Behavioural Processes 167 (2019) 103908

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Beyond the "Conceptual Nervous System": Can computational cognitive neuroscience transform learning theory?

Fabian A. Soto*

Department of Psychology, Florida International University, 11200 SW 8th St, AHC4 460, Miami, FL 33199, United States

ARTICLE INFO

Keywords:
Learning theory
Associationism
Connectionism
Theory underdetermination
Model identifiability
Computational cognitive neuroscience

ABSTRACT

In the last century, learning theory has been dominated by an approach assuming that associations between hypothetical representational nodes can support the acquisition of knowledge about the environment. The similarities between this approach and connectionism did not go unnoticed to learning theorists, with many of them explicitly adopting a neural network approach in the modeling of learning phenomena. Skinner famously criticized such use of *hypothetical* neural structures for the explanation of behavior (the "Conceptual Nervous System"), and one aspect of his criticism has proven to be correct: theory underdetermination is a pervasive problem in cognitive modeling in general, and in associationist and connectionist models in particular. That is, models implementing two very different cognitive processes often make the exact same behavioral predictions, meaning that important theoretical questions posed by contrasting the two models remain unanswered. We show through several examples that theory underdetermination is common in the learning theory literature, affecting the solvability of some of the most important theoretical problems that have been posed in the last decades. Computational cognitive neuroscience (CCN) offers a solution to this problem, by including neurobiological constraints in computational models of behavior and cognition. Rather than simply being inspired by neural computation, CCN models are built to reflect as much as possible about the *actual* neural structures thought to underlie a particular behavior. They go beyond the "Conceptual Nervous System" and offer a true integration of behavioral and neural levels of analysis.

1. Introduction

In the last century, learning theory has been dominated by what Hall (2002, 2016) calls the *Conceptual Nervous System* (CNS) approach, which highlights the formation of associations as an elemental learning mechanism that can support the acquisition of knowledge about our environment. In Hall's words (2002, p. 1):

> Specific accounts differ in many ways (as we shall see), but the central assumption of all associative analyses of conditioning has been that the effects observed can be explained in terms of the operation of a *conceptual nervous system* that consists of entities (to be referred to as *nodes*) among which links can form as a result of the training procedures employed in conditioning experiments. The existence of a link allows activity in one node to modify the activity occurring in another node to which it has become connected.

Hall also highlights the value of conditioning studies to reveal the principles by which such elemental learning mechanisms operate:

> Conditioning studies are seen as a tool that can tell us about the association between particular event representations (sometimes called nodes); but the principles revealed by these studies will have relevance to the specification of a conceptual nervous system consisting of a huge array of such nodes corresponding to all perceivable stimuli (and possibly, all behavioural outputs). Psychological phenomena are assumed to be determined by the activation of these nodes, and behavioural adaptation by the formation of connections among them, and the propagation of activation around the network. These notions will now seem familiar, being those popularised rather later (e.g., Rumelhart and McClelland, 1986) under the heading of connectionism; but they were anticipated by students of animal learning.

While Hall does a fantastic job at clearly articulating the assumptions behind the CNS approach, he himself admits that this is not his invention, but rather a common approach taken by researchers of animal learning (Hall, 2016). The work of Mackintosh, Wagner, Rescorla, Dickinson, and their students clearly falls within this general approach,

* Corresponding author.
E-mail address: fasoto@fiu.edu.

https://doi.org/10.1016/j.beproc.2019.103908
Received 2 December 2018; Received in revised form 8 May 2019; Accepted 11 July 2019
Available online 03 August 2019
0376-6357/

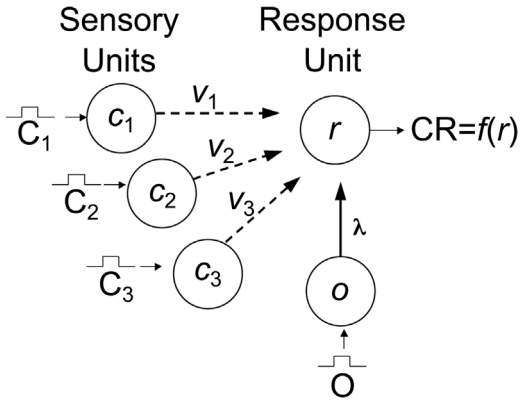

Fig. 1. Prototypical CNS model of Pavlovian conditioning, which can also be interpreted as a neural network model (see Vogel et al., 2004).

but the general ideas were first popularized by Pavlov (1927) himself. In addition, as Hall points out, there is a clear conceptual relation between the CNS approach from learning theory and connectionism (Rumelhart and McClelland, 1986), which was made explicit in the 1990s with the work of Schmajuk and others (Schmajuk, 1997), who pioneered the development of neural network models of associative learning. This type of model is still very popular in the field, with Mondragón et al. (2017) very recently proposing the use of deep neural networks to model associative learning processes.

The CNS approach has proven to have high heuristic value, as attested by the proliferation of quantitative CNS models of Pavlovian and instrumental conditioning in the last decades. Fig. 1 offers a prototypical CNS model of Pavlovian conditioning (adapted from Vogel et al., 2004). In this prototypical model, the physical presentation of environmental cues and an unconditioned stimulus or outcome (C_i and O, represented by square signals), activate their internal representations (c_i and o) or "sensory units." Through a network of associative connections, such sensory units affect activity of a "response unit," which is responsible for generation of a conditioned response, or CR. The strength of the connection between an outcome and the response unit is represented by λ, whereas the modifiable strength of connections between cue representations and the response unit is represented by v_i. Associative learning would happen because the response unit can, through training, be activated by the c units, whereas without training the response unit is only activated by the outcome.

As many readers might know, the idea of a *conceptual nervous system* was first proposed by Skinner in *The Behavior of Organisms* (Skinner, 1938), and used throughout his career to criticize the inference of neural mechanisms from behavior alone:

> the traditional 'C. N. S.' might be said to stand for the Conceptual Nervous System. The data upon which the system is based are very close to those of a science of behavior, and the difference in formulation may certainly be said to be trivial with respect to the status of the observed facts. […] In any one of these examples the essential advance from a description of behavior at its own level is, I submit, very slight. An explanation of behavior in conceptual terms of this sort would not be highly gratifying. But a Conceptual Nervous System is probably not what the neurologist has in mind when he speaks of the neural correlates of behavior.

While Skinner was not against using neurobiological mechanisms to explain behavior, he proposed that inferring such mechanisms from behavior was problematic. Despite Skinner's criticism, the CNS approach has not only survived for 80 years since the publication of *The Behavior of Organisms*, but it has been wholeheartedly adopted to such an extent that Skinner's original name for the approach, proposed in a tongue-in-cheek manner, is now used by theorists as a good way to summarize what they are indeed attempting.

Here, I would like to propose that Skinner was onto something: there are important theoretical problems with the CNS approach, and with computational cognitive modeling in general, and the time is perhaps ripe to address them.

2. Underdetermination of scientific theory

The underdetermination of scientific theory by evidence is a problem widely discussed in philosophy of science (for recent reviews, see Stanford, 2017; Turnbull, 2018). In short, it refers to the general problem that one cannot decide in favor of one theory over others based on empirical evidence alone. Philosophers distinguish among many forms of underdetermination, and some of them seem frankly too abstract and esoteric to be a source of worry for cognitive modelers (Turnbull, 2018). For example, holist underdetermination, also known as the Duhem-Quine thesis, proposes that when a theory is tested, a number of auxiliary assumptions and hypotheses must be held to carry out the experimental test. When a researcher encounters evidence that seems to falsify the theory, it is unclear whether such evidence truly falsifies the target theory or any of the auxiliary hypotheses. For example, a possibility is to conclude that the instruments used to test the theory were not functioning correctly. In that and similar concrete cases, the usual practice is to perform independent tests of the auxiliary hypotheses. In addition, rejection of such auxiliary hypotheses usually has the undesirable consequence of being more costly for science than rejecting the target theory (Laudan, 1990). In any case, this is not the type of underdetermination that I will focus on here.

Other forms of underdetermination seem rather trivial. For example, practical underdetermination refers to the case in which the currently available evidence cannot decide between competing theories. This is a common situation that has the simple solution of performing additional tests of the available theories.

Here, I will focus on a third form of underdetermination, which different scholars refer to as contrastive underdetermination (Stanford, 2017), equivalence underdetermination (Turnbull, 2018), or theory underdetermination (Lyre, 2011). According to theory underdetermination, any body of observable evidence can always support two or more theories equally well, even though the are ontologically different by proposing completely different unobservable constructs. I will not address the abstract case of an unconceived alternative to a proposed theory (Stanford, 2017), but instead focus on the more concrete case in which two already-proposed theories are known to be empirically equivalent. This form of "local" theory underdetermination tends to be uninteresting for philosophers, who find general arguments about "global" or pervasive forms of underdetermination more compelling (Grünbaum, 2018). However, concrete examples of "local" underdetermination are usually the most interesting for cognitive researchers, especially when the empirically-equivalent theories propose alternative, competing mechanisms of cognitive processing or representation. Here, I will focus specifically in such non-trivial cases of local theory underdetermination. In addition, I will not address cases in which two theories are so closely related that they can be considered simply two variants of the same theory (Norton, 2008). All the case studies included in the next section involve mechanisms that are considered to provide conceptually different and competing explanations of the same behavior, and have driven enough research and theoretical development to ensure that their differences are clearly important for learning scientists.

As indicated by Grünbaum (2018), while cases of practical underdetermination motivate researchers to design new experiments to differentiate two theories, when evidence is found that two hypotheses are more permanently underdetermined, the usual reaction has been one of "despair" and of discarding the theoretical problem as ill-defined. Grünbaum (2018) gives two historical examples of this situation in cognitive psychology: the debate between serial and parallel processing theories of attention and executive control in the 70s and 80s

(Anderson, 1978), and the debate about mental imagery in the 90s (Ganis and Schendan, 2011).

While theory underdetermination is a serious and concrete problem for theory development, not all philosophers of science have a pessimistic outlook on this issue. Perhaps the most famous argument against the severity of theory underdetermination is provided by Laudan and Leplin (1991). These authors indicate that there are several ways in which a theory can obtain evidential support over other, competing theories. For example, a theory may be preferred because it is related to other theories that have supporting evidence from a different set of observable phenomena. In addition, the development of new technologies to obtain data may expand the set of observable phenomena, dissolving the equivalence of two theories. I will discuss later an approach to theory development that embraces both of these approaches to solve issues of underdetermination, but first I will provide some concrete examples of how theory underdetermination currently affects learning theory.

3. Theory underdetermination problems in learning theory

A problem that has become increasingly evident in recent decades is that theory underdetermination problems are pervasive in computational cognitive modeling in general (Anderson, 1978, 1990; Ashby and Helie, 2011; Jones and Dzhafarov, 2014; Navarro et al., 2004; Townsend, 1990, 2008; Van Zandt and Ratcliff, 1995; White and Poldrack, 2013), and in CNS models of learning in particular. That is, after extensive study and testing of computational cognitive models of behavior, we have discovered that many of them can mimic each other's predictions, and in some cases they may make identical predictions. In addition, many models can offer multiple explanations for the same pattern of data, each in terms of a different cognitive process. In this case, it is impossible to learn which of the processes caused a particular data pattern.

