A Bio-Logical Theory of Animal Learning

David Guez
Faculty of Applied Science
The University of Canberra
Canberra, Australia
david.guez@canberra.edu.au
david.guez@mac.com

Abstract
This article provides the foundation for a new predictive theory of animal learning that is based upon a simple logical model. The knowledge of experimental subjects at a given time is described using logical equations. These logical equations are then used to predict a subject's response when presented with a known or a previously unknown situation. This new theory successfully anticipates phenomena that existing theories predict, as well as phenomena that they cannot. It provides a theoretical account for phenomena that are beyond the domain of existing models, such as extinction and the detection of novelty, from which "external inhibition" can be explained. Examples of the methods applied to make predictions are given using previously published results. The present theory proposes a new way to envision the minimal functions of the nervous system, and provides possible new insights into the way that brains ultimately create and use knowledge about the world.

Keywords
associative learning, extinction, habituation, latent inhibition, novelty detection, operant conditioning, Pavlovian conditioning, superconditioning

October 28, 2008; revised and accepted August 13, 2009
Biological Theory 4(2) 2009, 148–158.

Understanding the phenomena of how a brain acquires, orders, and applies knowledge of the world, or learns, remains a challenge. The development of learning theories is regarded as important for advancing these areas, as they provide a framework within which new hypothesis can be formulated and tested. Two main categories exist at present (Wagner 2003): the associative theories of learning, consisting themselves of the elemental theories of learning (e.g., Atkinson and Estes 1963; Rescorla and Wagner 1972) and the configural theories of Pavlovian conditioning (e.g., Pearce 1987; for review see Pearce and Bouton 2001), and the nonassociative theories of learning (e.g., Gallistel 1990). However, despite their impressive predictive abilities, problems still exist (e.g., Miller et al. 1995; Goddard 2003; Haselgrove et al. 2004), prompting for their continued improvement and the proposal of new theories that can successfully predict phenomena such as overexpectation (Lattal and Nakajima 1998), superconditioning (Rescorla 2003a), and external inhibition, as well as currently unpredicted phenomena such as extinction renewal (Pearce and Bouton 2001) and novelty detection.

Here the foundation for a new theory of animal learning is presented within the associative framework. This new theory is based upon logical rather than mathematical formalism, is not limited to Pavlovian conditioning, and excludes the use of free parameters. The primary aim of this article is to provide an introduction to this new theory of learning. While an in-depth literature review is beyond the scope of this current work, key experimental results have been used to measure its success and to illustrate how this new theory expands the predictive boundaries of existing learning theories.

The capacity for a species to predict environmental events can be considered a tremendous evolutive advantage. Organisms with a nervous system can construct a dynamic representation of reality ("Umwelt"; Von Uexküll 1965) by way of their sensory systems, which in turn enables them to predict the world around them, rightfully or wrongfully. This capacity to predict is dependent on learning and memory. It could therefore be said that the function of learning and memory is to generate predictive templates that enable an animal to predict future events based on its present and past experiences. To predict is equivalent to using a conditional proposition: if A then B. In consequence, if a minimalist approach is chosen, learning must enable the creation of causality-like links using conditional propositions that can be called "rules." If these rules are considered general by default, learning must enable the alteration of the field of application of the rule in order to restrict the domain of application of the causal link. The two processes necessary for learning are therefore of an associative type: The first associates a stimulus with a consequence, and the second restricts the field of application of the causal link by associating the exception to the rule with a particular situation (i.e., habituation/extinction). Thus, within the present framework, learning is always considered to be of an associative nature.

Interestingly, there are analogues to the logical conditional proposition existing within natural systems. One simple example of this is two neurons connected together. If we call (A) the presynaptic neuron and (B) the postsynaptic neuron, when (A) is activated (i.e., when it produces an action potential) the postsynaptic neuron will in turn be activated (assuming that the synaptic strength is strong enough). However, the activation of (B) does not have a bearing on the activation of (A), i.e., the activation of (B) will not trigger an action potential in (A). Therefore, in a logical sense, it can be written that if (A) then (B). However, when (B) is active at the same time as (A), long-term potentiation of the synapses occurs (Bliss and Gardner-Medwin 1973; Bliss and Lomo 1973). The repeated occurrence of such an event strengthens the synapse between (A) and (B) in a manner consistent with "Hebb's law" (Hebb 1949). Such a phenomenon could be interpreted as the building (learning) of a new conditional proposition or "rule." This biological analogue to the logical conditional proposition suggests the possibility of a learning theory that is based upon logical representations of the world, and therefore can be successfully translated into biological terms.

Logical Formalism

In the following discussion, the conventions of Boolean logic (Boole 2003) are used. To avoid confusion "$+$" indicates arithmetic addition, "$\dot{+}$" indicates the logical operator inclusive "or," "$\rightarrow$" indicates the logical operator "conditional proposition" (also called "conditional statement" or "material implication"), and "$\Longleftrightarrow$" indicates logical equivalence. Note the importance of distinguishing the *logical implication*, which is a tautology (a logical proposition that is always true) and a rule of inference, from the *conditional proposition*, which is not. For a list of the logical operators used herein and their truth table, please see "Conventions" (Box 1).

Assume that a particular event is to be predicted based upon the past experiences of a subject. A logical equation is one way of representing this event. This equation would be based upon the predictive stimuli that are already available to a subject, known as predictive events. As this equation would define what the subject knows of a particular event at a given time, it can therefore be used to predict a subject's response to a new situation that involves some of this past knowledge.

In order for a subject to predict future outcomes, it is necessary for it to assume that its knowledge at the time is true, consciously or not. For example, if it was known by a subject that the event A is conditional upon events B or C occurring ($B \rightarrow A$ and $C \rightarrow A$), in order to predict the occurrence of event A in function of the events B and C, the subject would need to postulate that ($B \rightarrow A$ and $C \rightarrow A$) is true. It can

Box 1. Conventions.

