Inferring Causality and Making Predictions. Some Misconceptions in the Animal and Human Learning Literature

by Helena Matute, Miguel A. Vadillo

Researchers may disagree in their theories of human and animal learning. But all of them agree at least in one thing: Learning is an adaptive tool that improves and organism’s ability to survive in a changing environment. The process of learning provides an animal with (either explicit or implicit) knowledge that captures, among many other things, the statistical relationships between different significant events that occur in the environment. This process takes place, for example, in Pavlovian conditioning experiments, in which humans and other animals learn the statistical relationship between a conditioned stimulus (CS) and an unconditioned stimulus (US) and adapt their response to the CS as a result of this experience. A similar process can be observed in experiments in which human subjects have to learn the relationship between different cues and outcomes and are asked to rate the perceived strength of those relationships.

However, as we have just stated, there is little consensus regarding the mechanisms by which organisms acquire and use this information. And, what is more important in our opinion, there are some conceptual contradictions in the terms used by theorists that will necessarily make the consensus more and more unlikely. As we will argue below, many researchers use common concepts related to associative learning inconsistently. Specifically, some authors implicitly assume that responding to the predictive value of a cue is the same as predicting an outcome, and a similar lack of clear-cut definitions is also evident when it is assumed that animals learn predictive relations between events or that they learn causal relations. Most theories of human and animal learning use terms such as prediction, predictive value and causality as if they were synonymous.

**Normative analyses and experiments with humans**

Regardless of whether people and animals ordinarily make this distinction in their daily lives, making a prediction is clearly different from assessing the predictive value of a cue. Imagine for example that a study shows that people living in your country have a probability of 10% of suffering skin cancer. Should you protect your skin? Of course you should, because there is a moderate likelihood that you will suffer skin cancer otherwise. Moreover, you should do so regardless of the probability of suffering cancer in other countries. It does not matter whether people living abroad suffer cancer with a probability of 0%, 10%, or 20%. You should always be equally careful because you are likely to be affected by this disease. Your prediction of the likelihood of suffering skin cancer (and your preparatory behavior to avoid it) should be unaffected by what happens in circumstances different from yours.

This knowledge of what happens in other countries is, however, extremely important if you have to assess whether or not living in your country is a good predictor of the likelihood of suffering skin cancer. If the probability of suffering cancer is 0.10 both in your country and in other countries, you cannot say that living in your country is a good predictor of developing skin cancer. Knowing that one person lives in your country does not help you decide whether he or she will be more likely to suffer cancer. Therefore, one thing is to make a prediction (how likely it is to suffer cancer) and a different one to assess the predictive value of a cue (whether or not living in your country increases the odds of being affected by skin cancer). Whereas predicting an outcome in a given situation requires taking into account only the probability of that outcome given the cues that define such situation [i.e., p(outcome|cue)],
assessing the predictive value of those cues requires taking into account whether the probability of the outcome is greater or smaller in the presence than in the absence of those cues. In other words, in order to assess the predictive value of a cue, you need to know whether the probability of the outcome is different in the presence than in the absence of that cue. This predictiveness or cue-outcome contingency is usually measured by the statistical index $\Delta p$, which is equal to the probability of the outcome given the cue minus the probability of the outcome in the absence of the cue \[i.e., \Delta p = p(\text{outcome}|\text{cue}) - p(\text{outcome}|\text{no cue})\].

Data gathered in our laboratory shows that the distinction between making predictions and assessing predictiveness and causal relations is a distinction that people draw spontaneously, and not a merely theoretical distinction that people are potentially able to understand but that they don’t use in their daily lives (Matute, Vegas, & De Marez, 2002; Vadillo, Miller, & Matute, 2005). For example, if people are told that 50% of the patients taking a medicine develop and allergic reaction and they are asked to assess how likely it is that a new, unknown patient taking the medicine will suffer the allergy, their response is close to the objective 50% regardless of the probability of developing the allergy in patients not taking the medicine. However, if they are asked to say whether they think that taking the medicine is a good predictor of developing the allergy, then they take into account what happens with patients not taking the medicine.