This issue seems to undermine the most important goals of cognitive modeling. Some key theoretical questions that cognitive models aim to answer are about representations and operations over those representations: Are representations supporting associative learning elemental or configural? Is learning supported by error-driven strengthening of associations or by probabilistic inference? Is improvement in a task due to perceptual or decisional processes? Are interactions between two stimulus properties happening at the level of perceptual or decisional processing? Is information processing in a given task sequential or parallel? Note that all of these questions offer a binary decision between two theoretically interesting possibilities (e.g., elemental vs. configural representations, association vs. inference, perceptual vs. decisional, sequential vs. parallel). Such questions are at the core of how cognitive modeling has proceeded since its inception. Additionally, all of these questions have proven to be difficult to answer, due to problems of theory underdetermination (e.g., Anderson, 1978; Ghirlanda, 2015; Jones and Dzhafarov, 2014; Navarro et al., 2004; Silbert and Thomas, 2013, 2017; Townsend, 1990; Van Zandt and Ratcliff, 1995). These problems are much more concrete than how underdetermination is usually discussed by philosophers of science. They go beyond vague claims of “yet unknown alternative theories,” showing instead that our current alternative explanations of behavior are very difficult–in some cases impossible–to tell apart using behavioral data alone.

Why are such underdetermination problems, affecting key theoretical issues, so pervasive in cognitive theory? One possibility is that this has to do with the nature of cognitive theory itself, as expressed by Skinner's critique of the CNS approach (Skinner, 1938). More specifically, all computational cognitive models essentially specify a function that maps input to output as a set of cognitive operations. We know from work in computer science that the exact same function can be implemented using many different operations, and there is no way to discriminate among such implementations from information about inputs and outputs alone (Anderson, 1990). This problem is compounded by the fact that most cognitive models do not work with a physical, measurable representation of the input, but rather start with assumptions about input representation (Grünbaum, 2018). This leads to the development of competing theories that differ both in the format of input representations and the cognitive operations that map such inputs to outputs. Deciding between such theories is difficult, as a behavioral pattern that seems to contradict the input representation assumed by a theory can be explained away through assumptions of cognitive operation, and vice-versa (for examples, see Grünbaum, 2018).

In favor of the idea that theory underdetermination problems might be particularly severe in cognitive modeling, several philosophers of science have voiced complaints that strong cases of theory underdetermination, in which two theories are proven to be empirically equivalent, seem rather rare and are circumscribed to specific areas of study (Earman, 1993; Lyre, 2011; Stanford, 2001; Turnbull, 2018). In comparison, the relatively young area of cognitive modeling has already accumulated many examples (e.g., Anderson, 1978; Ghirlanda, 2015; Jones and Dzhafarov, 2014; Navarro et al., 2004; Silbert and Thomas, 2013, 2017; Townsend, 1990; Van Zandt and Ratcliff, 1995) and a rich discussion around them (Anderson, 1990; Ashby and Helie, 2011; Love, 2015; Navarro et al., 2004; Townsend, 2008; Van Zandt and Ratcliff, 1995; White and Poldrack, 2013).

Anderson (1990) suggested that underdetermination problems were particularly severe for models proposed at an implementational level of analysis, such as connectionist models (see also Massaro, 1988), which do not truly describe neurobiological mechanisms, but rather the way in which algorithms might be implemented through neurally inspired architectures. Such models are flexible enough to implement the same set of cognitive operations using radically different architectures and internal representations, which explains the controversy between localist and distributed representation within connectionist modeling (Page, 2000).

The arguments advanced by these authors can be made more concrete if we come back to the prototypical CNS model of Pavlovian conditioning presented in Fig. 1. Even if this simple architecture is used, formalizing a model requires describing rules at each stage of processing: (1) rules to represent the physical presentation of cues and outcomes as activation patterns across the sensory units, (2) rules that determine what connections exist between nodes and whether those connections are fixed versus modifiable, (3) rules to change the modifiable connections as a function of experience with environmental events, (4) rules by which the activation of sensory units together with their associative strength produce a level of activation in the response unit, and (5) rules that transform the activity of the response unit into a measurable behavioral response (i.e., the conditioned response). Rules in (1) represent what we called earlier the format of input representations, whereas rules in (2)-(5) represent cognitive operations that transform inputs into outputs.

As hinted by Anderson (1990), the problem can only get worse when one considers more complex connectionist models, which usually include one or more hidden layers and many output units. For such models, rules in (1)-(3) must be described for each additional layer, and it is difficult to predict how the representations learned by the model in hidden layers will look like after exposure to different environmental contingencies. These models are not only more prone to problems of underdetermination than traditional two-layer CNS models, but also much more obscure and difficult to interpret.

Examples of theory underdetermination (i.e., model mimicry and unidentifiability) are so numerous in cognitive modeling that they cannot possibly be covered here (e.g., Anderson, 1978; Ghirlanda, 2015; Jones and Dzhafarov, 2014; Navarro et al., 2004; Silbert and Thomas, 2013, 2017; Townsend, 1990; Van Zandt and Ratcliff, 1995). Instead, in this section I will provide the reader with a detailed explanation of key examples that have arisen in learning theory and

related areas in the last couple of decades. The examples cover different types of underdetermination problem. In the first example, two competing models make the same predictions for a restricted set of behavioral phenomena. In the second example, two competing models make the same predictions for the full set of behavioral phenomena addressed by the models. In the final example, sets of parameter representing different explanations of behavior are unidentifiable within a single model.

3.1. Error-driven versus probabilistic learning of associations

We start with an example in which two competing models make the same predictions for a restricted set of behavioral phenomena. The two models have been proposed as alternative answers to the theoretical question: Is learning supported by error-driven strengthening of associations or by probabilistic inference? These can be considered alternative views of how the parameters v_i in Fig. 1 are learned from experience.

Traditionally, the Rescorla-Wagner model (Rescorla and Wagner, 1972) has been taken by researchers as the main source of predictions and explanations from an associative perspective. The Rescorla-Wagner model states that the change in associative strength between a cue C_i and an outcome O in a learning trial, or Δv_i, is given by:

$$\Delta v_i = \alpha_i \beta \left(\lambda - \sum_j v_j \right), \tag{1}$$

where α_i and β are learning parameters determined by the salience of the cue i and the outcome, repectively, λ is the maximum amount of associative strength that the cue can acquire, and $\sum_j v_j$ is the sum of associative strength of all the cues presented during the trial. Two fundamental learning principles embedded in this model should be highlighted. First, the model proposes an error-correction rule where the amount of learning in each trial is proportional to the difference between the outcome expected on the basis of previous learning (i.e., $\sum_j v_j$) and the outcome actually experienced at the end of the trial (i.e., λ). This allows reproducing the typical shape of learning curves in human causal learning. Second, the prediction error is computed over all the potential causes present during a trial, or $\sum_j v_j$, which helps to account for stimulus competition phenomena such as blocking (see Soto, 2018).

Besides its success explaining Pavlovian conditioning phenomena, the Rescorla-Wagner model can also explain a considerable amount of experimental observations in human causal learning (for reviews, see Allan, 1993; Young, 1995). However, a group of experimental observations in this area, collectively known as retrospective revaluation, cannot be explained by the Rescorla-Wagner model and most other traditional associative learning models. In a backward blocking experiment (e.g., Chapman, 1991; Shanks, 1985), for example, participants are first presented with a number of trials involving a compound of two potential causes, AB, followed by the outcome. The result is that participants judge that both A and B have some ability to cause the outcome. In a second set of trials, participants are now presented with A alone followed by the outcome, which leads to judgments of lowered causal strength for B, even when this cue was never trained in these latter trials. This and other effects found in causal learning cannot be explained by traditional associative theories, which assume that experience only modifies connections between stimuli that are physically present in a trial.

Although some researchers took retrospective revaluation as the main evidence against an associative explanation of human causal learning, modifications of models in this tradition (e.g., Aitken and Dickinson, 2005; Van Hamme and Wasserman, 1994) can give an account of learning about events in their absence. These models keep many of the relevant features of traditional associative theories, including the assumptions that learning is associative and driven by error, as in Eq. (1).

A competitor to such associative models are rational probabilistic theories of learning (e.g., Allan, 1980; Cheng and Novick, 1992; Cheng, 1997; Soto et al., 2014; Tenenbaum and Griffiths, 2001). All of them share the underlying idea that learning agents are intuitive statisticians, who keep a record of event frequencies and use them to compute probabilities and other statistics describing the causal relation between two events. One of them is the probabilistic contrast model (Cheng and Novick, 1992), developed to predict how a rational observer would judge covariation between a cause and an outcome. The probabilistic contrast for a potential cause C_i, considering a background of other potential causes B, is given by the following equation:

$$v_i = P(O|C_i, B) - P(O| \neg C_i, B), \tag{2}$$

where P represents probability, the symbol | should be read as "conditional", and ¬ as "not." Thus, the probabilistic contrast measures the difference between the probability of observing the outcome after a cue has been presented in some background B, or $P(O|C_i, B)$, versus the probability of observing it when the background is presented alone, or $P(O| \neg C_i, B)$. The predictions of the model depend strongly on the choice of the focal set B. Given the right choice of focal sets, the model predicts correctly both stimulus competition and retrospective revaluation effects.

It has been found that the predictions of the Rescorla-Wagner learning rule approach the predictions of the probabilistic contrast model when it reaches equilibrium after long training with the same event contingencies (Chapman and Robbins, 1990; Cheng, 1997; Danks, 2003). In fact, the result seems more general than this, as error-driven associative learning models can be developed that approximate the predictions of other probabilistic theories (Danks et al., 2003; Yuille, 2005).

What this means is that any prediction made by considering learning as the result of probabilistic inference, can also be made by considering learning the result of associative error-driven learning. The reverse is not true, as the Rescorla-Wagner and other associative models can make predictions about phenomena that are outside the scope of probabilistic models, such as trial order effects (see Allan, 2003). However, the price for that explanatory power are several free parameters not present in probabilistic models. For the specific empirical data that probabilistic models are designed to address, it is difficult to decide whether learning is driven by error or by probabilistic inference, as attested by the survival of the probabilistic approach up to this day.

One could consider this a case of underdetermination that is "trivial," in the sense that the two theories thus related are actually the same theory, with the probabilistic contrast model being a way to characterize the long-range behavior of the Rescorla-Wagner model. However, note that the two models arrive at the same behavior under radically different assumptions about the processes that underlie learning.

3.2. Configural and elemental representation in associative learning

In this section, we will cover an example of two competing theories that make the same predictions for the full set of behavioral phenomena that they address. That is, no past or future behavioral experiment can truly decide between the representations proposed by the two theories. The two theories have been proposed as alternative answers to the theoretical question: Are representations supporting associative learning elemental or configural? These can be considered alternative views on how environmental cues are represented by sensory units in Fig. 1. This question has a long history in associative learning research (for a review of the early literature, see Kehoe and Gormezano, 1980), and offers perhaps the clearest example of theory underdetermination within learning theory.

Due to its simplicity, perhaps the best-known compound generalization design is the summation experiment (Aydin and Pearce, 1995,

1997; Kehoe et al., 1994; Perez et al., 2018; Rescorla, 1997; Soto et al., 2009; Rescorla and Coldwell, 1995; Whitlow and Wagner, 1972). In a summation experiment, two stimuli A and B are both independently associated with a given reinforcer. A summation effect is found when the conditioned response to the compound AB is stronger than the conditioned response to each of the elements A and B.