1. "+," **inclusive OR**, $A + B$ is read as A **OR** B. The truth table for $A \dot{+} B$ (A OR B) is given below, by convention true = 1 and false = 0.

A	B	$A \dot{+} B$
0	0	0
0	1	1
1	0	1
1	1	1

2. "×," **exclusive AND**, $A \times B$ or AB is read as A **AND** B. The truth table for AB (A AND B) is given below.

A	B	AB
0	0	0
0	1	0
1	0	0
1	1	1

3. $\bar{A}$ is the **negation of *A***, if A is true then $\bar{A}$ is false, if A is false then $\bar{A}$ is true. $\bar{A} = 1 - A$
4. "→" is the **conditional proposition**, $A \rightarrow B$ is read **if** A **then** B. The truth table for $A \rightarrow B$ (if A then B) is given below.

A	B	$A \rightarrow B$	$\bar{A} \dot{+} B$
0	0	1	1
0	1	1	1
1	0	0	0
1	1	1	1

Note that $A \rightarrow B$ is equivalent to $\bar{A} \dot{+} B$

therefore be written that

$$(B \rightarrow A)(C \rightarrow A) = 1 \Longleftrightarrow (\bar{B} \dot{+} A)(\bar{C} \dot{+} A) = 1$$
$$\Longleftrightarrow \bar{B}\bar{C} \dot{+} A = 1. \quad (1)$$

Equation (1) is valid if

$$\bar{B}\bar{C} = 1 - A \Longleftrightarrow \bar{B}\bar{C} = \bar{A} \Longleftrightarrow \overline{\bar{B}\bar{C}} = A$$
$$\Longleftrightarrow A = B \dot{+} C. \quad (2)$$

It is therefore possible to write A as a function of B and C:

$$A = f(B, C) = B \dot{+} C. \quad (3)$$

That is, A is true if B or C is true; $f(B, C)$ is a representation of A. The possible logical states of this equation, or logical states of A in function of B and C, can be represented in a Table of Karnaugh (Figure 1(a)) (Karnaugh 1953). This figure represents the predictive universe for a given representation at a given time. (The definitions of terms used throughout this article are explained in Box 2).

Equation (3) is obtained using what is deemed to be the objective knowledge of the subject and it is not the reflection of an absolute knowledge of the world. In other words, this equation depicts event A in function of the events B and C. In absolute terms it is possible that other events such as X or Y may also predict event A, but if this is ignored by the subject such events cannot be included in the equation.

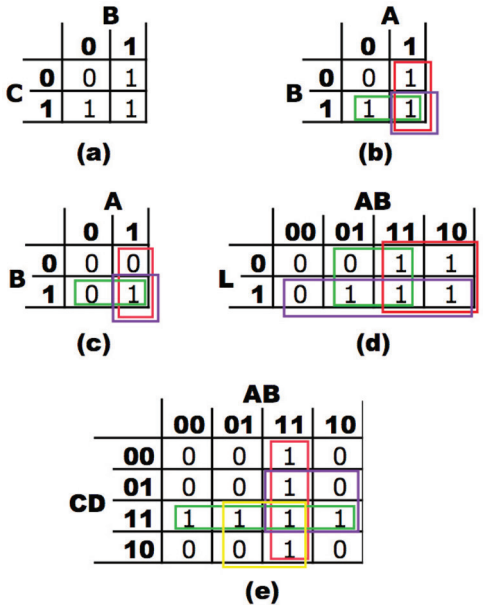

Figure 1.
(a) Truth table for $A = f(B, C) = B \dot{+} C$ (Karnaugh form); tables representing the predictive universe of (b) Food $= f(A, B) = A + B$; (c) Food $= f(A, B) = AB$; (d) Food $= f(A, B, L) = LB + A$; (e) Food $= f(A, B, C, D) = AB \dot{+} CD$. The first line and the first column of these tables give the possible state of predictor events or stimuli (0 if false, 1 if true). The tables also give the status for all combinations of predictor events for the predicted event (0 if false, 1 if true).

When animals learn, they associate a stimulus (e.g., A) with an event. The most commonly used event in experimental situations is the distribution of food (F). If A is presented immediately before or at the same time as food, A will become a predictor of food and so it can be written that $A \rightarrow F$. However, within the present theoretical framework these two modes of presentation (simultaneous or sequential) are not considered equivalent.

The Impact of Temporal Relationships Between Events or Stimuli

Within the present theory, temporal relationships between stimuli or events influence what is deemed to be learned by the animal, and consequently, its logical transcription. Consider two events A and B. If A and B occur sequentially with A preceding B without overlaps, then it can be written that $A \rightarrow B$. If A and B occur simultaneously or in succession but with an overlap, then it can be written that $A \rightarrow B$ and $B \rightarrow A$. Thus, within this theory association is directional (A toward B or B toward A) or bidirectional (A toward B and B toward A).

Within this theory, animals learn conditional propositions (if A then B) and their relations. Therefore, $A \rightarrow B$ also means that whenever A occurs, the neural representation of B will become active regardless of the "real" occurrence of B (prior

Box 2. Definitions and Premises.

Definitions
- Rule: A purported causality link (true or false).
- World representation: An ensemble of rules that predict the world around an animal.
- Predictive universe: An ensemble of predictions that can be inferred by a given representation of the world.
- Habituation: The waning of a reflex following repeated monotonous nonpertinent stimulations.
- Parallel predictive universe: An ensemble of rules identical to that of the general case minus the restricted rule or rules in a particular context. It is context-dependent and derived from a restricted world representation.

Premises
- Learning principle: Two events or objects will undergo association if and only if there is temporal contiguity between their two neural representations.

