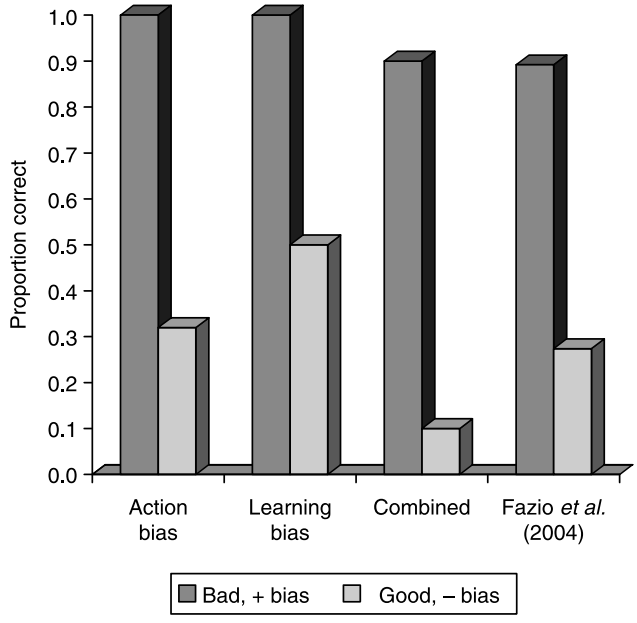

Figure 2. Mean proportions of targeted 'bad beans' correctly classified despite positive bias, and 'good beans' correctly classified despite negative bias, in Studies 1, 2, and 3 (simulations) and in human data from Fazio *et al.* (2004), Experiment 5.

three studies, along with comparable human data from Fazio *et al.* (2004, Experiment 5), that is the means for 'told positive, actually negative' versus 'told negative, actually positive'). As can be seen, the combined bias conditions of Study 3 are stronger (with respect to the suppression of identification of good beans) than either the action or learning biases considered separately, and essentially produce a more exaggerated pattern of that found in the human data.

DISCUSSION

These simulations were designed to help elucidate the processes by which false beliefs, expectancies and prejudices, may resist correction. Crucial to this enterprise is a principle that arises from the dynamic interaction between action and feedback in all forms of experiential learning, and lies at the heart of the BeanFest paradigm as well as many real-life social encounters. If expectancies about the valence of objects both guide whether such objects are approached or avoided *and* are shaped by feedback contingent on the learner's own behaviour, this sets up a fundamental asymmetry between the acquisition and maintenance of positive as opposed to negative expectancies and attitudes.

Taken as a group, these three studies strongly support our hypothesis that, under contingent feedback, false positive expectancies about the valence of a subset of objects would tend to be corrected by experience, but false negative expectancies would be much less likely to be corrected. In other words, if learners expect specific objects to be good, and therefore approach them, but then discover they are bad, they will correct their expectancies. However, if they expect specific objects to be bad, they will tend to avoid them, and so receive less feedback that could correct such expectancies if the objects are actually good. We sought to distinguish between a bias towards approach or avoidance at the level of action selection from an expectancy-confirmation bias at the level of how feedback was processed by the network. We stress that we made no assumption that the action and learning biases were manipulated at comparable strength (since there is no principled way of scaling this). Because the chosen strength of the two parameters was arbitrary, it is misleading to draw inferences about their comparable effectiveness when considered singly. Likewise, not too much reliance should be placed on the visual similarity between the data from any one of the three studies and the human data from Fazio *et al.* (2004) illustrated in Figure 2, since, here too, the strength of any effect may reflect the strength of the experimental manipulation. Nonetheless, our data suggest that the combined effect of the two forms of bias (Study 3) is stronger than that of either considered individually. Hence the two forms of bias should be regarded as reflecting complementary, rather than competing, processes.