Similarly, there is a difference between assessing whether a cue is a good predictor of an outcome and assessing whether a cue is a cause of the outcome. This difference can be easily understood with the aid of a real-world example related to the Simpson’s paradox. During the 70’s it was discovered that women applying to study in Berkeley University were more likely to be rejected than men. Although this evidence was taken as evidence of discrimination against female applicants, a closer look at the data showed that women were more likely to be rejected because they tended to apply for more selective programs with higher rejection rates. In this example, knowing that a woman has applied for Berkeley is a datum that provides us with important predictive information: If we know that a given woman has submitted an application to Berkeley, then we know that the probability that she will be rejected is higher than it would have been if the applicant had been a man. However, there is no causal relation between being a woman and being rejected. The reason why sex is a good predictor of rejection is that sex covaries with a factor that causes rejection. Thus, in order to infer a causal relation it is not enough to pay attention to the predictive value of the cue, it is also necessary to check that there are no confounding variables that could artificially increase the covariation between the cue and the outcome (see Cheng, 1997; Cheng & Novick, 1992).

In spite of these normative and descriptive differences between predicting, assessing predictive value and estimating the strength of causal relations, researchers often use these concepts, either explicitly or implicitly, as synonymous. For example, researchers studying causal learning have sometimes used in their experiments instructions or test questions suggesting scenarios where the predictive value of the cues is more important than their causal status. Similarly, theories proposed to account for causal learning have been assumed to be adequate to account for participants predictions and for judgments of cue-outcome predictiveness.

**Predictions and predictive value in animal conditioning experiments**

So far, we have shown that making a prediction, assessing the predictive value of a cue and assessing the strength of a causal relation are different things, in spite of what learning theories posit. And that there is also evidence showing that people make these distinctions spontaneously. But the next question we could ask is: Is this distinction relevant for animal learning researchers? What do animals do when they give a conditioned response? Are they predicting the unconditioned stimulus? Or are they responding to the value of the conditioned
stimulus as a predictor of the unconditioned stimulus? These two ideas are often confounded in the animal literature.

Many published reports on Pavlovian conditioning start their introductions by remarking how associative learning allows animals to prepare for future events adaptively. From this point of view, one would expect animals in a classical conditioning experiment to be preparing themselves for the occurrence of the US after they experience the CS. Therefore, it is predicting the US what should be important for them. However, there is some evidence inconsistent with this perspective.

Imagine that a rat receives a footshock after a light in 50% of the trials in which the light is presented. In principle the rat should be moderately afraid after the presentation of the light, because this light is followed by the aversive US in half of the occasions. Moreover, if the rat were simply predicting the US, then, as we have discussed above, the probability of the US in the absence of the light should be completely irrelevant. If the footshock follows the light with a probability of 0.50, then the rat should be afraid, no matter whether the footshock never occurs in the absence of the light or whether it occurs constantly in the absence of the light. In other words, when predicting whether a US will follow a CS, the predictive value of the CS should not be important or, at least, not as important as the probability of the US after the CS.

Famous experiments performed by Rescorla during the 60’s seem to be inconsistent with this perspective. Rescorla (1968) found that for a given probability of the US given the CS, the conditioned response (CR) was negatively correlated with the probability of the US in the absence of the CS. That is, a rat shows stronger fear to a light that is followed by a US in 50% of the trials if the US never occurs in the absence of the light than if it occurs in the absence of the light with a probability of 0.50. This seems to indicate that an animal’s CR does not reflect a prediction of the US. The animal, on the contrary, seems to be assessing the CS’s predictiveness or predictive value.

The Rescorla-Wagner model of animal conditioning: Predictive value or prediction?

A few years after those experiments were published, Rescorla and Wagner (1972) proposed their famous model of animal conditioning which explained the sensitivity to the CS-US contingency. According to this model, the strengthening of the CS-US association in a given trial depends on the extent to which the US is unpredicted after the presentation of the CS. In each trial, the change in the strength of the association, $\Delta V$, is given by the equation

$$ \Delta V = \alpha \cdot \hat{\alpha} \cdot (\hat{e} - V_T) $$

where $\alpha$ and $\hat{\alpha}$ are learning rate parameters dependent on the salience of the CS and the US respectively, $\hat{e}$ is the maximum associative strength that the US can support, and $V_T$ is the associative strength of all the stimuli (CS, experimental context, etc.) that are present in that trial. The term $(\hat{e} - V_T)$ measures the degree of surprise produced by the presentation of the US: The difference between the US that is actually presented ($\hat{e}$) and the US expected on the basis of the stimuli presented in that trial ($V_T$).