Elemental models of associative learning assume that each cue is represented as a separate entity, and that the total associative strength of a compound is the sum of the associative strengths of its components. That is, a compound AB is represented by two separate sensory representations, *a* and *b*, each of them acquiring an independent connection to the outcome through associative learning.

The Rescorla-Wagner model (Rescorla and Wagner, 1972), described in the previous section, is the most popular *elemental* theory of associative learning. Due to this popularity, two common mistakes are made regarding what theorists mean when they talk about an elemental model.

First, many elemental models include *unitized* representations of compounds, in addition to any representation of the individual components. For example, Rescorla and Wagner immediately proposed a unique-cue extension of their model (Wagner and Rescorla, 1972), graphically depicted in Fig. 2, which assumes a unique element or "configural cue" representing a particular compound. According to this theory, the presentation of a compound stimulus (AB) produces the activation of representations for each individual stimulus (*a* and *b*) and also a configural representation of the compound itself (*ab*). In a summation experiment, the unique-cue model predicts that activation of the response unit will be equal to the algebraic sum of associative strengths of each cue representation, including the configural cue; that is: $r(\mathrm{AB}) = v_a + v_b + v_{\mathrm{ab}}$.

Second, each sensory unit in an elemental model can itself be composed of a large number of representational elements, in what Wagner and Brandon (2001) called a *componential representation*. In the simplest case, each of these elements is either active (1) or inactive (0), and each active element can establish an association with active elements representing a second stimulus. This is depicted graphically in Fig. 3, where a sensory unit is composed of a number of elements, represented by small circles that can be either active (black circles) or inactive (white circles). The figure also shows that some of the active elements can themselves activate elements in a second representation, as a consequence of their learned connections. The Rescorla-Wagner model had a single element representing each stimulus, but more recent elemental models represent individual stimuli through collections of multiple elements (e.g., Harris, 2006; Harris and Livesey, 2010; McLaren and Mackintosh, 2002; Thorwart et al., 2012; Wagner, 2003).

Elemental models are usually contrasted to *configural* models of associative learning. Configural models are characterized by the assumption that stimulus compounds are processed as unique stimuli or

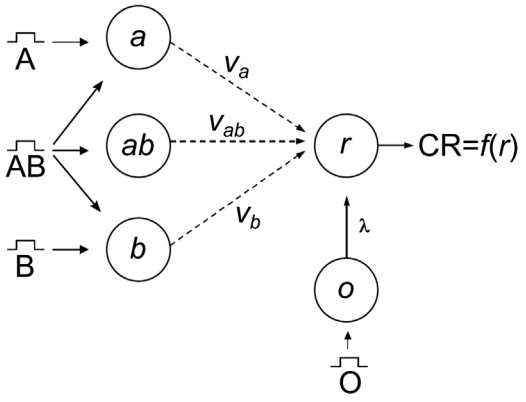

Fig. 2. Graphical representation of the unique-cue extension of the Rescorla-Wagner elemental model of associative learning.

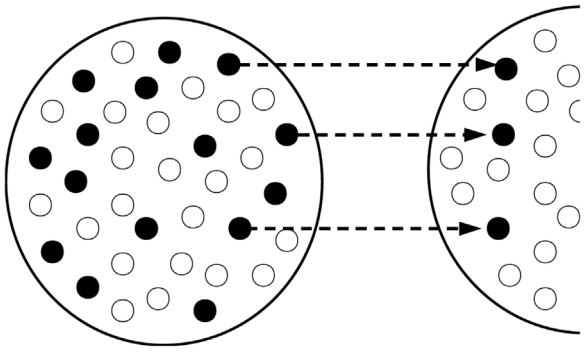

Fig. 3. Graphical representation of a componential representation, in which each stimulus representation (large circles) is composed of a number of elements (small circles) that can be either active (black) or inactive (white). Dotted arrows represent associations between the elements representing one stimulus and those representing a second stimulus.

configurations, distinct from their components. Thus, each novel compound is assigned its own representation and establishes its own association with an outcome. Still, two different compounds may activate each other's configural representations as the result of a process of configural generalization. That is, one compound can activate the configural representation of another as a function of their similarity.

The most popular configural model was proposed by Pearce (1987, 1994, 2002). At the heart of this theory is a distinction between the associative strength directly acquired by a stimulus (v_i) and that indirectly acquired through generalization (g_i). Generalized associative strength is determined by the similarity between two configurations, represented by S_{ij} (similarity between configuration *i* and *j*):

$$S_{\mathrm{ij}} = \frac{N_C}{N_i} \times \frac{N_C}{N_j}, \tag{3}$$

where N_C is the number of cues that are shared between compounds *i* and *j*, whereas N_i and N_j represent the total number of cues in compounds *i* and *j*, respectively. That is, the similarity between two compounds is determined by the proportion of cues in the first compound that are common to both, multiplied by the proportion of cues in the second compound that are common to both. These similarity values can be used to determine how the associative strength of previously-experienced configurations, indexed by *j*, is generalized to the presentation of configuration *i*:

$$g_i = \sum_j S_{\mathrm{ij}} v_j. \tag{4}$$

When Eqs. (3) and (4) are applied to a summation experiment, the prediction is different from that of the Rescorla-Wagner model. According to configural theory, when A and B have both independently acquired a strong associative strenght of $v_a = v_b = 1$, the generalized associative strength to compound AB is $g_{ab} = S_{ab\to a} v_a + S_{ab\to b} v_b = 0.5 \times 1 + 0.5 \times 1 = 1$. That is, an averaging of the associative strength of each individual cue.

Pearce's learning rule is very similar to that proposed by the Rescorla-Wagner model (see Eq. (1)):

$$\Delta v_i = \beta[\lambda - (v_i + g_i)], \tag{5}$$

where v_i is the associative strength directly acquired by configuration *i*, g_i is the associative strength generalized from other similar configurations, and other parameters are defined as in Equation (1). Pearce's learning rule does not include a parameter α, as it assumes that all possible configurations have equal salience.

Fig. 4 is a graphical depiction of the assumptions behind Pearce's configural theory, represented through the prototypical CNS model architecture presented earlier. Importantly, other configural models share the same overall parameterization and assumptions, but differ in

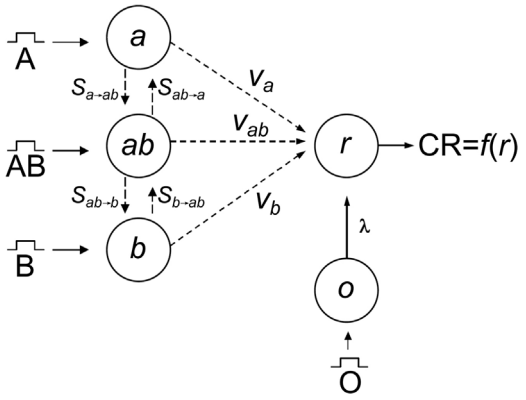

Fig. 4. Graphical representation of Pearce's configural model of associative learning.

the way in which similarities are computed, replacing Equation (3) with other equations (Estes, 1994; Kinder and Lachnit, 2003).

Comparison with Fig. 2 immediately reveals why elemental and configural models might be difficult to tell apart: it seems feasible to take the assumptions and parameterization of configural models, and change Eq. (3) so that the similarity values S_{ij} would allow to make the same predictions as the unique cue model of Fig. 2. All that is needed is a rule that sets $S_{a\to ab}$ and $S_{b\to ab}$ to zero, and $S_{ab\to a}$ and $S_{ab\to b}$ to one. While this would not constitute Pearce's specific flavor of configural theory anymore, it would still be a configural model that acts as the elemental unique-cue model in a summation experiment.

The opposite is also true, as argued by Wagner and Brandon (2001; see also Wagner, 2003, 2007). These authors implemented Pearce's configural theory using a componential elemental stimulus representation, as shown in Fig. 5. The key is to assume that when two (or more) cues are presented in compound, they mutually inhibit some proportion of the elements in the other cue's representation, so that the total number of elements is kept constant. The proportion of elements representing a cue that is kept active when that cue is presented in a compound is equivalent to the similarity parameter in Pearce's model.

In sum, theoretical exploration led to the discovery that it is possible to develop elemental models which behave like Pearce's configural model (e.g., Harris, 2006; Wagner, 2003, 2007), and configural models that behave like Rescorla and Wagner's elemental model (Kinder and Lachnit, 2003). From a behavioral point of view, one can think of elemental and configural models as ways of implementing different generalization rules. What is evident is that similar generalization rules can be implemented using either of the two forms of representation. Once this was discovered in the 2000s, interest seemed to shift toward

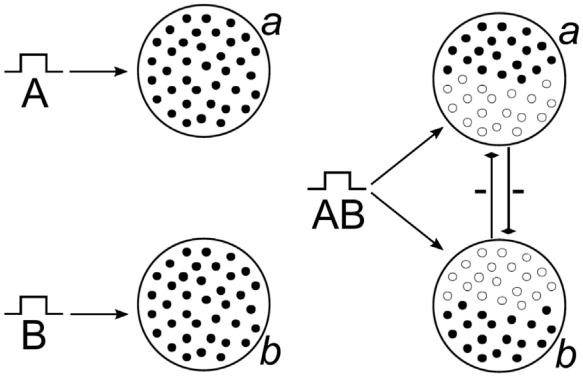

Fig. 5. Representation of stimuli A and B when presented separately (left) or in compound (right), according to the inhibited elements model developed by Brandon and Wagner (2001), as an elemental version of Pearce's configural theory.

trying to determine under what circumstances one form of representation would be favored over the other. Several models were developed, proposing mechanisms underlying the shift (Harris, 2006; Harris and Livesey, 2010; Soto et al., 2014, 2015a). The similarity hypothesis (Perez et al., 2018) was particularly popular among such models, proposing that elemental generalization (e.g., summation) should occur with very dissimilar stimuli, but configural generalization (e.g., averaging) should occur with very similar stimuli.

Thus, associative learning theory kept doing business as usual, expanding the theoretical edifice of the CNS, while at the same time simply ignoring the original question: Are compound representations elemental or configural? Perhaps the hope was that, if it kept going like this, the theoretical expansion would eventually yield a clear winner. It was in this scenario that in 2015 a paper came out that put the final nail in the coffin of the elemental/configural controversy. In a theoretical *tour de force,* Ghirlanda (2015) showed that elemental and configural models in general are indistinguishable through behavioral data, as it is always possible to create an elemental model that implements a set of configural generalization rules, and it is always possible to create a configural model that implements elemental generalization rules. Ghirlanda's general proofs of these statements are beyond the scope of this paper, but in essence what he has proven is that our previous insight that the elemental unique-cue model can be implemented using the configural architecture and parameterization of Fig. 4, is in general true for any elemental model. Vice-versa, he has also proven that Wagner and Brandon's (2001) insight that Pearce's configural theory can be implemented using a componential elemental architecture and parameterization as in Fig. 5, is in general true for any configural model.