Total learning.
- Total learning is not considered equivalent to asymptotical learning, even if such an equivalence could be possible with regards to a single stimulus training. This is especially not the case when considering the training of a compound stimulus. In addition, total learning assumes that all relationship between events have been acquired.
- Rules are general by default.
- Restricting rules (using habituation/extinction) is equivalent to creating a parallel or conditional representation upon a context, and therefore a parallel or conditional predictive universe. In this conditional predictive universe, the restricted rule is interdicted.

In relation to prediction.
- The occurrence of an event is predicted in relation to a given purported predictor for an animal's given representation of the world, not the animal's absolute response to it, although in the right conditions they become correlated (e.g., the prediction of food, in the case of an animal in need of food).
- If two or more representations apply to a particular test stimulus (i.e., if the test stimulus is ambiguous in terms of representations), the predictions of all relevant representations are summed.

to extinction taking place), as B is a prediction made due to the occurrence of A. In addition, B would be associated with another event even in the absence of B when A is presented in conjunction with this other event, as it is the presentation of A that activates the representation of B (before extinction of $A \rightarrow B$ takes place). In other words, the neuronal networks that learn also support the memory and the structure of the memory of what has been learned, with each predictor causing the recall of what it predicts within a defined structure.[1] This leads on to what can be defined as the "Learning Principle": *Two events will undergo association (i.e., linked by conditionals propositions) if and only if there is temporal contiguity between their two neural representations.* (For a strong argument in favor of this rule of learning, see Miller and Barnet 1993.)

Examples of Sequential Presentation (Total Learning)

Let us suppose that a subject learns to associate A with food and then B with food (both in a sequential manner). It can be written that $A \rightarrow F$ and $B \rightarrow F$. Thus,

$$(A \rightarrow F)(B \rightarrow F) \Longleftrightarrow (\bar{A} \dot{+} F)(\bar{B} \dot{+} F) \tag{4}$$

$$\Longleftrightarrow \bar{A}\bar{B} \dot{+} \bar{A}F \dot{+} \bar{B}F \dot{+} F$$

$$\Longleftrightarrow \bar{A}\bar{B} \dot{+} F. \tag{5}$$

Equation (5) is valid if

$$\bar{A}\bar{B} \dot{+} F = 1 \quad \text{if } F = \overline{\bar{A}\bar{B}} = A \dot{+} B. \tag{6}$$

This logical equation, with regards to the prediction of food, can again be represented in a table. If A, B, and the compound stimulus AB are tested, it is possible to predict the level of response to each stimulus by reading Figure 1(b). If A was presented, it can be seen that A predicts food two out of two times (red rectangle), B predicts food two out of two times (green rectangle), and AB predicts food one out of one time (purple rectangle). Thus, from the learning perspective of the animal there is a certainty of food regardless of the test performed, and therefore all three stimuli will induce the same response. In other words, at total learning the system behaves like a digital system for what it knows.

Let us now suppose that the subject learned using the exact same stimuli (A and B) that the compound stimulus AB predicts food by the presentation of AB just prior to food being offered. The simultaneous presentation of A and B introduces an uncertainty: Is it A alone, B alone, or both of A and B that predict the deliverance of food? Without additional information the subject cannot discriminate between these possibilities. Therefore, it can be written in relation to the prediction of food that the subject has learned $(A \rightarrow F) \dot{+} (B \rightarrow F)$. Therefore,

$$(A \rightarrow F) \dot{+} (B \rightarrow F) \Longleftrightarrow (\bar{A} \dot{+} F) \dot{+} (\bar{B} \dot{+} F)$$

$$\Longleftrightarrow \bar{A} \dot{+} \bar{B} \dot{+} F. \tag{7}$$

For this equation to be true,

$$\bar{A} \dot{+} \bar{B} \dot{+} F = 1 \quad \text{if } F = \overline{\bar{A} \dot{+} \bar{B}} = AB. \tag{8}$$

We can also represent this equation in a table (Figure 1(c)).

If we now test A, B, and the compound stimulus AB, the level of response to each stimulus can be predicted using Figure 1(c). For example, when A is presented the event "food" is predicted one out of two times (red rectangle), B predicts

Table 1. Experimental design for Lattal and Nakajima (1998).

Phase 1	Phase 2	Test
$L+$, $A+$, $B+$	$LB+$	A, B

Note: All letters represent auditory or visual stimuli; "+" the delivery of reinforcement.

Table 2. Experimental design for Rescorla (2003b).

Conditioning	Test
$AB+$, $CD+$	AD, BC

Note: All letters represent auditory or visual stimuli; "+" the delivery of reinforcement.

food one out of two times (green rectangle), and AB predicts food one out of one time (purple rectangle). Thus, it can be deduced that $AB > A = B$, as AB is, from what the subject has learned, the certainty of the event "food," while A or B predict food with the same uncertainty ($A = B = 1/2$ and $AB = 1$). It is important to notice that the absolute values of the responses are not what is being predicted, but rather their relations. Thus, as stated previously, at total learning the system behaves like a digital system in relation to what it knows (AB), but it also behaves like an analogue system for what it does not know (A and B individually). This example also provides an interpretation of the phenomenon that Pavlov called "overshadowing" (Pavlov 1927). In this case, both stimuli are equally salient and if A is trained in association with B in the compound stimulus AB, then $A = 1/2$ (Figure 1(c), red rectangle). However, if A is trained separately to the stimulus B, then $A = 1$ (Figure 1(b), red rectangle). In order to test the above propositions, they were applied to some previously published training designs and their success determined. Some examples are given below.

