

ASSOCIATIVE AND CONNECTIONIST ACCOUNTS OF BIASED CONTINGENCY DETECTION IN HUMANS*

SERBAN C. MUSCA, MIGUEL A. VADILLO, FERNANDO BLANCO AND HELENA MATUTE

Laboratorio de Psicología del Aprendizaje, Universidad de Deusto, 24, Avenida de las Universidades, 48007 Bilbao, Spain

Associative models, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972), correctly predict how some experimental manipulations give rise to illusory correlations. However, they predict that outcome-density effects (and illusory correlations, in general) are a preasymptotic bias that vanishes as learning proceeds, and only predict positive illusory correlations. Behavioural data showing illusory correlations that persist after extensive training and showing persistent negative illusory correlations exist but have been considered as anomalies. We investigated what the simplest connectionist architecture should comprise in order to encompass these results. Though the phenomenon involves the acquisition of hetero-associative relationships, a simple hetero-associator did not suffice. An auto-hetero-associator was needed in order to simulate the behavioural data. This indicates that the structure of the inputs contributes to the outcome-density effect.

1. Introduction

Perceiving contingency between potential causes and outcomes is of crucial importance in order to understand, predict, anticipate and control our environment. However, there is little agreement on the mechanisms that underlie this ability, and the research on human contingency perception is in its flourishing years.

As with many other cognitive phenomena, one way to gain a better understanding of the ability under scrutiny is to find variables that affect it in a systematic way, be able to predict their influence, and finally comprehend why these variables do have an effect. Personality or mood variables (e.g. Alloy & Abramson, 1979), the valence of the outcome (e.g. Alloy & Abramson, 1979; Aeschleman, Rosen & Williams, 2003), the density of the cue/response (e.g.

* Support for this research was provided by Grant SEJ2007-63691/PSIC from the Spanish Government and Grant SEJ406 from Junta de Andalucía. Fernando Blanco was supported by a F.P.I. fellowship from Gobierno Vasco (Ref.: BFI04.484).

Allan & Jenkins, 1983; Matute, 1996), the density of the outcome (e.g. Alloy & Abramson, 1979; Allan & Jenkins, 1983; Matute, 1995) are all factors that have an influence on the ability to correctly perceive contingency in humans.

In the following, after describing the general methodology used in behavioural experiments that study contingency perception, we will focus on the influence that density of the outcome has on a judgment of contingency. We will present the widely accepted associative account of Rescorla and Wagner (1972) and also behavioural data that are not accounted for by this model. We will then present simulations conducted with two different neural network models designed to encompass those behavioural results not accounted for by the associative model, and the surprising results the simulations yielded.

1.1. Studies of Contingency Judgment

Experimental studies of contingency judgment in humans generally involve the use of a 2-phase task. During the first phase (the training), covariational information is given to the participants in successive trials. In each trial a cue (e.g. ingestion of strawberries) is either present or absent, and the outcome of interest (e.g. allergic reaction) either occurs or does not occur (cf. Table 1). Trials where both the cue and the outcome are present (*a* trials), trials where the cue is present and the outcome absent (*b* trials), trials where the cue is absent and the outcome present (*c* trials), and trials where both the cue and the outcome are absent (*d* trials) are presented in random order for a total of (*a+b+c+d*) trials.

Table 1. Trial types that make up the covariational information that is given to the participants in contingency judgment experiments.

		Outcome (allergic reaction)	
		present	absent
Cue (eaten strawberries)	present	a	b
	absent	c	d

In the subsequent test phase participants are to judge the degree of the causal relationship between the cue and the outcome (e.g. to what degree they think the ingestion of strawberries is the cause of the allergic reaction).

Of course, this is just a general outline, and many variants of the task exist. For instance, the subjective contingency can be assessed throughout the learning phase by presenting the participants with the cue and asking them to predict what the outcome would be before displaying the actual outcome. In another task that has been used extensively, during the training phase the cue

(present/absent) is replaced by the participant's response (i.e. response/no response).

Participants generally get things right, but under certain conditions participants' judgments diverge from the ideal judgment one is expected to give based on the objective covariational information presented during the training phase (López, Cobos, Caño, & Shanks, 1998). However, in order to observe this discrepancy, one has to dispose of a measure of the ideal judgment expected.