The conclusion is clear: we could *never* solve the issue of whether the stimulus representation in associative learning is elemental or configural using behavioral data alone, at least not in the way in which the problem has been formulated up to this point. It is always possible to take an elemental theory and implement it using the language of configural theory, obtaining a model in which configurations are the representations associated with an outcome, as in Fig. 4. Conversely, it is always possible to take a configural theory and implement it using the language of elemental theory, obtaining a model in which elements are the representations associated with an outcome, as in Figs. 2 and 5 . The elemental/configural controversy and the models that started it have had a tremendous heuristic value, as much research has accumulated in the last 30 years surrounding this issue. However, it is undeniable that not having an answer to the question that guided all this research, and actually concluding that *we can never have an answer*, is a rather unsatisfactory state of affairs.

3.3. Beyond traditional CNS models: Perceptual versus decisional interactions between stimulus components

In this final example, I wanted to address the case in which sets of parameters representing different explanations of behavior are unidentifiable within a single model. These problems are very common in mathematical modeling, but I do not know of any examples of unidentifiability problems like this within learning theory, at least not including a formal treatment with proofs of the problem's generality. However, I suspect that this is not due to the absence of such problems in learning theory. Rather, it is likely a consequence to the way in which modeling is approached in this area of research: by contrasting qualitative patterns of behavior of two or more competing models, rather than fitting individual models to data.

This approach is so widespread in learning theory that it is a common practice to skip the formalization of some aspects of a model. In particular, many models of associative learning do not formalize a function mapping the output of the model, which is usually in units of associative strength, and the behavioral conditioned response. This function, which is described as $CR = f(r)$ in Fig. 1, is in most cases

implicitly assumed to be a simple identity function. Some nonlinear choices for $f(r)$, which would change most models' predictions, are absolutely necessary for the appropriate statistical description of discrete behavioral measures from the continuous variable r. Such discrete measures are common in learning research, including key pecks, eyeblinks, lever presses, and button-presses. Fitting models to such data would require making the explicit choice of a nonlinear function for $f(r)$, to link output in units of associative strength to the probability of a given response.

One area of research in which this is commonly done is psychophysics. In particular, signal detection theory (Green and Swets, 1966; Macmillan and Creelman, 2005) assumes that a stimulus presentation produces an internal continuous variable, representing sensory evidence. This variable is similar to r[1], but it also includes added random variability, assumed to be due to internal sources of noise. The sensory evidence variable is transformed into a discrete choice (e.g., stimulus absent = 0 vs. stimulus present = 1) by a decision process, implemented by comparison with a threshold τ:

$$f(r) = \begin{cases} 0, & r < \tau \\ 1, & r \geq \tau \end{cases} \quad (6)$$

Signal detection theory is extraordinarily successful in part because of its ability to dissociate between perceptual and decisional factors influencing detection and discrimination behavior. There is also a number of successful applications of signal detection theory to the study of stimulus discrimination and generalization in the learning literature (Allan et al., 2008; Alsop, 1998; Blough, 1967, 2001; Siegel et al., 2009; Soto and Wasserman, 2011; Terman and Terman, 1972).

Here I will describe an underdetermination problem that I have encountered in my own work with signal detection theory models. I will review an extension of signal detection theory that is used to answer the theoretical question: Are interactions between two stimulus properties happening at the level of perceptual or decisional processing?

The multidimensional version of signal detection theory, named general recognition theory (GRT; for reviews see Ashby and Soto, 2015; Soto et al., 2017), allows researchers to study how two or more stimulus components interact during processing. According to GRT, two stimulus components are *separable* when changes in one of them do not affect processing of the other. The theory thus provides a formal definition of configural representation, in the form of failures of separability. Such formal definition and the model-based experimental designs and analyses provided by GRT have been very useful to study topics such as configural encoding of faces and other visual objects (Mestry et al., 2012; Richler et al., 2008; Wenger and Ingvalson, 2002). There have also been attempts to link the dual concepts of separability/integrality and elemental/configural processing (Lachnit, 1988; Soto et al., 2014, 2015a) but, to the best of my knowledge, GRT has never been applied to study configural encoding in associative learning.

The main contribution of GRT to the study of configurality is the proposal that interactions between two stimulus components (i.e., lack of separability, or integrality) can occur both at the level of perceptual representations and at the level of decisional strategies. *Perceptual separability* refers to the case in which representation of a stimulus component does not depend on variations in another stimulus component, whereas *decisional separability* refers to the case in which the behavioral response rule related to a stimulus component does not depend on variations in another stimulus component. The main goal of modeling and experimental design within this framework is to dissociate between these different sources of component interaction.

A concrete example can clarify the distinction between perceptual and decisional separability. Suppose that an animal is trained to give a response in the presence of stimulus A, and is tested with a new compound AB. This design was first proposed by Pavlov (1927) and it usually results in lower responding to AB than to A, in what is called an *external inhibition* effect. According to an elemental analysis, responding to AB should be the same as responding to A, because the representation of A does not change in the presence of B and thus its associative strength transfers perfectly to the new compound. Looking at Fig. 2, the presentation of AB fully activates the representations *a, b* and *ab,* with the result that all the associative strength of A, v_a, is transferred to the compound. Thus, an elemental model proposes that A is perceptually separable from B, so no external inhibition should occur. According to a configural analysis, responding to AB should be different than responding to A, because the configural representation of AB is unique and only part of the associative strength of A will transfer to it. Looking at Fig. 4, the presentation of AB directly activates only the new configural representation *ab*. However, due to generalization, the configural representation *a* is partially activated as a function of $S_{ab \rightarrow a}$. Thus, a configural model proposes that A is not perceptually separable from B, because it is represented by the full activation of *a* when alone, but only a partial activation of *a* when B is also present. The prediction is that an external inhibition effect should occur.

Both the elemental and configural analyses assume that the associative strength of A can be directly assessed through a behavioral response, through the implicit assumption of an identity function $f(r) = r$. GRT, on the other hand, does not make that assumption. Instead, it assumes a response function like that proposed in Eq. (6). In addition, GRT assumes that this decision process itself can be affected by the presence or absence of B. There would be two values of the threshold τ, one for the case in which B is absent, or $\tau_{A\neg B}$, and another for the case in which B is present, or $\tau_{A \wedge B}$. Decisional separability holds when both values of τ are the same, it fails if they are different.

GRT was first proposed more than thirty years ago (Ashby and Townsend, 1986), and since then it has been applied to dozens of studies (for a list, see Ashby and Soto, 2015). However, it was not until this decade that Silbert and Thomas (2013) discovered that, in the most commonly-applied GRT model, perceptual and decisional separability cannot possibly be dissociated. That is, the same behavioral pattern can be found in the presence of perceptual separability and the absence of decisional separability, and vice-versa. An intuitive understanding of this result can be obtained by using the external inhibition example. Imagine that we assume the Rescorla-Wagner model of Fig. 2, so we know that the representation of A is perceptually separable from B. In this case, the output of the model in terms of r is the same when A is presented or AB is presented. However, GRT assumes that the decisional threshold τ can have different values in the presence and absence of B. If that was the case, one would observe different proportions of responses to A and AB, even under the elemental model. On the other hand, imagine that we assume Pearce's configural model from Fig. 4, so we know that the representation of A is not perceptually separable from B. In this case, the output of the model in terms of r is different when A is presented or AB is presented. However, again one could choose the values of the two decisional thresholds $\tau_{A\neg B}$ and $\tau_{A \wedge B}$, so that the proportion of responses to A and AB are equivalent, even under the configural model. This would be equivalent to assuming that the presence of B changes how much sensory evidence (or associative strength) is necessary to make a response to A, which seems like a reasonable assumption. For example, an animal could increase its general tendency to respond in the presence of a novel stimulus (B), as a way to foster exploration, or reduce that tendency as a way to show caution. In sum, the dependence of the threshold τ on the context provided by B is an explicit acknowledgement that increments or decrements in responding during generalization tests could be due to factors other than stimulus representation. As indicated earlier, such assumption has proven useful

[1] Indeed, a simple linear observer model from psychophysics, which is an extension of signal detection theory, is similar to the Rescorla-Wagner model but with C_i representing stimulus components, v_i fixed rather than modifiable (all v_i usually represented together in the template vector **v**), and without input from an outcome to the response unit. In addition, the variable r is assumed to be perturbed by additive Gaussian noise.

in past applications of signal detection theory to learning research (Allan et al., 2008; Alsop, 1998; Blough, 1967, 2001; Siegel et al., 2009; Soto and Wasserman, 2011; Terman and Terman, 1972).

What Silbert and Thomas (2013) discovered is that problems of identifiability between perceptual and decisional processes are always present when using the most common GRT model. That is, it is never possible to know the status of perceptual and decisional forms of separability (or in our case, configurality). This identifiability issue can be solved by re-parameterizing the model (Soto et al., 2015b), but only at the cost of making assumptions about the stimulus representations that some theorists find too strong (Silbert and Thomas, 2017).

While these parameter identifiability problems are less serious than the model equivalence problem discussed in the previous section, they also seem more pervasive in cognitive modeling. Although the specific example covered here does not come specifically from CNS modeling, it is close enough to such models that we could describe it using Fig. 1. It is clear that parameter identifiability problems could be easily found in learning theory, if we use fully-formalized models including a perhaps nonlinear $f(r)$. As indicated above, I suspect that we have not discovered these problems in CNS models only because our approach to modeling has stopped us from doing so.

4. Computational cognitive neuroscience modeling: constraining theory with neurobiology

What can be done when theory underdetermination seems to get in the way of scientific progress? One common answer is to despair and discard the original theoretical problem (e.g., configural vs. elemental representation of compounds) as ill-defined (Grünbaum, 2018). Another common answer is to hold an instrumentalist view of scientific theory, in which all models are considered as formal descriptions of the transformation of certain inputs into an output, and the actual unobservables proposed by the theory are deemed irrelevant. Both of these answers implicitly assume that the key theoretical issues that we have been pursuing in the last decades are rather trivial (e.g., configural vs. elemental representation, associative vs. inferential learning, etc.). This might be satisfactory to some, but it is unacceptable for those of us who deeply care about such theoretical questions, and do not want to throw the theoretical baby out with the underdetermination bathwater.

As indicated earlier, not all philosophers of science agree on the seriousness of the problem of theory underdetermination. Laudan and Leplin (1991) famously proposed that, even when it has been shown that two theories are empirically equivalent for the set of observable phenomena that they address, scientific advances may change this state of affairs. In particular, they argue that new sources of knowledge and technology could provide evidential support for one of the theories over the other. One way in which this can happen is if only one of the theories provides links to a third well-supported theory that deals with a different body of observable phenomena. For example, a psychological theory could be preferred over another because it provides links to well-tested neurobiological theories, or theories in the social sciences. Another way in which this can happen is through development of technologies to measure the unobservable entities proposed by the two theories. This is usually not a possibility for cognitive models, which propose entities that are unobservable *in principle*. How can one observe a representation, or an algorithm? One can surely infer them from behavioral data, under the correct assumptions, but direct measurement of the entities proposed by cognitive theory seems implausible.