Confronting the Data

Within the Scope of Actual Theories—Two Simple Examples

The following example of overexpectation is from Lattal and Nakajima (1998). The protocol was as follows (see also Table 1):

1. Subjects learned to associate L, A, and B with food (F).
2. Subjects then learned that the compound stimulus LB predicts food, not L or B alone (L or B alone were not presented in this phase of training).

Therefore, a representation of F can be derived as a function of A, B, and L. This function can be defined by

$$F = f(A, B, L) = LB \dot{+} A. \tag{9}$$

This follows from

$$(A \rightarrow F)(L \rightarrow F \dot{+} B \rightarrow F). \tag{10}$$

This proposition is equivalent to

$$(\bar{A} \dot{+} F)(\bar{L} \dot{+} F \dot{+} \bar{B} \dot{+} F) \iff \bar{A}\bar{L} \dot{+} \bar{A}\bar{B} \dot{+} \bar{A}F \dot{+} \bar{L}F \dot{+} \bar{B}F \dot{+} F, \tag{11}$$

which in turn is equivalent to

$$\bar{A}\bar{L} \dot{+} \bar{A}\bar{B} \dot{+} F. \tag{12}$$

Equation (11) is valid for

$$F = \overline{\bar{A}\bar{L} \dot{+} \bar{A}\bar{B}} = A \dot{+} AB \dot{+} LB = A \dot{+} LB. \tag{13}$$

From this, Figure 1(d) can be constructed. In addition, the subjects have learned

$$L \rightarrow B \quad \text{and} \quad B \rightarrow L. \tag{14}$$

Or, that L is conditional upon B and B is conditional upon L.

If the response to A (red rectangle), B (green rectangle), and L (purple rectangle) were to be predicted, it is possible to deduce from Figure 1(d) that the response would be $A(4/4) > B(3/4) = L(3/4)$. This prediction agrees with experimental data from various authors (e.g., Lattal and Nakajima 1998; Rescorla 2003a) and conforms to what elemental theories call "overexpectation." The second example is from Rescorla (2003b). Here, the subjects learned (see Table 2) that the compound stimuli AB and CD each predict the appearance of food (F).

It can therefore be written in relation to the prediction of food that

$$F = f(A, B, C, D) = AB \dot{+} CD. \tag{15}$$

From this a predictive table can be constructed (see Figure 1(e)).

The subjects also learned that

$$(A \rightarrow B \text{ and } B \rightarrow A) \text{ and } (C \rightarrow D \text{ and } D \rightarrow C). \tag{16}$$

At test, subjects were confronted with the original compounds (AB, CD), transfer compounds (AD, CB), and each individual stimulus A, B, C, and D. From Figure 1(e) it is possible to predict that

1. $AB = CD = 4/4 = 1$ (red and green rectangles),
2. $AD = CB = 3/4 = 0.75$ (purple and yellow rectangles), and
3. $A = B = C = D = 5/8 = 0.625$.

AQ2

Therefore, $AB = CD > AD = CB > A = B = C = D$. This accurately supports the finding of Rescorla (2003b). In contrast, elemental theories of learning predict

$$AB = CD = AD = CB > A = B = C = D,$$

and configural theories of learning predict

$$AB = CD > AD = CB = A = B = C = D$$

(see Rescorla 2003b).

In order to further test the predictive abilities of this new theory, more challenging experimental situations involving habituation and/or extinction procedures were investigated.

The Treatment of Habituation or Extinction

It is proposed that the function of habituation or extinction is to restrict the field of application for a given rule. Habituation is referred to when this rule is preexisting (i.e., a rule that the great majority of a given species share and can be considered innate). Extinction is referred to when the rule has been acquired previously (for example through an experiment), and therefore is shared only by a limited number of individuals in a given species. In this context, the distinction between habituation and extinction is no more than rhetorical. It can therefore be postulated that the only discernable difference between habituation and extinction is the intrinsic pertinence of the rule that is being acted upon in relation to the survival of the species and their relative plasticity. However, in relation to learning, their consequences are similar.

It is generally accepted that after habituation or extinction, a rule (the causality-like link) does not disappear (Pavlov 1927; Brimer 1972; Rescorla 1996, 2003c). Instead the rule is not displayed in a particular context. Following on from this, it could be postulated that during extinction or habituation a parallel predictive universe is created. The predicted event would then be dependent upon two logical equations. For example, if the subjects first learn to associate A with food and B with food, the following equation can be written

$$\text{Food} = f(A, B) = A \dot{+} B. \tag{17}$$

Subsequently, if the subjects later learn that the stimulus N in association with A does not predict food, but rather B presented alone predicts food, the stimulus N becomes an indicator of the predictive universe in association with the experimental context. As rules are general by default, in the universe defined by the occurrence of N the logical equation becomes

$$\text{Food} = f(B) = B. \tag{18}$$

Table 3. Experimental design for Rescorla (2004).

Phase 1	Phase 2	Test
$A+$, $B+$, $NA-$, $NB-$	$NA+$, $B+$	A, B

Note: All letters represent auditory or visual stimuli; "+" the delivery of reinforcement and "−" the absence of reinforcement.

In the general case the equation is still

$$\text{Food} = f(A, B) = A \dot{+} B. \tag{19}$$

In addition, the subjects have learned the relationship between A and N after being exposed to NA learning

$$(N \rightarrow A \text{ and } A \rightarrow N). \tag{20}$$

In the present theory, it is postulated that if a question (a test stimulus) is ambiguous in terms of a valid predictive representation, the predictions from each relevant representation are summed. This is because it is not possible for the subject to be certain which representation is solely valid while making the prediction. However, if a question (e.g., a test stimulus) is not ambiguous, only the stipulated representation is used. The following examples demonstrate how to predict the outcome of training sessions that include extinction in their design. The first is a "superconditioning" experiment reported by Rescorla (2004). The second is also from Rescorla (2000) and illustrates the particular predictive value of the present theory by comparison with current elemental theories.