Taking these findings as illustrative of processes underlying prejudice in humans, the negative action bias can be regarded as analogous to the situation where an individual avoids contact with members of a disliked group, and therefore never discovers their good qualities but continues to hold untested negative stereotypes about them. The process simulated by the learning bias is more subtle, and is best thought of as analogous to individuals discounting or distrusting information that contradicts their prior expectancies. There are reasons to suppose that this is a powerful process underlying the maintenance of prejudicial beliefs. Although prejudices can be protected, up to a point, by avoidance of contact with members of an out-group, contact may still happen

for reasons beyond the individual's personal choice or control (In our simulations, this is operationalised through the stochastic element in the action selection mechanism). Without any other form of 'protection', if such contact provides evidence that contradicts the initial prejudice, that prejudice will have to be weakened or abandoned.

Unfortunately, individuals who are strongly committed to their prior beliefs and prejudices are often not so easily swayed by evidence, but may employ a variety of cognitive devices to reinterpret such evidence in a way that leaves the central core of their belief system intact. For example, in their study of the authoritarian personality, Adorno, Frenkel-Brunswik, Levinson, and Sanford (1950), report an interviewee whose brightest class-mate at school was an African-American girl, but this is attributed to an attempt to overcompensate for inherent inferiority. In their words (p. 616):

> To the prejudiced, the Negro is 'dull'; if he meets, however, one of outstanding achievement, it is supposed to be mere overcompensation, the exception that proves the rule. No matter what the individual is or does, he is condemned.

In a similar vein, Billig (1985, p. 94) discusses the difficulties faced by prejudiced individuals in maintaining inflexible stereotypes when confronted by disconfirming evidence, or 'exceptions' to their imagined 'rule':

> This implies that a certain inventiveness will be required to maintain categorical distinctions: it is precisely this sort of inventiveness that can lead the serious racist to formulate complex theories about hidden racial conspiracies and concealed qualities of blood, which 'prove' that under the skin all Aryans, blacks or whatever really do have 'essential' racial qualities.

In a similar vein, Macrae, Bodenhausen, Schloerscheidt, and Milne (1999) have argued that social perceivers who encounter information inconsistent with their stereotypes of specific social groups may attempt to engage in 'inconsistency resolution' and 'individuation' – processes that, nonetheless, require executive cognitive operations (i.e. deliberative thought) and so may be impeded by particular manipulations of cognitive load. The full complexity of such rationalization processes and their consequences for intergroup behaviour lies well beyond the aims or pretensions of the simulations reported here. Furthermore, we were concerned more narrowly with the association of valence with specific objects, rather than with the content of stereotypes as such. Nonetheless, connectionist simulation can make an important contribution to the understanding of the processes underlying stereotype formation and change precisely because we are dealing with learning in a highly abstract and minimalist context. By definition, connectionist networks are not conscious, or motivated by concerns with protecting their self-esteem or social identity. They are merely systems designed to form simplified representations of patterns of information and generate outputs appropriate to such inputs. The way they achieve this is through 'automatic' rules defining how associations are formed from experience with objects. In general terms, therefore, if patterns characteristic of human cognition and behaviour can be simulated in such artificial learning systems, this adds force to the argument that we may be dealing with very general, potentially 'automatic', processes of information-handling that do not *necessarily* require modulation or control by any higher-order conscious function.

Previous experimental (e.g. Johnston & Hewstone, 1992; Weber & Crocker, 1983) and connectionist (Queller & Smith, 2002; Smith & DeCoster, 1998) research on stereotype change typically involves participants being presented with information about the attributes of specific target individuals belonging to particular social groups.

These attributes may be selected as consistent or inconsistent with some preformed stereotype, and are presented in such a way that participants are able to form associations among these attributes and/or between sets of attributes and a signifier of group membership. Issues considered by such researchers, but not addressed specifically in our present work, include the factors that can lead to moderation of beliefs about a group as a whole ('bookkeeping', Rothbart, 1981), as opposed to the 'subtyping' of a group into more and less prototypical members (Brewer, Dull, & Lui, 1981; Taylor, 1981). Queller and Smith (2002) and Van Rooy, Vanhoomissen, Labiouse, and French (2003) have persuasively argued that both 'bookkeeping' and 'subtyping' phenomena may fall out of connectionist learning processes without needing to invoke more complex attributional mechanisms (e.g. Ybarra, 2002), and stereotype change in response to disconfirming information should be seen merely as a continuation of the same learning processes that can account for the acquisition of stereotypes in the first place (Deutsch & Fazio, 2008).