Apparently, this model was able to explain Rescorla’s (1968) results in a quite simple manner. According to the model, when the US occurs frequently in a given context, both in the presence and in the absence of the CS, the experimental context becomes strongly associated with the US given that there are many context-US pairings. Because of this strong context-US association, the animal is able to predict the US in that context and therefore the US is no longer surprising. In other words, the term $(\hat{e} - V_T)$ gradually approaches zero. Thus,
eventually, the aforementioned equation will yield a value close to zero when the CS is presented, which means that the animal will not develop a strong CS-US association. This model predicts, therefore, that a low CS-US contingency should result in only weak conditioning to the CS.

However, the model also predicts that there should be a strong CR to the context in these situations. A strong context-US association would allow an animal to predict the US in that context. According to the Rescorla-Wagner model, if a light is followed by a footshock in 50% of the trials, and this footshock is also present with a probability of 0.50 in the absence of the light, the conditioned animal should be afraid at all times while being in that context because of fear being conditioned to the context. In other words, if an animal exposed to a null CS-US contingency is tested in the same context were it was conditioned, its total level of CR should not reflect CS-US predictiveness, but a prediction of the US in that situation. As we have shown, this is contrary to the pattern of results reported by Rescorla (1968), which showed an absence of fear in that situation.

In order to make sure that these are the actual predictions of the model, we have conducted some simulations of it. Below we show the results of some of them. These simulations illustrate how this model would expect animals to make accurate predictions of the USs if they are tested in the context where they were originally trained. Figure 1 shows the associative strengths that a CS would accrue under different CS-US contingencies. As there can be seen, the associative strength of the CS is asymptotically equal to the programmed CS-US contingency[1] and, therefore, this is not the information an animal should use when its goal is to predict the US.

![Figure 1. Simulation of the Rescorla-Wagner model showing the associative strength of the CS under several CS-US contingencies. The A series show the simulation in a conditioning situation where p(US|CS) = 1.00, p(US|~CS) = 0.50, \( \hat{\alpha}p = 0.50 \). In the B condition p(US|CS) = 1.00, p(US|~CS) = 1.00, \( \hat{\alpha}p = 0.00 \). In C, p(US|CS) = 0.50, p(US|~CS) = 0.00, \( \hat{\alpha}p = 0.50 \). In D, p(US|CS) = 0.50, p(US|~CS) = 0.50, \( \hat{\alpha}p = 0.00 \). Learning rate parameters were assigned.](http://www.interdisciplines.org/causality/papers/16/printable/paper)
the following values: ā_Cue = 0.8, ā_Context = 0.5, ā_Outcome = 0.6, and ā_NoOutcome = 0.6. For each condition, 10,000 iterations with randomized trial orders were performed.

Figure 2 shows the sum of the associative strength of the CS and the associative strength of the context. As can be seen, this sum of associative strengths always tends to be equal to the probability of the US given the CS \( p(US|CS) \), regardless of cue-outcome contingency.

Figure 2. Simulation of the Rescorla-Wagner model showing the sum of the associative strength of the CS and the associative strength of the context under several CS-US contingencies. As in Figure 1, the A series show the simulation in a conditioning situation where \( p(US|CS) = 1.00, p(US|\neg CS) = 0.50, \Delta p = 0.50 \). In the B condition \( p(US|CS) = 1.00, p(US|\neg CS) = 1.00, \Delta p = 0.00 \). In C, \( p(US|CS) = 0.50, p(US|\neg CS) = 0.00, \Delta p = 0.50 \). In D, \( p(US|CS) = 0.50, p(US|\neg CS) = 0.50, \Delta p = 0.00 \). Learning rate parameters and number of iterations were set to the same values as in the simulations reported in Figure 1.