Here I argue that a novel approach to the development and implementation of cognitive models would facilitate solving problems of theory underdetermination in the way proposed by Laudan and Leplin (1991). Computational cognitive neuroscience (CCN; Ashby and Helie, 2011; Ashby, 2018) proposes that cognitive models should be constrained using neurobiological data. Just like CNS and other cognitive models, the main goal of CCN modeling is to explain behavioral phenomena. In addition, the target explanation is clearly cognitive, in the sense that it should involve representations of inputs and operations on those inputs to produce a behavioral output. However, the model is built using knowledge from experimental and computational neuroscience, with two consequences. First, by using the language of computational neuroscience to implement cognitive models, one immediately links the resulting model to a large theoretical edifice supported by a body of observations from neuroscience. In addition, using such language allows one to include in the model relevant knowledge from experimental neuroscience. This is the first type of knowledge that Laudan and Leplin (1991) argued can provide evidential support for underdetermined theories. Second, implementing cognitive models using the language of computational neuroscience makes the model completely observable *in principle*. That is, it opens the possibility to observe and study each part of the model. Thus, new neuroscientific discoveries can push an updating or a complete reorganization of the model. While the same algorithm can be implemented in many different ways, the reverse is not true, and a change in implementation usually means a change in the computational mechanisms that produce a certain behavior.

This second point is not novel. Encouraged by the development of neuroimaging technologies, several authors have proposed that one way to constrain cognitive models is through neuroscientific data (Love, 2015; Townsend, 2008; White and Poldrack, 2013) and we have already seen some success stories (Ganis and Schendan, 2011). What is new about the CCN approach is that it proposes that the interaction between cognitive theory and neuroscience should work both ways: not only with cognitive models providing predictions for neuroscientific experiments, but with known neuroscientific results providing constraints during cognitive model development.

An important point is that if two models are committed to different descriptions of the brain, then theory underdetermination at the behavioral level is not a serious problem anymore. Instead, it is at worst a form of *transient* underdetermination, which can be solved with future data obtained with a precise-enough instrument (Grünbaum, 2018). This is a much better scenario than having important theoretical questions that cannot be answered. The important point is that, by making a commitment to offer some neurobiological detail, one expands the different types of data that will be accepted. This expansion of phenomena, as indicated by Laudan and Leplin (1991), can be proposed strategically to solve known underdetermination problems with CNS models.

CCN models do not constitute a wild departure from the traditional CNS models from associative learning theory, and one could argue that many learning theorists have already implicitly adopted the goals of CCN modeling. For example, CCN models do not require us to abandon associationism, but rather to promote a neurobiological version of associationism, similar to that proposed by Hebb (1955). Most animal learning researchers would not frown upon this, as Pavlov himself seemed interested in conditioning as a way to study associations learned in the nervous system, rather than in some abstract mental space (Pavlov, 1927).

What, then, is new about the CCN approach? The main departure from traditional approaches is that CCN modeling is committed to seriously consider data and theory about neural computation during model development. While CNS models usually make claims about "biological plausibility," in practice most models include features that are known to be incompatible with current neuroscientific knowledge (Ashby and Helie, 2011). In contrast, CCN models are not just neurally inspired, but incorporate real neuroscientific knowledge about brain areas involved in a particular behavior, their connectivity, their characteristic patterns of neural activity, synaptic learning mechanisms, etc. This means that CCN models can and should make predictions about both behavioral and neural data, providing the foundation for a true integrative research approach.

4.1. How to build a CCN model

This section provides a short summary of how CCN models are usually built. In its more general form, a CCN model is simply a computational cognitive model that explains some behavior while at the same time being constrained by neurobiological data. Defined this way, the approach to build such a model is quite flexible, as each aspect of the model can be built with different levels of neurobiological realism. After all, the main goal is not to obtain a neurobiologically realistic model, but rather a model that can be understood at the level of cognitive computation, but is constrained by neurobiology.

One of the few groups who have explicitly described principles for model development are Ashby and collaborators (Ashby and Helie, 2011; Ashby, 2018), who propose the following three: do not contradict neuroscientific knowledge (the *neuroscience ideal*), do not include neuroscientific detail that does not add explanatory value (the *simplicity heuristic*), and do not modify the model architecture across simulations (the *set-in-stone ideal*). The reasoning behind each of these principles is better explained below, by going through the steps that one would usually take to build the model. However, it is important to explain why it is wise to follow these principles in the first place: they represent a balance between constraining theory with neurobiology (the *neuroscience* and *set-in-stone* ideals), while at the same time focusing on the goal of cognitive modeling and explanation (the *simplicity heuristic*).

Most CCN models take the form of neural networks, being very similar to traditional connectionist models. The general recipe to build a CCN model involves three steps: (1) define the processing units and what they do, (2) define their connectivity, and (3) relate the model activity to behavior and other observable measures.

4.1.1. Define the processing units and what they do

The first step during development is usually defining the model's units and what they do. Here at least three choices are available from the computational neuroscience literature: spiking neurons, population models, and microcircuits.

Models of spiking neurons are those that attempt to reproduce the dynamic behavior of single neurons in the brain. We would choose spiking neurons as units only if we want the model to explain single-neuron recording data or when the temporal dynamics of single neurons offer mechanistic explanations for behavioral data. In general, we can generate such predictions by using so-called “point-neuron” models, which model a single point in a neuron's membrane. Thus, we avoid modeling the propagation of potentials through the neuronal dendrites, soma, and axons. Many of such models have been proposed in the literature, so an important question here is which one to choose. At one extreme, biophysically-accurate models such as the famous Hodgkin-Huxley model involve parameters that are directly related to measurable physical quantities. They can reproduce the behavior of many neurons in the brain, but at the cost of being complex and difficult to simulate. At the other extreme, very simple phenomenological models like the leaky integrate-and-fire (LIF) model are mathematically simple and easy to simulate, but they include so little biological detail that they can only roughly simulate the simplest behavior of real neurons. A number of options exist between those two extremes, and fortunately some of them involve adaptive extensions of the LIF model that, while being relatively simple and easy to simulate, are able to reproduce the dynamic behavior of neurons as well or better than the most complex biophysical models (Brette, 2015; Brette and Gerstner, 2005; Gerstner and Naud, 2009; Izhikevich, 2004; Naud et al., 2008). Following the *simplicity heuristic*, if our goal is to simulate single-neuron recordings, then such adaptive LIF models should be chosen.

The fact that phenomenological single-neuron models can reproduce many different types of neuronal responses means that these models are highly flexible and dependent on parameter values. Paradoxically, this seems to generate models that are less rather than more constrained than a simple CNS model. Here is where the *set-in-stone* idea shows its usefulness. One should choose the model's parameters to make sure that it reproduces the dynamics of neurons in the region of the brain that is being simulated (e.g., through fitting to neurophysiological data). Once those parameters are fixed, they should not be changed to fit behavioral data or other observable measures.

Sometimes, we might decide that the spiking behavior of single-neurons is not very important for the model's computations, which can be implemented through more coarse variables such as firing rates or the activity of large populations of neurons. Two types of simplifications can be used in this case. First, one may simplify across space and model the activity of a population of neurons (for a review, see Deco et al., 2008). Several statistics can be chosen to describe the population, such as mean voltage or proportion of neurons spiking at a given time. This would be a good choice if the model dynamics within a trial are important to explain the behavioral data of interest (e.g., response times), or if one wants to make predictions of neural data for which such dynamics is important (e.g., local field potentials, electrocorticography, EEG, etc.). If neither behavioral or neural data warrant the development of a temporally-precise model, then the best choice is to simplify across time and model the firing rate of a neuron or the activity of a population as a variable that adopts a single value in each trial, and only changes across trials. Noise can be added to such a variable, which is usually assumed to be Poisson for the firing rates of single neurons, and Gaussian for populations of neurons.

Finally, one may use as a unit a model implementing a relatively complex computation, assumed to be the output of a circuit not explicitly modeled. For example, changes in the response of a V1 neuron to gratings that change in orientation could be modeled through a large network receiving inputs from thalamus (e.g., Zhu et al., 2009), as a filter whose inputs are image patches (e.g., Marcelja, 1980), or through a simple phenomenological model in which neural responses follow a bell-shaped function of orientation (Pouget et al., 2003). One more time, the *simplicity heuristic* tells us here that, unless the details of how V1 neurons get to fire the way they do are important to understand behavior, we should choose the simplest phenomenological model. Other examples of these models are those computing divisive normalization (Carandini and Heeger, 2012) and reward prediction errors (Suri, 2002), both of which assume that the computation is carried out by a circuit that is not explicitly modeled.

4.1.2. Define the network's connectivity

A second step is to define the network's connectivity. This is perhaps the most important point of departure from connectionism, as connectivity should not be left to the modeler to decide. Following the *neuroscience ideal*, the decision to connect two regions should be determined by the results of studies testing whether fibers effectively go from one region to the other (e.g., through tracing studies). When a connection is determined to exist, another issue to be solved is whether the synapses are excitatory or inhibitory. Here, again, the *neuroscience ideal* indicates that the decision should not be left to the modeler, but be constrained by neurobiological data. The determination of connections and their functional roles as excitatory or inhibitory determines the network's architecture. Finally, the *set-in-stone ideal* plays a very important role, and provides a departure from traditional connectionist modeling. Once the architecture has been decided based on neurobiological data, this architecture is fixed for all simulations of the model. That is, connections should not be added or subtracted only to fit behavioral data or other observable measures.

The final choice to make regarding connectivity is what synapses in the model are plastic, and how plasticity will be modeled. One could argue that plasticity is widespread in the brain, and thus most neurons should involve some form of plasticity. In practice, not all synapses are plastic in CCN models, as the *simplicity heuristic* is used to explore plasticity in only those synapses that are key for the model's explanation of behavioral phenomena.

As was the case for models of single neurons, there are multiple

models of synapses (for a review, see Van Rossum and Roth, 2010) and synaptic plasticity (e.g., Cooper and Bear, 2012; Kuśmierz et al., 2017; Morrison et al., 2008) to choose from, and this choice should be guided by the *simplicity heuristic*. However, the *neuroscience ideal* should bias us to choose models that do not contradict knowledge from neuroscience, rather than choosing arbitrary learning rules known to contradict principles of neural computation. Additionally, the *set-in-stone* ideal suggests that any free parameters should be fixed before simulations are performed.

It is important to note that some CCN models do not involve connectivity among units. Instead, such models propose that a neural population implemented in a given region is directly connected to behavior through some algorithm, like a "decoder" that "reads out" information about a variable and generates a behavioral response (e.g., Deneve et al., 1999; Seung and Sompolinsky, 1993). Such models have proven very useful to connect computational neuroscience to signal detection theory, and thus we will see an example of them later on.

4.1.3. Relate the model activity to behavior and other measures

At this point, the model would consist of a network of units whose activity could be measured directly. Because the final goal is to explain behavioral phenomena, CCN models must incorporate assumptions linking neural activity to behavioral measures. There are multiple ways of doing this, the simplest being to include a number of "decision neurons" and a decision rule that transforms their activity into discrete behavioral choices at a particular time. As indicated above, another option is to use a decoding algorithm (e.g., Deneve et al., 1999; Seung and Sompolinsky, 1993), perhaps together with a behavioral rule to convert the decoded variable into a discrete choice.

For the model to make predictions about indirect measures of neural activity, such as those obtained through functional MRI, it is necessary to build a measurement model connecting activity in the model and the actual observed variables. There are many such models and they vary greatly in complexity. They are beyond the scope of the current paper (for reviews, see chapter 3 of Ashby, 2011; Ashby and Waldschmidt, 2008; Henson and Friston, 2007), but it should be noted that these measurement models are not part of the actual CCN model. Instead, they are proposed to link the CCN model to observable variables, and therefore they are not subject to the three principles of CCN modeling presented above.