Example 1. The design of this "superconditioning" experiment was as follows. First, the experimental subjects learned to associate A, B with the delivery of food (F) and NA and NB with its absence (see Table 3). In this condition, in the general case the subjects learned that A and B predicted the delivery of food ($A \rightarrow F$ and $B \rightarrow F$), whereas in the "experimental context" (plus N) the delivery of food did not have a predictor $A \rightarrow \bar{F}$ and $B \rightarrow \bar{F}$.

In addition, they also learned that $N \rightarrow (A \oplus B)$ ($\oplus$ is the sign for exclusive "or"). That is, N is conditional upon the event A or the event B, but not both at the same time (A and B are mutually exclusive in relation to N). Following this conditioning, the subjects learned that NA was followed by the delivery of food, and B was followed by the delivery of food. The individual stimuli A and B were then tested.

Based on the present theory, after this first conditioning, the subjects built the two following representations:

(i) Representation 1 (the general case)

$$F = f(A, B) = A \dot{+} B. \tag{21}$$

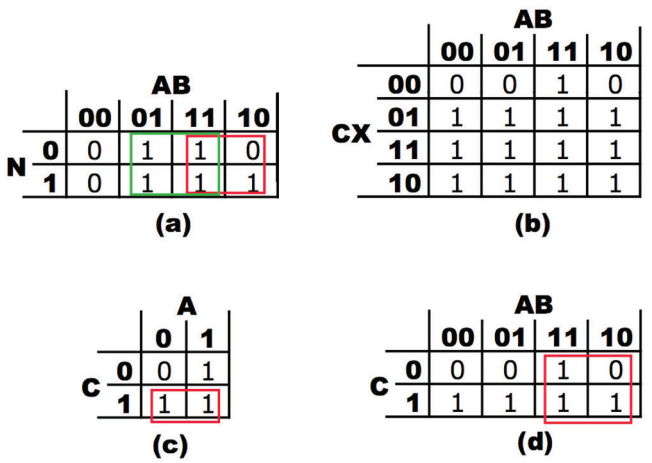

Figure 2.
(a) Predictive universe of Food $= f(A, B) = NA \dot{+} B$; (b, c, d). Predictive universes given respectively by the general case, the representation defined by D (+ context) and the representation defined by B (+ context).

(ii) Representation 2 (the "experimental context" and N): F, the delivery of food is not predicted. After the second conditioning, Representation 1 (the general case) is changed to

$$F = f(A, B) = NA \dot{+} B. \tag{22}$$

As seen here a stimulus can simultaneously be an indicator of a predictive universe that restricts a rule and a predictor of F. Finally in Representation 2, F is now predicted as

$$F = f(A) = A. \tag{23}$$

Note that the relationship between the elements has now changed to $N \rightarrow (A \oplus B)$ and $(A \rightarrow N)$. Figure 2(a) represents the predictive universe given by Representation 1. The predictive universe for Representation 2 is quite simple: if A then F.

In Representation 1, A predicts that F is true in three out of four cases (Figure 2(a), red rectangle) and B predicts that F is true in four out of four cases (Figure 2(a), green rectangle). However, as $A \rightarrow N$, when A is tested, Representation 2 is evoked. ("Experimental context" is also a factor, but since in this case it is stable it can be ignored.) In this representation, F is true if A is true, therefore A predicts that in this representation F is true one out of one time. The relationship existing between B and N is *not* $B \rightarrow N$ but $N \rightarrow B$ since in the second phase of conditioning B was presented alone and not in conjunction with N. This indicates that within the experimental context B did not predict N whereas the prediction of B by N was unaffected by the second phase of training (see Table 3). In other words, the predictive universe defined by N is not active when B is presented alone. Thus, it can be written that in relation to the prediction of food, $A = 3/4 + 1/1 = 7/4 = 1.75$, whereas in relation to the prediction of food, $B = 4/4 = 1$. Therefore, $A > B$ as was demonstrated

Table 4. Experimental design for Rescorla (2000).

Pretraining	Phase 1	Test
$A+, C+, X+, XB-, XD-$	$AB+$	AD, BC

Note: All letters represent auditory or visual stimuli; "+" the delivery of reinforcement and "−" the absence of reinforcement.

experimentally and in accordance with both the present theory and the elemental theories of associative learning.

Example 2. The experimental design was as follows. In pretraining sessions the subjects learned that A, C, and X predicted the delivery of food (F). Compound stimulus XB and XD ended with no delivery of food. Pretraining was followed by conditioning sessions of AB stimulus followed by the delivery of food. Subjects were then tested with the compound stimuli AD and BC (see also Table 4).

Based on the present framework, after pretraining the subjects know
(i) in the general case,

$$F = f(A, C, X) = A \dot{+} C \dot{+} X; \tag{24}$$

(ii) in the representation defined by B or by D (and context),

$$F = f(A, C) = A \dot{+} C. \tag{25}$$

In addition, due to the alternated exposure to XD and XB, D and B are mutually exclusive (if D then $\bar{B}$ and if B then $\bar{D}$). After conditioning the subjects know

(i) in the general case,

$$F = f(A, B, C, X) = AB \dot{+} C \dot{+} X; \tag{26}$$

(ii) in the representation defined by B (and context),

$$F = f(A, C) = A \dot{+} C; \tag{27}$$

(iii) in the representation defined by D (and context),

$$F = f(A, B, C) = AB \dot{+} C. \tag{28}$$

These representations can be translated into Figures 2(b), (c), and (d).

After training, the subjects were presented with the test stimuli AD and BC. Each test stimulus unambiguously calls for at least a particular restricted representation (due to the presence in each test stimulus of an indicator of restricted representation, viz. B and D). Thus, the general representation is of no use; only the representations governed by D or B are used to make predictions. Furthermore, as no ambiguities exist for

each test stimulus in relation with the pertinent representation, no interrepresentation summation is expected (see Box 2).