An objective measure has to take into account both the probability of a present outcome when the cue was present — that is $p(O|C)$ — and the probability of a present outcome when the cue was absent — that is $p(O|noC)$. Indeed, the fact that the outcome is present when the cue is present does not mean that the cue is the *cause* of the outcome if the outcome is present just as many times when the cue is not present. Based on this reasoning, the ΔP index was proposed by Jenkins & Ward (1965; see also Allan, 1980; Cheng & Novick, 1992) as a measure of contingency:

$$\Delta P = p(O|C) - p(O|noC) = a/(a+b) - c/(c+d) \quad (1)$$

This index has a value of 0 if the presence of the cue is not the cause of the presence of the outcome, that is, if the presence of the outcome is not contingent on the presence of the cue.

1.2. Illusory Correlation

As the ΔP formula hints, to one ΔP value may correspond many different distributions of trial types (see Table 1 for the four trial types), with different cue probability — *cue density*, $p(C) = (a+b)/(a+b+c+d)$ — and/or outcome probability — *outcome density*, $p(O) = (a+c)/(a+b+c+d)$.

As noted in the introduction, it is documented that these densities (among other variables) do bias the perceived contingency in the sense that they incorrectly affect participants' judgment of contingency. *Illusory correlation* refers to the phenomenon whereby in a noncontingent situation (i.e. a situation of stochastic independence between cue and outcome) participants incorrectly perceive a contingency between the cue and the outcome. The contingency between cue and outcome perceived by the participants is illusory because considering the covariational information supplied to the participants ΔP is nil.

In the following we will consider to a greater extent one of the possible causes of illusory correlation, the outcome density.

2. Outcome-density Effect

The outcome-density effect is an illusory correlation that has its roots in the probability of occurrence of the outcome (or *outcome density*). Though participants' judgments of contingency should not differ while the contingency is kept constant, it is documented that participants' judgments of contingency differ following the probability of the outcome, $p(O)$. This bias was called outcome-density effect (e.g. Alloy & Abramson, 1979; Allan & Jenkins, 1983; Matute, 1995). For instance, while ΔP is nil both for $a = 15, b = 5, c = 60, d = 20$ and for $a = 5, b = 15, c = 20, d = 60$, $p(O)$ is of 0.75 in the former case, and of 0.25 in the latter. In the example given here, were the participants to rate the contingency as higher when $p(O) = 0.75$ than when $p(O) = 0.25$ one would speak of an outcome-density effect.

2.1. An Associative Account: The Rescorla-Wagner Model

The Rescorla-Wagner model (hereafter RW) proposed by Rescorla & Wagner (1972) is one of the most widely used associative models when it comes to simulate how people learn to associate potential causes and effects (here, cue and outcome). Sutton & Barto (1981) have shown that it is formally equivalent to the delta rule (Widrow & Hoff, 1960) used to train two-layer distributed neural networks through a gradient descent learning procedure. In the RW model, the change (ΔV_C^n) in the strength of the association between a potential cue C and a potential outcome after each learning trial takes place according to the equation:

$$\Delta V_C^n = k \cdot (\lambda - \Delta V_C^{n-1}), \quad (2)$$

where k is a learning rate parameter that reflects the associability of the cue, α , and that of the outcome, β ($k = \alpha \cdot \beta$ in the original RW model); λ reflects the asymptote of the curve (which is assumed to be 1 in trials in which the outcome is present and 0 otherwise), and ΔV_C^{n-1} is the strength with which the outcome can be predicted by the sum of the strengths that all the cues that are present in the current trial had in trial $n-1$.

The RW model correctly predicts that outcome density manipulations give rise to illusory correlations. This is illustrated in Figure 1, by manipulating the outcome density in a case where the total covariational information corresponds to a noncontingent situation (i.e. $\Delta P = 0$). The parameter k was set to 0.3 for the cue and to 0.1 for the context. We run 3000 replications. As can be seen in Figure 1, the associative strength developed between the cue and the outcome,

which corresponds to an illusory correlation in this case (because ΔP is nil), is stronger and more long-lasting for the case when the outcome density is higher.