According to the Rescorla-Wagner model, the CR in a given situation depends on the associative strength of all the stimuli present in that situation (including the context). Therefore, this model predicts that the total strength of the CR when the CS is presented in the training context should be dependent on the probability of the US given the CS. The model, therefore, expects animals to give responses that accurately prepares them for the potential subsequent US.

**Resolving the paradox**

With the preceding paragraphs we are not trying to conclude that the Rescorla-Wagner model is wrong, but simply that there is an obvious contradiction between the experiments performed by Rescorla (1968) and the Rescorla-Wagner model, and that this contradiction has not been made explicit in the animal or the human learning literature, among other things, because researchers have not drawn an appropriate distinction between predicting a US and assessing the CS-US predictiveness.
The origin of this contradiction might perhaps be found in the measure of CR that Rescorla (1968) used in his experiments. In experiments were aversive stimuli are used as USs, researchers often measure the animal’s CR by using a conditioned suppression paradigm. In this preparation the animal is first taught to press a bar to obtain food pellets. Then the animal is exposed to several pairings of the CS and the aversive US. Finally, in the test phase the animal is again allowed to bar-press for food and after it has spent some time pressing the bar, the CS is presented. If the animal is scared of the CS, then it will probably freeze and give fewer bar-press responses during the CS than in the immediately preceding pre-CS interval. Researchers can thus calculate a suppression ratio as a measure of the amount of fear the animal is showing. What the suppression ratio measures is to what extent the number of bar-press responses is lower during the CS than during the pre-CS period. Thus, the suppression ratio does not provide a direct and absolute measure of how scared the rat is during the CS. It is only measuring whether the rat is more or less scared when the CS is presented than it was previously in that context when the CS was not present. Imagine, for example, that the animal had been receiving footshocks both in the presence and in the absence of the CS and that there is a null CS-US contingency. In this situation, the rat might be equally scared and freezing both when the CS is off and when it is on. If the rat is equally scared during the pre-CS and the CS periods, then the suppression ratio would yield a null CR. However, this would not mean that the animal was not predicting the US, but simply that it was not predicting it more strongly during the CS than before the CS.

Therefore, the contradiction between the Rescorla-Wagner model and Rescorla’s (1968) experiments could be due to the dependent variable used by Rescorla (1968). In order to decide to what extent Rescorla’s (1968) experiments are actually contradictory with the predictions of the Rescorla-Wagner model one would need a measure of CR that directly assessed the intensity of the CR without making reference to the pre-CS level of responding.

Concluding comments

With the preceding paragraphs we have tried to show how the lack of distinction between concepts such as prediction and assessment of predictive value introduced confusion in the study of animal conditioning and human learning. Classical conditioning was supposed to provide animals with a means to successfully predict significant events. Rescorla’s (1968) experiments showed that animals were actually assessing CS-US predictiveness. This meant that they were not predicting the US (at least not optimally), but this implication went unnoticed. Moreover, when the Rescorla-Wagner model was proposed, it was thought that it provided an accurate explanation for the sensitivity of CR to the CS-US predictiveness (or contingency). But the model was actually predicting that animals tested in the same context where they had been conditioned should show conditioned responses that reflected an accurate prediction of the US; not CS-US predictiveness. Again, the lack of a clear conceptual distinction between predicting and assessing predictive value (or predictiveness) obscured this feature of the Rescorla-Wagner model.

It is unfortunate that after so many years since the Rescorla-Wagner model was published and after so many experiments testing the model there is still no empirical evidence with animals that would help resolve this paradox. However, as we have shown in the introduction of this paper, there is a growing body of literature showing that at least humans are able to flexibly make predictions, assess predictiveness or assess causal value as a function of what they believe will be more adequate at each time. A closer relationship between researchers investigating animal conditioning and those investigating human learning might help us detect and solve some of these conceptual problems in the future.

References


[1] This only holds if certain parameter values are used. If $\alpha$ is allowed to have different values when the US is present and when it is absent, then the associative strength of the CS would no longer be equal to $\bar{A}p$. However, it would still be dependent on this statistical index (i.e., a greater $\bar{A}p$ would give rise to greater associative strength).

[2] Again, this only holds if $\alpha$ is assumed to have the same value in all trials.