In the representation defined by D (+ context), BC is not relevant as stimulus D is absent. Note that in this case the context has not changed throughout the experiment and so it becomes irrelevant. In contrast, AD is relevant as it contains stimulus D. In the representation governed by D, the test stimulus AD is equivalent to A. Therefore, in this representation, $AD = A = 3/4$ (see Figure 2(d), red rectangle). In the representation that is governed by B, AD is also not relevant since B is absent. In this representation, the test stimulus BC is equivalent to C, as B is an indicator of this representation; thus $BC = C = 2/2$ (see Figure 2(c), red rectangle). It follows that $AD = 3/4 = 0.75$, whereas $BC = 2/2 = 1$. Thus it is predicted that $BC > AC$, which is in accordance with the experimental results. In contrast, elemental theories of learning predict $AD = BC$ for the Rescorla-Wagner (Rescorla and Wagner 1972) model, and $AD > BC$ for the Mackintosh model (Mackintosh 1975; see Rescorla 2000).

Outside the Scope of Current Theories–Two Examples

Learning and re-learning Higgins and Rescorla (2004) have reported an interesting phenomenon that is beyond the interpretive power of current predictive learning theories. The authors have reproduced and completed earlier results obtained by Freberg (reported by Rescorla 1981). The experimental details were as follows (see Higgins and Rescorla 2004):

The first experimental group received the following training (Experiment 1):

1. Simultaneous presentation of an almond flavor (A) with the appetitive compound polycose (P). The two compounds were mixed together in water.
2. Extinction of the association of A with P by the presentation of A alone.

The second experimental group received the following training (Experiment 2):

1. Sequential presentation of the almond flavor (A) followed by presentation of polycose (P). A solution of water containing the almond flavor was provided before a second solution of water containing polycose.
2. Extinction of the association of A with P by the presentation of A alone (not followed by P).

In both cases, the authors attempted retraining using the same initial protocol and compared the efficiency of this retraining between the two groups. Surprisingly, retraining failed for Group 1 (the group receiving simultaneous training), while it was successful for Group 2 (the group receiving sequential training). Thus, A did not reacquire its predictive value using simultaneous training (Group 1), whereas when sequential training was used (Group 2), re-acquisition did occur.

This surprising result can be easily explained using the framework of the present theory. Group 1 (simultaneous presentation) learned that $A \rightarrow P$ and $P \rightarrow A$, but Group 2 (sequential presentation) learned that $A \rightarrow P$ only. In terms of the prediction of P, both groups learned the same thing: $A \rightarrow P$. However, for Group 1, there is some uncertainty as to what this simultaneous presentation means, while for Group 2 no such problem exists. When extinction was performed, A was presented alone in both groups. This effectively reduced the uncertainty for Group 1. Group 1 therefore was aware that $P \rightarrow A$ and that the proposition $A \rightarrow P$ was false. In relation to Group 2, the results were quite different, as Group 2 did not have a predictor for P at this time. Thus, when retraining was attempted, it did not challenge the representation that Group 1 had constructed because A was presented simultaneously with P (meaning $P \rightarrow A$), and therefore no new association of A and P was needed. However, the existing representation of Group 2 was challenged through retraining because the events that occurred were not predicted and therefore, a new re-association of A and P was necessary. In the context of the present theory, it is possible to predict that if sequential training is performed as "retraining" for Group 1, a re-association would be made between A and P in the form of $A \rightarrow P$ to the cost of $P \rightarrow A$. Moreover, the presentation of P alone in addition to A alone, during the extinction process for Group 1, should enable successful retraining using the simultaneous protocol utilized in this experiment.

Novelty detection: The known, the unknown, and the not known This theory also offers a theoretical account of the phenomenon known as "novelty detection." If the notion of "known, unknown, and not known" (Guez 2000) were considered in the framework of the present theory, "the known" would be contained in a representation of reality structured as multiple chains of conditional propositions. "The unknown" would be represented as a structure that is partially predicted by the available representation of reality, and therefore could be integrated into it provided that some changes were made. "The not known" is all that is beyond the predictive universe as defined by the representation of reality, which is in use at that given time.

If the environment known to an animal is represented by multiple chains of conditional propositions, the detection of novelty will take place when predictions based on this representation fail. The extent of the response will be determined by the extent to which this representation fails to predict a particular event. The greater the failure, the closer the unknown is to the not known, and so the greater the response. Obviously however, some not known will never induce a response if the perceptual world of a given animal ("Umwelt"; Von Uexküll 1965) cannot describe them, and therefore cannot be integrated into its representation of reality.

Consequence: External inhibition The detection of novelty implies that a representation of reality has failed. Therefore, it is imperative for the animal to evaluate to which extent this failure must change its representation of the world before using the rules that are part of this representation. This simple observation explains the phenomenon that Pavlov called "external inhibition" (Pavlov 1927). Pavlov observed that tests would fail if they were performed in a room other than the one where training was performed, and/or if a loud noise (e.g., a truck passing in the street) occurred during testing. It is likely that such a loud noise is not predicted by the representation of reality of a laboratory animal. If it is the case, the noise is detected as novelty, and consequentially disrupts the acquisition or the display of a known rule until the animal can evaluate its impact upon itself.

Furthermore, in the case of a laboratory animal with a relatively poor representation of reality (i.e., with no experience of the natural world), such an effect could be quite dramatic.