The RW model correctly predicts and simulates a large set of associative learning phenomena (for a review see Miller, Barnet, & Grahame, 1995; López et al., 1998). However, in the following we will focus on a set of data that are not accounted by this model, and see why the characteristics of the model make it impossible for it to simulate these behavioural data.

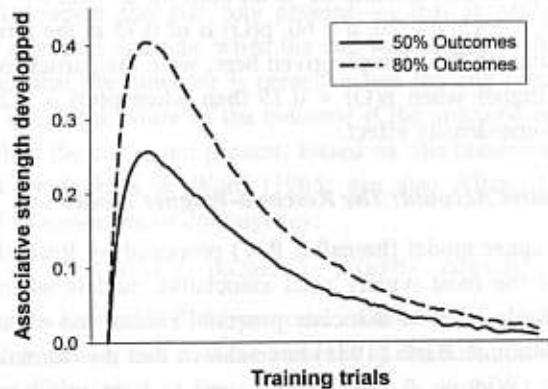


Figure 1. Illusory correlation (associative strength developed between the cue and the outcome) in the RW model in a noncontingent situation (see text for details).

The simulation presented above hints two characteristics of the RW model. One of them is that it predicts that outcome-density effects are a preasymptotic bias that should vanish as learning proceeds (e.g. López et al., 1998; Shanks, 1995). Indeed, from inspection of Figure 1 one can see that even when the outcome density is high, with enough training the associative strength developed between the cue and the outcome finally goes down to zero.

Another characteristic of the RW model is that the associative strength developed between the cue and the outcome is positive, so that with a low outcome density (e.g. 20%) the model will not yield a negative illusory correlation but a positive one.

While data contradicting these predictions of the model were at first scarce and were considered as anomalous, behavioural results at odd with these two RW model characteristics accumulated over the years. For instance, there is evidence that the outcome-density effect sometimes persists even with extensive training (Shanks, 1987) and that it does not disappear but, on the contrary, become stronger with more training trials (Allan, Siegel & Tangen, 2005). Also,

some experiments with noncontingent situations that comprise a condition of low outcome density yielded a negative illusory correlation (Shanks, 1985, 1987; Allan et al., 2005; Crump, Hannah, Allan, & Hord, 2007).

In view of these results that the RW model cannot simulate and account for, the two aforementioned characteristics of this model appear as limitations. In the following we consider another type of modelling, one using simple distributed artificial neural networks. Using this more powerful metaphor, we investigated what the simplest connectionist architecture should comprise in order to encompass these results.

2.2. Connectionist Simulations: What is the Minimal Model?

The delta rule (Widrow & Hoff, 1960) on top of being equivalent to the learning rule used by the RW model is also the ancestor of the generalized delta rule (Rumelhart, Hinton & Williams, 1986) that once discovered in the late 1980's gave rise to the connectionist revolution we all know of. The generalized delta rule allows training networks of more than two layers, some of which have a complex nonlinear behaviour. These are more powerful simulation tools than the RW model, but when using this class of models one has to keep the simulation model to a minimum and analyse the constraints and the degrees of freedom it allows for.

In accordance with this principle, and because of our previous work and inclinations (Musca, 2005; Musca & Vallabha, 2005) we chose to tackle the problem at hand with 3-layer distributed neural networks trained with a backpropagation learning algorithm that minimizes the cross-entropy cost function (Hinton, 1989). Because the problem involves the learning of cue-outcome pairs and the structure of the outputs is of interest (the outcome density is a property of the outputs), the minimal model to be used is a hetero-associator (Bishop, 1995). For both the hetero-associator initially considered and the augmented auto-hetero-associator used afterwards (see below), the learning rate was set to 0.1, the momentum to 0.7 and the activation of the bias cell to 1. For both types of architecture used, 50 replications were run, with matched connection weights between the two conditions that were contrasted (i.e. low vs. high outcome density).

2.2.1. Translating the Problem into "Neural Networks Language"

An important part of the modelling is the translation of the problem into neural networks language. This involves creating a training set, that is, choosing the input and output vectors and the way they are related one to another.