The fact that feedback was contingent on approach constitutes the principal distinctive feature of the simulations reported here and the human experiments on which they are based. Applying Hertwig, Barron, Weber, and Erev's (2004) terms, we are concerned with learning 'from experience' unlike much previous research involving learning 'by instruction'. Under the contingent feedback conditions of our paradigm, learners have to *do something* to gain information about the valence of attitude objects. They need to make choices under conditions of uncertainty, and if they choose incorrectly, this will be costly (and even doing nothing can be somewhat costly). We suggest that this corresponds – perhaps more closely – to the situation faced in the real world by learners of many species who have to navigate between safety and danger, to distinguish prey from predator, or friend from foe. Once learning is constrained by its dependence on the learner's own behaviour, it becomes easier to account, not just for *change* in expectancies, but also for why some kinds of expectancies *resist change* more than others. Because of this constraint, the resistance to change of negative prejudices can be explained in terms of learning processes that are at once simple and extremely general.

Acknowledgements

This research was supported by Grant R000223077 from the Economic and Social Research Council, United Kingdom to the first author and by Grant MH38832 from the National Institute of Mental Health to the third author.

References

Abelson, R. P. (1959). Modes of resolution of belief dilemmas. *Journal of Conflict Resolution*, *3*, 343–352.

Adorno, T. W., Frenkel-Brunswick, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper and Row.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison Wesley.

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, *126*, 925–945.

Billig, M. (1985). Prejudice, categorization and particularization: From a perceptual to a rhetorical approach. *European Journal of Social Psychology*, *15*, 79–103.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, *41*, 656–670.

Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, *35*, 867–881.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*, 20–33.

Deutsch, R., & Fazio, R. H. (2008). How subtyping shapes perceptions: Predictable exceptions to the rule reduce attention to stereotype-associated dimensions. *Journal of Experimental Social Psychology*, *44*, 1020–1034.

Eiser, J. R. (1994). *Attitudes, chaos and the connectionist mind*. Oxford: Blackwell.

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29*, 1221–1235.

Eiser, J. R., Stafford, T., & Fazio, R. H. (2008). Expectancy-confirmation in attitude learning: A connectionist account. *European Journal of Social Psychology*, *38*, 1023–1032.

Ellis, R. & Humphreys, G. (Eds.), (1999). *Connectionist psychology: A text with readings*. Hove: Psychology Press.

Fazio, R. H. (2001). On the automatic evaluation of associated evaluations: An overview. *Cognition and Emotion*, *15*, 115–141.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*, 293–311.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. New York: Free Press.

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: III. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*, 360–368.

Macrae, C. N., Bodenhausen, G. V., Schloerscheidt, A. M., & Milne, A. B. (1999). Tales of the unexpected: Executive function and person perception. *Journal of Personality and Social Psychology*, *76*, 200–213.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751–783.

Queller, S., & Smith, E. R. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology*, *82*, 300–313.

Read, S. J., & Miller, L. C. (1998). *Connectionist models of social reasoning and social behavior*. Mahwah, NJ: Erlbaum.

Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 145–182). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, *70*, 893–912.

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a connectionist network. *Journal of Personality and Social Psychology*, *74*, 21–35.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press.

Taylor, S. E. (1981). The interface of cognitive and social psychology. In J. Harvey (Ed.), *Cognition, social behavior, and the environment* (pp. 182–211). Hillsdale, NJ: Erlbaum.

Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, *9*, 231–274.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, *110*, 536–563.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*, 961–977.

Ybarra, O. (2002). Naive causal understanding of valence behaviors and its implications for social information processing. *Psychological Bulletin*, *128*, 421–441.

Received 27 March 2008; revised version received 7 August 2008