Conclusion

The present theory demonstrates a large predictive and explanatory scope, and proposes a representational structure used by the brain to order knowledge extracted from the world in a logical framework. It has predictive power with respect to a variety of experimental designs, both inside (Lattal and Nakajima 1998; Rescorla 2000, 2003b) and outside the framework of current theories (Higgins and Rescorla 2004), and is able to theoretically explain phenomena such as the detection of novelty. A highlight of this model is probably its capacity to deal successfully with extinction without using the concept of unlearning. Instead, the emphasis is on the way the brain structures and uses information gathered from the surrounding environment. In the present framework, extinction only restricts a rule within a particular context; it does not erase the rule. This explains why a rule is masked rather than erased by extinction, and also allows for the prediction of phenomena such as renewal and spontaneous recovery from extinction (Pearce and Bouton 2001). A second highlight of the present theory is that it does not postulate major differences between events; in other words, it does not postulate that learning occurs only with reinforcement. A third highlight is the rejection of the notion of a nonassociative form of learning, which is a step toward a more generalized approach of learning. Finally, the theory is sensitive to the order in which the different events are learned: It is theoretically different to learn $NA-$ in the first phase then $A+$ in the second experimental phase, than $A+$ in the first phase then $NA-$ in the second ("+" indicating the delivery of the reinforcer, "−" its absence).

The extension of the concept of "association" to the concept of "rules" of the form "if A then B" enables this theory to deal with problem solving and causal reasoning, as recently evidenced in rats (Blaisdell et al. 2006). It also allows the use of meta-rules ("rules about rules"). This new feature appears necessary in order to deal with the pretraining effects of unrelated stimuli on subsequent training that uses different stimuli (e.g., Beckers et al. 2006). Within the present framework, meta-rules are expected to be context-dependent, since they are the expression of a change in the default behavior of the learning system, and therefore can be said to result from a restriction of the default learning rules.

Extending the concept of "association" to the concept of "rule" offered here in the form of an "if-then relationship" should not be viewed as a large departure from the basic concept of association. Instead, it merely adds to it and formalizes a notion of direction. If in more traditional associative theory we say that A is associated with B, it could mean in the present theory that if A then B, or if B then A, or both. The introduction of this limited notion of "rule" introduces more precision as to what is actually going on. For some, the inability of this theory to fully describe a learning/performance curve as a mathematical model could be seen as a weakness. But no doubt one could argue that describing mathematically a curve of any sort should not be confused with an explanation of the underlying mechanism that underpins this curve. A mathematical description is not an explanation per se of a phenomenon; it is just one possible way to describe it.

A distinct advantage of using a logical representation is that it allows the deduction of the minimum neuronal network necessary for solving a given problem. The construction of such a network is dependent on regarding the actions of individual neurons or a pair of neurons as equivalent to logical operators. When neuronal inputs are synchronized the neuron acts as an "and-gate" ($\times$), whereas when the same inputs are desynchronized the neuron acts as an "inclusive-or-gate" ($\dot{+}$). The construction of a conditional proposition ($\rightarrow$) requires the use of a minimum of two neurons, with a strong connecting synapse that allows the activation of the postsynaptic neuron following the activation of the presynaptic neuron. Since the logical operators "inclusive or" ($\dot{+}$), "and" ($\times$), and "the conditional proposition" ($\rightarrow$) can all be described using only one of them and the logical operator "no," these logical operators allow the reconstruction of the minimum biological network that is equivalent to the learned logical equation. In actuality, real networks may be more complex, but logically they can be regarded as perfectly equivalent to the minimum network. Obviously this is not to say that this kind of limited network is "the network" that could solve all problems. Such a network should be seen as part of a bigger network in a similar way that the limbic system is part of the brain in vertebrates.

Throughout this work I have assumed total learning, and limited my analysis to it. It is likely that the scope of this theory could be extended to incorporate partial learning situations by introducing at least some new premises, one of

which would be to consider that each partial rule defines a separate representation, and therefore its own predictive universe. The consequence of this would be that if an animal partially learned that $A \rightarrow F$ and $B \rightarrow F$ to the same performance level, the test response to the presentation of A or B alone or the compound AB would be of the form $A = B < AB$ because the test stimulus AB would call on two representations and therefore the prediction of each separate predictive universe would be summated (see Box 2). The second would be that the integration of knowledge occurs secondarily when dealing with partial learning; an illustration of such a mechanism is the blocking phenomenon (Kamin 1969). An example of blocking training is as follows: In phase 1 of training a stimulus A is paired with a US (e.g., food, F), so that the subjects learn that $A \rightarrow F$. In phase 2, a compound stimulus AX is paired with the same US (typically the number of trials in phase 1 is far larger than the number of trials during phase 2). After phase 2 training the subjects have learned that $A \rightarrow F$ (first predictive universe) and $A \rightarrow F$ or $X \rightarrow F$ (second predictive universe). At test stimulus X will activate the representation of A and in consequence the two predictive universes for the prediction of F are activated. So A and AX are predictors of F ($A \rightarrow F$ and $A \rightarrow F$ or $X \rightarrow F$); this is equivalent to

$$\begin{aligned}&(\bar{A} \dotplus F)(\bar{A} \dotplus F \dotplus \bar{X} \dotplus F)\\ &\quad \Longleftrightarrow \bar{A} \dotplus \bar{A}F \dotplus \bar{A}\bar{X} \dotplus \bar{A}F \dotplus F \dotplus \bar{X}F\\ &\quad \Longleftrightarrow \bar{A} \dotplus F. \end{aligned} \tag{29}$$

Equation (29) is said to be valid for

$$\bar{A} \dotplus F = 1 \Longleftrightarrow F = A. \tag{30}$$

Stimulus X will fail to elicit a response as it is not part of the equation predicting F ($F = A$). Nevertheless, X has been "associated" with F (the subjects have learned that $A \rightarrow F$ and $A \rightarrow F$ or $X \rightarrow F$), but this "association" will not be expressed until A and X fully predict each other due to their compound presentation in the form of AX, creating the rules $A \rightarrow X$ and $X \rightarrow A$. However, the rule $X \rightarrow A$ must at least partially exist after compound training, and therefore some weak response to X should be observed because X predicts A and A predicts F in full.