One element of importance is that because of its mode of functioning a neural network cannot learn at the same time (i.e. as part of the same problem) something and its contrary. In other words, two trials such as “cue present-outcome present” and “cue present-outcome absent” cannot coexist in a training base, unless they occur in a different context. This is a supposition that has to be done, but we think it is a sensible one. After all, you will burn your fingers when touching an oven or not depending on the context, that is whether it has been used and is hot or whether it has not been used for a long time, but you will not be able to tell whether your fingers will be burned or not when touching the oven if you do not dispose of the context information. When a hetero-associator neural network is trained, its task can be understood just as this: give the right output given the input at hand. So, if the input at hand gives different outputs depending on the context where it occurs, this context must be specified. With these considerations in mind, we decided that the training base will have as many different contexts as training trials (i.e. that no training exemplar had the same context as another training exemplar). While this is not the most economical solution it has the advantage of avoiding possible biases due to the sharing of context between trials.

The training base comprised 100 training exemplars. The input of each exemplar is a 102-component vector made of two parts. The first part is a 100-component context vector that contains only one 1 component and 99 0 components in such a way that the context vectors of all the training exemplars are orthogonal. The second part of the input vector is a 2-component vector that codes for the cue, with 1 0 being cue present and 0 1 being cue absent. The output vectors are 2-component vectors that code for the outcome, with 1 0 being outcome present and 0 1 being outcome absent.

The dependent variable that we used, which we call “contingency estimation” is computed according to the following reasoning. After training, the network is probed without any context and the activation of the first output node is recorded, first with the *cue present* input (i.e. 1 0) and then with the *cue absent* input (i.e. 0 1). The network’s *contingency estimation* is the difference in activation between the first recording and the second recording. Inspired by the ΔP index (see Equation 1), this index is computed as:

$$\text{Contingency estimation} = \text{activation (O|C)} - \text{activation (O|noC)} \quad (3)$$

The covariational information given to the networks corresponds to a noncontingent situation, with an outcome density that was varied with two

values, 40% (low) and 60% (high)^a. In terms of types of trials (see Table 1) the low outcome density condition corresponds to $a = 32$, $b = 48$, $c = 8$, $d = 12$, and the high outcome density condition corresponds to $a = 48$, $b = 32$, $c = 12$, $d = 8$.

2.2.2. Three-layer Hetero-associative Network

Starting with random connection weights — uniformly sampled between -0.5 and 0.5 — the 3-layer network with 102 input units, 10 hidden units and 2 output units was trained with the abovementioned parameters.

The dependent variable contingency estimation was computed at three points during training: when the root mean squared error on the training set (RMS Error) was of 0.1, of 0.01 and of 0.001, which corresponds roughly to 10, 20 and respectively 100 training epochs. As the results depicted in Figure 2 show, be it with little or extensive training the hetero-associative network failed to exhibit the expected outcome-density effect.

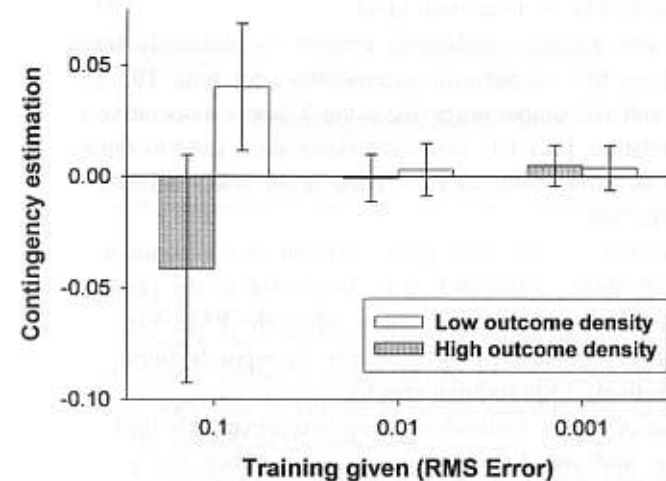


Figure 2. Results of the simulation with a hetero-associative network: the outcome-density effect is not simulated (whiskers represent .95 confidence interval limits).