4.2. Re-visiting examples of underdetermination in learning theory

4.2.1. Error-driven versus probabilistic learning of associations

Would a CCN model of associative learning look more like error-driven learning algorithms or probabilistic inference theories? We do not need to guess the answer to this question, as several models of Pavlovian conditioning have been proposed in the computational neuroscience literature. These models involve different neural substrates for different forms of associative learning: the circuitry of the cerebellum for eye-blink conditioning (Dean et al., 2010; Lepora et al., 2010), the amygdala for fear conditioning (e.g., Carrere and Alexandre, 2015; Krasne et al., 2011; Moustafa et al., 2013), the basal ganglia for appetitive conditioning (Maia, 2009; Niv, 2009; Suri, 2002), and so on. However, they have in common that learning of associations between neural representations of cues and the outcome are modified by a three-term rule, which takes into account not only activity in pre- and post-synaptic neurons, but also modulation by a prediction error term, similar to the parenthetical term in Eq. (1).

One consequence of this is the expectation that prediction error signals should be observable in the brain during learning. Such prediction error signals are indeed found in many brain areas during learning, including during human causal learning situations like those that prompted the development of the probabilistic contrast model and similar rational models (Corlett et al., 2004; Fletcher et al., 2001; Turner et al., 2004). In particular, prediction error signals are found during retrospective revaluation procedures (Corlett et al., 2004), which offers evidence for modified associative models rather than probabilistic inference as the substrate for such learning phenomena.

An important advantage of constraining models by neurobiology is that different modelers tend to converge on similar architectures and mechanisms (Ashby and Helie, 2011; Ashby, 2018). When an area is involved in more than one form of learning, then all the models involved should share a set of computational principles implemented in the circuitry of that area. For example, models of category learning in the basal ganglia (Seger, 2008; Shohamy et al., 2008) borrow many of their features from models of appetitive conditioning and reinforcement learning (Maia, 2009; Niv, 2009; Suri, 2002). Most CCN models of category learning are quite similar to one another, and incorporate error-driven learning through dopaminergic input within the circuitry of the basal ganglia (Ashby et al., 1998; Soto and Wasserman, 2012b; Villagrasa et al., 2018).

4.2.2. Elemental versus configural representation

Would a CCN model of associative learning implement configural or elemental processing? Unfortunately, this is a question that has not yet been answered.

A simple strategy to start answering this question within a CCN framework would be to implement the representations in CNS models as neural encoding models (for reviews, see Pouget et al., 2003; van Gerven, 2017). That is, each of the "representational nodes" in Figs. 2-5 could be implemented using models of single neurons or neural populations that are selective to the same features as the unobservable sensory units in the original models.

Take as an example the original configural representation of AB presented by Pearce (Pearce, 1987, 1994, 2002) and the elemental version of the same representation envisioned by Wagner and Brandon (Wagner and Brandon, 2001), featured in Figs. 4 and 5, respectively. While both produce the same behavioral predictions in a summation experiment, Pearce's implementation would assume (i) neurons selective to A in a context-specific manner, being more active during presentation of A than AB, (ii) neurons similarly selective to B, and (iii) neurons selective to AB but also active when the components are presented in isolation.

On the other hand, Wagner and Brandon's implementation assumes (i) neurons selective to A in a context-invariant manner ("invariant A" neurons), being highly active when A is presented regardless of the presence of B, (ii) neurons similarly selective to B, (iii) "A not B" neurons that are highly active when A is presented alone but inactive in the presence of B (i.e., extreme context-specificity), and (iv) similar "B not A" neurons.

Are neural representations of compounds more similar to the configural or elemental versions of Pearce's model? The answer to this question seems to be "it depends." For example, Tsunoda et al. (2001) measured the activity of columns in inferior temporal cortex, while monkeys saw objects that were progressively simplified by subtracting features from them. The representation found was similar to what would be predicted from Wagner and Brandon's model: some columns seemed selective for a feature regardless of the presence of other features (i.e., "invariant A" neurons), whereas others showed extremely context-specific selectivity to a feature, so that they were active only when other features were absent (i.e., "A not B" neurons). Thus, a CCN model of compound conditioning with complex visual stimuli represented in inferior temporal cortex would favor an elemental representation over a configural representation, despite both theories offering the same behavioral predictions. Of course, in practice one would consider the results from more than one study, but the example works for didactic purposes.

Would we expect the same results if the components belonged to different sensory modalities? The most likely answer is "no," as in those cases stimuli are represented in different brain regions, which prevents local interactions between representations. Within a particular area of

sensory cortex, the response of a neuron to its preferred stimulus depends on the presence of other stimuli through a process of divisive normalization (e.g., see Carandini and Heeger, 2012). This mechanism might explain why some neurons show context-specificity, either extreme ("A not B" neurons) or mild ("A" representation in Pearce's model). However, such effects seem to be local, as the same neural response is unlikely to be affected by the presentation of a stimulus from a different sensory modality. This represents a neurocomputational implementation of the hypothesis that stimulus representations interact with one another more strongly (i.e., they are more configural) the more similar the stimuli are (Perez et al., 2018). Thorwart et al. (2012) have proposed an elemental model of associative learning that implements something similar to divisive normalization among its mechanisms. This is a case in which the transition from a CNS to a CCN implementation would be relatively straigthforward. This discussion shows an advantage of applying the CCN approach in learning theory in particular, as many models in this tradition have been developed with a decidedly neural "flavor".

One may argue that in many cases the data necessary to constrain a CCN model are not available and, given the slow pace of progress in neuroscience, they might not be available for a very long time. Although this could be true in some cases, CCN models offer the advantage of producing predictions that can be tested using either behavioral or neurobiological data. Although specific assumptions may remain untestable for a while, incorrect assumptions are more easily detected when they impact both behavioral and neural data.

4.2.3. Perceptual versus decisional interactions between stimulus components

Can we use CCN models to disentangle the study of sensory representation from decisional strategies? The answer is "yes." There are cases in which implementing a computational cognitive model as a CCN model allows one to more easily study the model's components in isolation. For example, we have recently proposed a CCN version of GRT that permits the study of separability and configurality in brain representations, and it formally links such properties to perceptual separability and similar concepts from the original theory (Soto et al., 2018). A schematic representation of the theory and its link to signal detection theory is presented in Fig. 6. In this example, an encoding model represents the level of anger in a face through four neurons, which together form an encoding model (Pouget et al., 2003; van Gerven, 2017). Each neuron in the figure is represented by a curve with a different color and has a preferred level of anger: the blue neuron has a preferred level around 2, the red neuron has a level around 4, and so on. The output of the encoding model is a pattern of neural activity, which includes neural noise that is added during processing in the visual system. Note that the level of anger is not explicitly represented in this pattern of neural activity. Rather, it is distributed across the responses of several neurons (four in this case). To make a decision on whether or not the face is angry, the brain must take that implicit information and make it explicit. To do this, it "decodes" level of anger from the pattern of neural activity. The brain's goal is to estimate the level of anger shown in the stimulus as precisely as possible, but because there is noise in the neural representation, the final perceptual evidence variable will also be noisy. That is, decoding produces a random perceptual evidence variable. This is exactly what signal detection theory assumes is the representation of anger, which allows us to link a theory of brain representation to GRT. Within this neural version of GRT, it is possible to define a form of separability for neural representations, named *encoding separability*, and study how it relates to perceptual separability. Such neural sensory representations can be studied directly, without the influence of decision processes, as no overt decision is necessary to get access to brain representations through the tools available in neuroscience, such as functional MRI. Thus, the kind of identifiability problem shown by the original GRT (Silbert and Thomas, 2013, 2017; Soto et al., 2015b) vanishes when one uses the neurocomputational version of GRT.

In addition, in some cases a clearly-specified CCN model might facilitate making inferences about neurocomputational mechanisms from behavior, an approach that has shown some success in psychophysics (e.g., Ling et al., 2009; Paradiso, 1988; Series et al., 2009). More specifically, the model presented in Fig. 6 also allows one to determine under what circumstances it is possible to make valid inferences about encoding and perceptual separability from behavioral measures (see Soto et al., 2018). Thus, rather than giving up on the goal of inferring mechanism from behavior, the CCN approach allows one to make stronger inferences as a result of the strong constraints placed on theory by neurobiology.

5. Potential pitfalls of the CCN approach

Despite the many advantages of applying CCN modeling in learning theory, it is important to highlight that the approach is not without problems. In this last section, I would like to point out what I believe are some potential pitfalls of the CCN approach.

5.1. Giving up CNS models

To be clear, I do not want to advocate for a complete abandonment of CNS modeling. I myself have published work that applies the CNS approach to the study of object category learning in pigeons (Soto and Wasserman, 2010b, 2014, 2012a; Soto et al., 2012), which serves to show how useful I consider this approach. Instead, I would like to propose that there is a natural progress in theory development from CNS to CCN models. Given that the CNS models available in learning theory are already "neurally inspired", I believe that implementing such learning theories as CCN models is in many cases both straightforward and beneficial. For example, Ed Wasserman and I translated our CNS model of category learning in pigeons into a neurocomputational model (Soto and Wasserman, 2012b), which served to explain the origin of a stimulus representation that was mostly assumed in the original model, and to link the theory to knowledge from comparative neuroscience, which in turn generates hypotheses about the evolution of category learning and clues about the usefulness of pigeons as animal models to study the mechanisms of category learning in people (Soto and Wasserman, 2014).

In sum, rather than giving up CNS models, I believe that we should focus on implementing them as CCN models and constraining them with knowledge from neurobiology.

5.2. Modest contributions

The contributions of a CCN model can be rather modest, compared to what we have become accustomed to from cognitive modeling. For

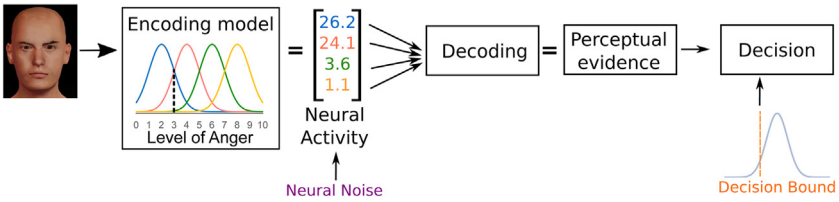

Fig. 6. A schematic representation of an encoding model representing face anger with four populations of neurons (blue, red, green, and yellow curves), whose neural activity is then used by the brain to decode evidence of anger. Adapted from Figure 2 in (Soto et al., 2018), copyright by the authors under license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

example, most models of Pavlovian conditioning can account for dozens of behavioral results, observed across several animal species and experimental preparations. This goal is still too far-fetched for CCN modeling. This is partly a consequence of the requirement to be explicit about the neural substrates of behavioral phenomena. For example, traditional CNS models can ignore the evidence indicating that different neural circuits are involved in eyeblink conditioning (Freeman, 2015) and fear conditioning (Herry and Johansen, 2014). After all, both circuits seem to implement a similar neurocomputational mechanism of learning (Gluck et al., 2001; McNally et al., 2011), and thus a single algorithmic theory can be proposed. On the other hand, a CCN model would require modeling each circuit independently to account for phenomena from each area of research.