This theory already shows great promise and offers a new way of thinking about learning. It expands the field of application of learning theories to a wider range of situations, proposing a new way to deal successfully with extinction. It can be extended by adding a more elaborate decision-making process by, for example, using a hierarchy of threshold-driven rules, or even be mathematically modeled in order to regain the capacity of describing learning curves. I believe that for all these objective reasons the present theory is well worth considering.

Acknowledgments

I would like to thank R. Guez for our useful discussions during the elaboration of this article, and C. L. Conway, D. Wheeler, J. C. Amundson, J. Zeil, G. Urcelay, J. Arndt, and R. R. Miller for their comments on an earlier version of this manuscript. I would also like to thank three anonymous reviewers for their very constructive comments. This work began at the Australian National University, was further developed at the State University of New York with the support of NIH Grant No. 33881, and was finally completed at the University of Canberra. I dedicate this work to my father who taught me so much.

Note

1. This is not to say that the mediated activation of the neural representation of B is of the same amplitude of the activation of the neural representation of B when B is physically present. If such was the case it will be probably equivalent to a hallucination.

References

Atkinson R, Estes W (1963) Stimulus sampling theory II. In: Handbook of Mathematical Psychology (Luce R, Bush RR, Galanter E, eds), 121–268. New York: Academic Press.

Beckers T, Miller R, De Houwer J, Urushihara K (2006) Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. Journal of Experimental Psychology: General 135: 92–102.

Blaisdell A, Sawa K, Leising K, Waldmann M (2006) Causal reasoning in rats. Science 311: 1020–1022.

Bliss T, Gardner-Medwin A (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the unanesthetized rabbit following stimulation of the perforant path. Journal of Physiology 232: 357–374.

Bliss T, Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of anesthetized rabbit following stimulation of perforant path. Journal of Physiology 232: 331–356.

Boole G (2003) The Laws of Thought. New York: Prometheus Books.

Brimer C (1972) Inhibition and learning. In: Disinhibition of Operant Response (Boakes R, Halliday M, eds), 225–227. New York: Academic Press.

Gallistel C (1990) The Organization of Learning. Cambridge, MA: MIT Press.

Goddard M (2003) Latent inhibition of US signal value. Quarterly Journal of Experimental Psychology B 56: 177–192.

Guez R (2000) Les traces et l'art en question. In: Approche du rôle des traces au niveau plastique dans l'oeuvre visuelle (Berthet D, ed), 57–79. Paris: L'Harmattan.

Haselgrove M, Aydin A, Pearce J (2004) A partial reinforcement extinction effect despite equal rates of reinforcement during Pavlovian conditioning. Journal of Experimental Psychology: Animal Behavior Processes 30: 240–250.

Hebb D (1949) The Organization of Behavior. New York: Wiley.

Higgins T, Rescorla R (2004) Extinction and retraining of simultaneous and successive flavor conditioning. Learning and Behavior 32: 213–219.

Kamin LJ (1969) Predictablility, surprise, attention and conditioning. In: Punishment and Aversive Behavior (Campbell BA, Church RM, eds), 279–296. New York: Appleton-Century-Crofts.

Karnaugh M (1953) The map method for synthesis of combinational logic circuit. Transactions of the AIEE Part 1 72: 593–599.

Lattal K, Nakajima S (1998) Overexpectation in appetitive Pavlovian and instrumental conditioning. Animal and Learning Behavior 26: 351–360.

Mackintosh N (1975) A theory of attention: Variations in associability of stimuli with reinforcement. Psychological Review 82: 276–298.

Miller R, Barnet R (1993) The role of time in elementary associations. Current Directions in Psychological Science 2: 106–111.

Miller R, Barnet R, Grahame N (1995) Assessment of the Rescorla-Wagner model. Psychological Bulletin 117: 363–386.

Pavlov I (1927) Conditioned Reflexes. Oxford: Oxford University Press.

Pearce J (1987) A model for stimulus generalization in Pavlovian conditioning. Psychological Review 94: 61–73.

Pearce J, Bouton M (2001) Theories of associative learning in animals. Annual Review of Psychology 52: 111–139.

Rescorla R (1981) Advances in analysis of behavior. In: Simultaneous Associations 2 (Harzem P, Zeiler M, eds), 47–80. New York: Wiley.

Rescorla R (1996) Preservation of Pavlovian associations through extinction. Quarterly Journal of Experimental Psychology B 49: 245–258.

Rescorla R (2000) Associative changes in excitors and inhibitors differ when they are conditioned in compound. Journal of Experimental Psychology: Animal Behavior Processes 26: 428–438.

Rescorla R (2003a) Contemporary study of Pavlovian conditioning. Spanish Journal of Psychology 6: 185–195.

Rescorla R (2003b) Elemental and configural encoding of the conditioned stimulus. Quarterly Journal of Experimental Psychology B 56: 161–176.

Rescorla R (2003c) Protection from extinction. Learning and Behavior 31: 124–132.

Rescorla R (2004) Superconditioning from a reduced reinforcer. Quarterly Journal of Experimental Psychology B 57: 133–152.

Rescorla R, Wagner A (1972) Classical conditioning II. In: A Theory of Pavlovian Conditioning: Variation in the Effectiveness of Reinforcement and Nonreinforcement (Black A, Prokasy W, eds), 64–99. New York: Appleton-Century-Crofts.

Von Uexküll J (1965) Mondes animaux et monde humain. Paris: Gonthier.

Wagner A (2003) Context-sensitive elemental theory. Quarterly Journal of Experimental Psychology B 56: 7–29.