The failure of the hetero-associator to produce the outcome-density effect is very surprising, because this kind of network builds its internal representations by taking into account the structure of the outputs, which is exactly what is manipulated when we use two different outcome densities. However this

^a We chose outcome density values close to 50% so as to avoid a ceiling effect in the simulations.

surprising pattern of results seems to be robust, so we had to understand why it occurs. Could it be that what is called outcome-density effect is in fact not an effect that comes only from the density of the outcome? Moreover, could it be that the density of the cue plays a role in what is called outcome-density effect?

2.2.3. Three-layer Auto-hetero-associative Network

In order to answer these questions we resorted to a slightly augmented architecture, an architecture that takes into account when building its internal representations both the structure of the outputs (in this it is a hetero-associator) and that of the inputs (in this it is an auto-associator). Thus it is an auto-hetero-associator that we used in this second neural network simulation.

An auto-hetero-associator is a 3-layer network very similar to a hetero-associator, but its output layer contains not only the hetero-associative units but also other units, corresponding to the input units. Its task is both to associate the current input with the current hetero-associative target and to re-create at the output layer the current input.

Starting with random connection weights — uniformly sampled between -0.5 and 0.5 — the 3-layer auto-hetero-associator with 102 input units, 10 hidden units and 104 output units (the same 2 hetero-associative units as in the previous simulation, plus 102 auto-associative units used to reproduce the 102 input units) was trained with the same parameters as the hetero-associator in the previous simulation.

The dependent variable contingency estimation was computed at five points during training, three corresponding to those used in the previous simulation, and 2 complementary intermediate ones: when the RMS Error was of 0.1, of 0.075, of 0.01, of 0.005 and of 0.001, which corresponds roughly to 10, 40, 200, 350 and respectively 1500 training epochs.

Inspection of Figure 3 allows noticing two remarkable findings. First of all, both positive and negative outcome-density effect are obtained. This is compatible with the behavioural results that exist in the literature where both positive and negative illusory correlations have been found when manipulating the outcome density.

Secondly, the effects do not appear immediately but only after quite a considerable amount of training. However they do not vanish but become stronger when more training is given. And, though this may be artefactual and should be investigated at more length, a negative outcome-density effect seems to be obtained easier, (i.e. before, with less training) than a positive one, a result that has been found in humans by Crump et al. (2007). The pattern of results

found in this simulation is compatible with the results of Shanks (1987) where a negative outcome-density effect was still present after very extensive training and with results of Allan et al. (2005) and Shanks (1985) showing that illusory correlation increases with training (but see López et al., 1998 for divergent results).

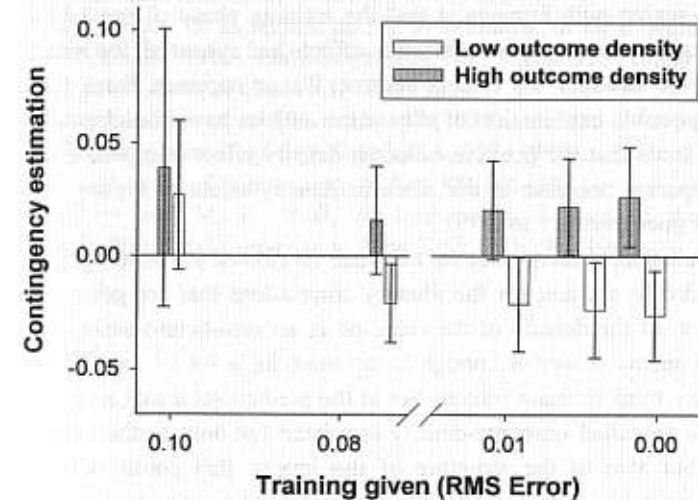


Figure 3. Results of the simulation with an auto-hetero-associative network: the outcome-density effect is simulated (whiskers represent .95 confidence interval limits). See text for details.

3. Conclusion

When trying to simulate those outcome-density effects found in humans the Rescorla-Wagner model cannot account for we had recourse to a distributed artificial neural network that is known to be sensitive to the structure of the outputs, that is, to the density of the outcome.

The simulation with this kind of neural network, a hetero-associator, failed to produce the expected outcome-density effects. This failure, in artificial neural networks terms clearly has only one implication: the so-called outcome-density effect is not an effect of the density of the outcome *per se*. Were it the case, the outcome-density effect would have been simulated with this class of networks.