In addition, one of the most important pitfalls of asking for "more neuroscience" in the explanation of behavior is that many complex behaviors emerge from the interaction of multiple neurocomputational mechanisms implemented in a variety of brain regions and circuits (Soto and Wasserman, 2010a, 2014). Cognitive models can directly describe such emergent computational principles, but most current CCN models cannot do the same, given our limited theoretical and technical tools.

In general, current CCN models are not models of a general class of behaviors or cognitive processes (e.g., a model of associative learning), but rather models of a specific experimental behavioral paradigm (e.g., a model of discrete-stimuli fear conditioning in rodents).

5.3. Forgetting about behavior

It is important to remember that the second "C" in CCN stands for "cognitive." This means that a CCN model must be able to explain behavioral phenomena. Without behavioral data to explain, a CCN model is just a computational neuroscience model. In addition, we must remember that our final goal should not be to stop at modest contributions like those described in the previous section. Rather, the final goal should be to obtain large-scale models involving multiple circuits, which can approximate behavior in naturalistic tasks.

Finally, the goal of a CCN model should be to explain real data from behavioral experiments, rather than focusing on the implementation of a particular computation that is only assumed to be performed by the brain. For example, a common question in computational neuroscience is how hypothetical neural populations and circuits can implement optimal statistical inference (e.g., Beck et al., 2008; Deneve et al., 1999). The resulting model adheres to the approach proposed here only if real behavior is well-described by such computations.

5.4. Proofs of concept

A common practice in computational modeling is focusing on "proofs of concept" showing that there is a kind of model (e.g., a neural network, or a Bayesian model) that can show a particular behavior. If the model shows the target behavior, it is said to be a good candidate for a mechanistic explanation. While this approach is quite common in computational modeling in general, and in neural network modeling in particular, I believe that it is far removed from the ideals of the CCN approach.

Proofs of concept are dangerously close to falling into the logical fallacy of "affirming the consequent". For example, we may know from simulation work with neural networks that a particular mechanism (e.g., deep learning) produces a representation with certain features. We then find that representations in some brain area have similar features, and conclude that this area implements a similar mechanism as the simulated neural network. There is no basis for such a conclusion, unless the model itself is built by constraining its mechanisms to reproduce key neurobiological properties of the target brain area. Unfortunately, neural networks are highly abstract models only inspired by neurobiology, and not constrained by it. Indeed, the constraints put into these models come mostly from engineering considerations, and biology is not considered seriously.

5.5. Model-based neuroimaging versus neurobiology-based models

Here, I have argued that one way to deal with underdetermination problems is to build neurobiology-based models. A different approach is to use tools from cognitive neuroscience, like functional MRI, to locate brain areas that might encode variables and vectors from a cognitive model (e.g., Love, 2015; Turner et al., 2017; White and Poldrack, 2013). In this approach, the scope of cognitive models is expanded to the explanation of neural data, and the addition of such data might allow to discriminate between two models that are indistinguishable from behavior alone. One might go so far as to hypothesize that a specific variable must correlate with activity in a specific brain area, which brings this approach closer to a CCN model.

I am very sympathetic towards this framework and I do believe that it can help to solve the kind of problem that I have identified here. However, this approach only gets us halfway towards a definitive solution. Without a model specifying how brain circuits compute such variables and vectors, we are still right in the conceptual nervous system territory.

The use of neuroimaging to test cognitive models requires making assumptions about how variables in a cognitive model are linked to neural activity, and that neural activity to the indirect measures provided by fMRI or EEG. Current practice among researchers is to simply assume that the relation between model parameters and neuroimaging measurements is straightforward (e.g., linear). It is not clear to what extent such simplifications have an impact on the inferences drawn from model-based neuroimaging. We are slowly starting to learn that neuroimaging has its own problems of theory underdetermination (Diedrichsen, 2018; Diedrichsen and Kriegeskorte, 2017; Liu et al., 2018; Soto et al., 2018; Popov et al., 2018), some of which involve distinguishing between computational and measurement models as sources of a particular pattern of results.

5.6. What about top-down constraints?

Developing CCN models can be considered a bottom-up approach to cognitive theory. It uses neurobiological constraints to discover principles of cognitive computation in real brains. An alternative is a top-down approach, which starts by creating models that characterize the computations that an organism must carry out in order to solve a specific environmental task (Anderson, 1990), and then uses those theories to constrain cognitive models (Griffiths et al., 2012). I think that the bottom-up and top-down approaches are complementary, rather than competitive. At the same time, there are several reasons why I think that a bottom-up CCN approach is preferable in learning theory.

First, top-down approaches have been severely criticized in recent years (Bowers and Davis, 2012; Jones and Love, 2011), with their flexibility pointed out as one of the most important issues. That is, models proposed at the "rational" or "computational" level of analysis do not seem to be more constrained than other cognitive models.

A second problem is practical: top-down approaches have been around for a while, but they have failed to provide constraints on algorithmic cognitive models. The reason seems rather simple: if two cognitive models are known to be empirically equivalent, they are both implementations of the same computational theory. Discovering that computational theory will not allow one to decide between the two implementations.

A final problem with top-down constraints from computational theory is that modelers tend to interpret them literally, leading to much theoretical confusion. For example, some theorists seem to propose not only that the brain carries out certain computations suggested by theory, but that it does so using the same algorithms used by human engineers to approximate such computations. For example, Bayesians

have proposed that the brain not only approaches optimal behavior, but that it does so using something like probabilistic Bayesian computation (Bowers and Davis, 2012), or using algorithms that are guaranteed to approach such Bayesian computation (e.g., Griffiths et al., 2012). Similarly, Gallistel and King (2009) not only argue that the brain must implement something like an addressable read/write memory and symbolic representation, but they think that their implementation must be relatively transparent. That is, they fail to consider the possibility that such operations could be implemented in the brain in a way that is different from human-engineered solutions (Dayan, 2009).

Is there any principled reason to consider such models and algorithms to be good candidates for cognitive mechanisms? It seems as though the only reason for doing so is because they work. This "engineering approach" to theory is pervasive in contemporary cognitive psychology and it seems to rely on the assumption that, if there is an algorithm that is designed to solve a particular computational task and if humans appear to face the same computational task, then the algorithm must be a good candidate for a psychological theory. This approach to theory is unlikely to solve issues of theory underdetermination, and it might instead exarcerbate them as it simply adds to an already rich "toolbox" of mechanisms that can be used by cognitive modelers to create their theories, without adding many constraints on how those theories are developed.

6. Conclusion

To summarize the main points of this article: the CNS approach (Hall, 2002, 2016) has dominated learning theory for the last century, with its most recent instantiation being the push for adopting deep neural networks as models of associative learning (Mondragón et al., 2017). I have argued that problems of theory underdetermination are inherent to computational cognitive modeling in general, and to the CNS approach in particular, and provided some examples of such problems. Underdetermination problems can be solved by putting additional constraints on models, and the CCN approach proposes using bottom-up constraints from computational and experimental neuroscience during model building. In addition, CCN modeling provides a framework for integration across disciplines interested in learning and behavior. Overall, learning theorists have much to win and little to lose from adopting a CCN approach to model development. The current level of knowledge and availability of new techniques in neuroscience mean that it is now possible to slowly move from modeling a "conceptual" nervous system to modeling the "real" nervous system. In the words of Hebb (1955):

> … the conceptual nervous system of 1930 was evidently like the gin that was being drunk about the same time; it was homemade and none too good, as Skinner pointed out, but it was also habit forming; and the effort to escape has not really been successful. Prohibition is long past. If we must drink we can now get better liquor; likewise, the conceptual nervous system of 1930 is out of date and – if we must neurologize – let us use the best brand of neurology we can find.

Acknowledgements

S.D.G. Thanks to Ed Wasserman for his comments on an early version of this paper.

References

Aitken, M.R.F., Dickinson, A., 2005 May. Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. Learn Behav. 33 (2), 147–159.
Allan, L.G., 1980. A note on measurement of contingency between two binary variables in judgment tasks. Bull. Psychon. Soc. 15 (3), 147–149.
Allan, L.G., 1993. Human contingency judgments: Rule-based or associative? Psychol. Bull. 114 (3), 435–448.
Allan, L.G., 2003. Assessing power PC. Learn Behav. 31 (2), 192–204.
Allan, L.G., Hannah, S.D., Crump, M.J.C., Siegel, S., 2008. The psychophysics of contingency assessment. J. Exp. Psychol. Gen. 137 (2), 226–243 URL http://www.sciencedirect.com/science/article/B6X07-4SYM07R-2/1/c1325c7b54d5bdd54be0a06f7551de12.
Alsop, B., 1998. Receiver operating characteristics from nonhuman animals: Some implications and directions for research with humans. Psychonomic Bull. Rev. 5 (2), 239–252.
Anderson, J.R., 1978. Arguments concerning representations for mental imagery. Psychol. Rev. 85 (4), 249.
Anderson, J.R., 1990. The adaptive character of thought. Lawrence Erlbaum Associates, Hillsdale, NJ.
Ashby, F.G., 2011 Feb. Statistical Analysis of fMRI Data. MIT Press.
Ashby, F.G., 2018. Computational cognitive neuroscience. In: Batchelder, W.H., Colonius, H., Dzhafarov, E.N. (Eds.), New Handbook of Mathematical Psychology. Vol. 2. Cambridge University Press, New York, NY.
Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., Waldron, E.M., 1998. A neuropsychological theory of multiple systems in category learning. Psychol. Rev. 105 (3), 442–481.
Ashby, F.G., Helie, S., 2011. A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. J. Math. Psychol. 55 (4), 273–289.
Ashby, F.G., Soto, F.A., 2015. Multidimensional signal detection theory. In: Busemeyer, J., Townsend, J.T., Wang, Z.J., Eidels, A. (Eds.), Oxford Handbook of Computational and Mathematical Psychology. Oxford University Press, New York, NY, pp. 13–34.
Ashby, F.G., Townsend, J.T., 1986. Varieties of perceptual independence. Psychol. Rev. 93 (2), 154–179.
Ashby, F.G., Waldschmidt, J.G., 2008. Fitting computational models to fMRI data. Behav. Res. Methods 40 (3), 713–721.
Aydin, A., Pearce, J.M., 1995. Summation in autoshaping with short- and long-duration Stimuli. Q. J. Exp. Psychol. 48B (3), 215–234.
Aydin, A., Pearce, J.M., 1997. Some determinants of response summation. Anim. Learn. Behav. 25 (1), 108–121.
Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E., Pouget, A., 2008. Probabilistic population codes for Bayesian decision making. Neuron 60 (6), 1142–1152.
Blough, D.S., 1967. Stimulus generalization as signal detection in pigeons. Science 158 (3803), 940–941.
Blough, D.S., 2001. Some contributions of signal detection theory to the analysis of stimulus control in animals. Behav. Process. 54, 127–136.
Bowers, J.S., Davis, C.J., 2012. Bayesian just-so stories in psychology and neuroscience. Psychol. Bull. 138 (3), 389–414.
Brette, R., 2015 Apr. What is the most realistic single-compartment model of spike initiation? PLoS Comput. Biol. 11 (4), e1004114 URL https://doi.org/10.1371/journal.pcbi.1004114.
Brette, R., Gerstner, W., 2005 Nov. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. J. Neurophysiol. 94 (5), 3637–3642 URL http://jn.physiology.org/content/94/5/3637.
Carandini, M., Heeger, D.J., 2012 Jan. Normalization as a canonical neural computation. Nat. Rev. Neurosci. 13 (1), 51–62 URL https://doi.org/10.1038/nrn3136.
Carrere, M., Alexandre, F., 2015. A pavlovian model of the amygdala and its influence within the medial temporal lobe. Front. Syst. Neurosci. 9, 41 URL https://www.frontiersin.org/articles/10.3389/fnsys.2015.00041/full.
Chapman, G.B., 1991 Sep. Trial order affects cue interaction in contingency judgment. J. Exp. Psychol. Learn. Mem. Cognit. 17 (5), 837–854.
Chapman, G.B., Robbins, S.J., 1990. Cue interaction in human contingency judgment. Mem. Cognit. 18 (5), 537–545.
Cheng, P.W., 1997 Apr. From covariation to causation: A causal power theory. Psychol. Rev. 104 (2), 367–405.
Cheng, P.W., Novick, L.R., 1992 Apr. Covariation in natural causal induction. Psychol. Rev. 99 (2), 365–382.
Cooper, L.N., Bear, M.F., 2012 Nov. The BCM theory of synapse modification at 30: interaction of theory with experiment. Nat. Rev. Neurosci. 13 (11), 798–810 URL http://www.nature.com/nrn/journal/v13/n11/full/nrn3353.html?WT.ec_id=NRN-201211.
Corlett, P.R., Aitken, M.R., Dickinson, A., Shanks, D.R., Honey, G.D., Honey, R.A., Robbins, T.W., Bullmore, E.T., Fletcher, P.C., 2004. Prediction Error during Retrospective Revaluation of Causal Associations in Humans:: fMRI Evidence in Favor of an Associative Model of Learning. Neuron 44 (5), 877–888.
Danks, D., 2003. Equilibria of the Rescorla-Wagner model. J. Math. Psychol. 47 (2), 109–121.
Danks, D., Griffiths, T.L., Tenenbaum, J.B., 2003. Dynamical causal learning. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing. MIT Press, Cambridge, MA, pp. 67–74.
Dayan, P., 2009 Oct. A neurocomputational jeremiad. Nat. Neurosci. 12 (10), 1207. https://doi.org/10.1038/nn1009-1207.
Dean, P., Porrill, J., Ekerot, C.F., J“orntell, H., 2010. The cerebellar microcircuit as an adaptive filter: experimental and computational evidence. Nat. Rev. Neurosci. 11 (1), 30.
Deco, G., Jirsa, V.K., Robinson, P.A., Breakspear, M., Friston, K., 2008. The dynamic brain: from spiking neurons to neural masses and cortical fields. PLoS Comput. Biol. 4 (8), e1000092. https://doi.org/10.1371/journal.pcbi.1000092.
Deneve, S., Latham, P.E., Pouget, A., 1999 Aug. Reading population codes: a neural implementation of ideal observers. Nat. Neurosci. 2 (8), 740–745.
Diedrichsen, J., 2018. Representational models and the feature fallacy. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences.
Diedrichsen, J., Kriegeskorte, N., 2017 Apr. Representational models: A common framework for understanding encoding, pattern-component, and representational-