In a second simulation we used an auto-hetero-associator, a type of network that takes into account both the structure of the outputs and of the inputs. With this distributed artificial neural network we were able to simulate the negative outcome-density effects that exist in the literature. Thus this model encompasses some important results that cannot be explained by the RW model. However,

one must keep in mind that the RW model can explain data in the literature that an auto-hetero-associator could not simulate without complementary suppositions.

Moreover, quite some training was needed to the auto-hetero-associator before the effects appeared. Thus one possible explanation of the fact that such results are scarce with humans is that the training phase of most behavioural experiments is never extensive. Once the effects had appeared, the more training was given the stronger the effects became. Taken together, these results may point at a possible explanation of why some authors have found with a limited number of trials that the positive outcome-density effect dropped down to zero (see the apparent decrease in the outcome-density effect in Figure 3 when the RMS Error goes from 0.1 to .075).

In conclusion, it seems that the minimal distributed artificial neural network that is needed to account for the illusory correlations that are generated by the manipulation of the density of the outcome is an auto-hetero-associator. While this model seems powerful enough to account for a lot of extant data in the literature, we think its main interest lies in the predictions it makes. For instance, whether the so-called outcome-density is related not only to the density of the outcomes but also to the structure of the inputs, this could be checked in simulation work that manipulates both cue-density and outcome-density.

References

1. S. R. Aeschleman, C. C. Rosen and M. R. Williams, *Beh. Proc.* **61**, 37 (2003).
2. L. G. Allan, *Bul. Psychon. Soc.* **15**, 147 (1980).
3. L. G. Allan and H. M. Jenkins, *Learn. and Motiv.* **14**, 381 (1983).
4. L. G. Allan, S. Siegel and J. M. Tangen, *Learn. & Behav.* **33**, 250 (2005).
5. L. B. Alloy and L. Y. Abramson, *J. of Exp. Psych.: Gen.* **108**, 441 (1979).
6. C. M. Bishop, (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.
7. P. W. Cheng and L. R. Novick, *Psych. Rev.* **99**, 365 (1992).
8. M. J. C. Crump, S. D. Hannah, L. G. Allan and L. K. Hord, *Quart. J. of Exp. Psych.* **60**, 753 (2007).
9. G. E. Hinton, *Artif. Intell.* **40**, 185 (1989).
10. H. M. Jenkins and W. C. Ward, *Psych. Monograph.* **79**, 1 (1965).
11. F. J. López, P. L. Cobos, A. Caño and D. R. Shanks, *In* M. Oaksford & N. Chater (Eds.) Oxford: Oxford University Press, 314 (1998).
12. H. Matute, *Quart. J. of Exp. Psych.* **48B**, 142 (1995).
13. H. Matute, *Psych. Sci.* **7**, 289 (1996).
14. R. R. Miller, R. C. Barnet and N. J. Grahame, *Psych. Bul.*, **117**, 363 (1995).
15. S. C. Musca, *In* A. Cangelosi, G. Bugmann & R. Borisyuk (Eds.), Singapore: World Scientific, 367 (2005).
16. S. C. Musca and G. Vallabha, *In* B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.) Mahwah, NJ: Lawrence Erlbaum Associates, 1582 (2005).
17. R. A. Rescorla and A. R. Wagner, *In* A. H. Black & W. F. Prokasy (Eds.), New York: Appelton-Century-Crofts, 64 (1972).
18. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *In* D. E. Rumelhart and J. L. McClelland (Eds.), Cambridge, MA: MIT Press, 318 (1986).
19. D. R. Shanks, *Mem. & Cogn.* **13**, 158 (1985).
20. D. R. Shanks, *Learn. and Motiv.* **18**, 147 (1987).
21. D. R. Shanks, *Quart. J. of Exp. Psych.* **48A**, 257 (1995).
22. R. S. Sutton and A. G. Barto, *Psych. Rev.* **88**, 135 (1981).
23. G. Widrow and M. E. Hoff, *In* Convention Record of the Western Electronic Show and Convention, New York: IRE, 96 (1960).