similarity analysis. PLOS Comput. Biol. 13 (4), e1005508.
Earman, J., 1993. Underdetermination, realism, and reason. Midwest studies in philosophy 18 (1), 19–38.
Estes, W.K., 1994. Classification and Cognition. Oxford University Press.
Fletcher, P.C., Anderson, J.M., Shanks, D.R., Honey, R., Carpenter, T.A., Donovan, T., Papadakis, N., Bullmore, E.T., 2001. Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nat. Neurosci. 4 (10), 1043–1048.
Freeman, J.H., 2015. Cerebellar learning mechanisms. Brain Res. 1621, 260–269.
Gallistel, C.R., King, A.P., 2009 May. Memory and the computational brain: Why cognitive science will transform neuroscience. John Wiley & Sons.
Ganis, G., Schendan, H.E., 2011. Visual imagery. Wiley Interdisciplinary Reviews: Cognitive Science 2 (3), 239–252 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.103.
Gerstner, W., Naud, R., 2009 Oct. How good are neuron models? Science 326 (5951), 379–380 URL http://infoscience.epfl.ch/record/142067.
Ghirlanda, S., 2015. On elemental and configural models of associative learning. J. Math. Psychol. 64-65, 8–16.
Gluck, M.A., Allen, M.T., Myers, C.E., Thompson, R.F., 2001 Nov. Cerebellar substrates for error correction in motor conditioning. Neurobiol. Learn. Mem. 76 (3), 314–341.
Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.
Griffiths, T.L., Vul, E., Sanborn, A.N., 2012 Aug. Bridging levels of analysis for probabilistic models of cognition. Curr. Directions Psychol. Sci. 21 (4), 263–268 URL http://cdp.sagepub.com/content/21/4/263.
Grünbaum, T., 2018. The two visual systems hypothesis and contrastive underdetermination. Synthese 1–24.
Hall, G., 2002. Associative structures in pavlovian and instrumental conditioning. In: Gallistel, R. (Ed.), Steven's Handbook of Experimental Psychology. Wiley, New York, pp. 1–45.
Hall, G., 2016. Mackintosh and associationism. In: Trobalon, J.B., Chamizo, V.D. (Eds.), Associative Learning and Cognition. Homage to Professor N. J. Mackintosh. In Memoriam (1935–2015). Edicions Universitat Barcelona, Barcelona, pp. 21–35.
Harris, J.A., 2006. Elemental representations of stimuli in associative learning. Psychol. Rev. 113 (3), 584–605.
Harris, J.A., Livesey, E.J., 2010. An attention-modulated associative network. Learn Behav. 38 (1), 1–26.
Hebb, D.O., 1955. Drives and the C. N. S. (conceptual nervous system). Psychol. Rev. 62 (4), 243–254.
Henson, R., Friston, K.J., 2007. Convolution models for fMRI. In: Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E. (Eds.), Statistical parametric mapping: The analysis of functional brain images. Elsevier, London, pp. 178–192.
Herry, C., Johansen, J.P., 2014. Encoding of fear learning and memory in distributed neuronal circuits. Nat. Neurosci. 17 (12), 1644.
Izhikevich, E., 2004 Sep. Which model to use for cortical spiking neurons? Neural Networks, IEEE Transactions on 15 (5), 1063–1070.
Jones, M., Dzhafarov, E.N., 2014. Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. Psychol. Rev. 121 (1), 1–32.
Jones, M., Love, B.C., 2011. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci. 34 (04), 169–188 URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8359929.
Kehoe, E.J., Gormezano, I., 1980. Configuration and combination laws in conditioning with compound stimuli. Psychol. Bull. 87 (2), 351.
Kehoe, E.J., Horne, A.J., Horne, P.S., Macrae, M., 1994. Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. Anim. Learn. Behav. 22 (1), 19–26.
Kinder, A., Lachnit, H., 2003. Similarity and discrimination in human Pavlovian conditioning. Psychophysiology 40 (2), 226–234.
Krasne, F.B., Fanselow, M., Zelikowsky, M., 2011. Design of a neurally plausible model of fear learning. Front. Behav. Neurosci. 5, 41 URL https://www.frontiersin.org/articles/10.3389/fnbeh.2011.00041/full.
Kuśmierz, L., Isomura, T., Toyoizumi, T., 2017 Oct. Learning with three factors: modulating Hebbian plasticity with errors. Curr. Opin. Neurobiol. 46 (Supplement C), 170–177 URL http://www.sciencedirect.com/science/article/pii/S0959438817300612.
Lachnit, H., 1988. Convergent Validation of Information Processing Constructs With Pavlovian Methodology. J. Exp. Psychol. Hum. Percept. Perform. 14 (1), 143–152.
Laudan, L., 1990. Demystifying underdetermination. In: Savage, C.W. (Ed.), Scientific theories, Minnesota studies in the philosophy of science. Vol. 14. University of Minnesota Press, Minneapolis, pp. 267–297.
Laudan, L., Leplin, J., 1991. Empirical equivalence and underdetermination. J. Philos. 88 (9), 449–472.
Lepora, N.F., Porrill, J., Yeo, C., Dean, P., 2010. Sensory prediction or motor control? Application of Marr-Albus type models of cerebellar function to classical conditioning. Front. Comput. Neurosci. 4, 140.
Ling, S., Liu, T., Carrasco, M., 2009. How spatial and feature-based attention affect the gain and tuning of population responses. Vision Res. 49 (10), 1194–1204.
Liu, T., Cable, D., Gardner, J.L., 2018 Jan. Inverted encoding models of human population response conflate noise and neural tuning width. J. Neurosci. 38 (2), 398–408.
Love, B.C., 2015. The algorithmic level is the bridge between computation and brain. Top. Cognit. Sci. 7 (2), 230–242.
Lyre, H., 2011. Is structural underdetermination possible? Synthese 180 (2), 235–247.
Macmillan, N.A., Creelman, C.D., 2005. Detection theory: A user's guide, 2nd Edition. Lawrence Erlbaum Associates, Mahwah, NJ.
Maia, T.V., 2009. Reinforcement learning, conditioning, and the brain: Successes and challenges. Cognit. Affect. Behav. Neurosci. 9 (4), 343–364.
Marcelja, S., 1980. Mathematical description of the responses of simple cortical cells. J. Opt. Soc. Am. 70 (11), 1297–1300.
Massaro, D.W., 1988. Some criticisms of connectionist models of human performance. J. Mem. Lang. 27 (2), 213–234.
McLaren, I.P.L., Mackintosh, N.J., 2002. Associative learning and elemental representation: II. Generalization and discrimination. Anim. Learn. Behav. 30 (3), 177–200.
McNally, G.P., Johansen, J.P., Blair, H.T., 2011 May. Placing prediction into the fear circuit. Trends Neurosci.
Mestry, N., Wenger, M.J., Donnelly, N., 2012. Identifying sources of configurality in three face processing tasks. Front. Percept. Sci. 3, 456.
Mondragón, E., Alonso, E., Kokkola, N., 2017. Associative learning should go deep. Trends Cogn. Sci. 21 (11), 822–825.
Morrison, A., Diesmann, M., Gerstner, W., 2008. Phenomenological models of synaptic plasticity based on spike timing. Biol. Cybern. 98 (6), 459–478.
Moustafa, A.A., Gilbertson, M.W., Orr, S.P., Herzallah, M.M., Servatius, R.J., Myers, C.E., 2013 Feb. A model of amygdala-hippocampal-prefrontal interaction in fear conditioning and extinction in animals. Brain Cogn. 81 (1), 29–43 URL http://www.sciencedirect.com/science/article/pii/S0278262612001418.
Naud, R., Marcille, N., Clopath, C., Gerstner, W., 2008. Firing patterns in the adaptive exponential integrate-and-fire model. Biol. Cybern. 99 (4), 335–347 URL http://www.springerlink.com/content/uw514r4j6323x763/abstract/.
Navarro, D.J., Pitt, M.A., Myung, I.J., 2004. Assessing the distinguishability of models and the informativeness of data. Cognit. Psychol. 49 (1), 47–84.
Niv, Y., 2009 Jun. Reinforcement learning in the brain. J. Math. Psychol. 53 (3), 139–154 URL http://www.sciencedirect.com/science/article/B6WK3-4VK6995-1/2/492309e97f849421a460c8b5a2c784b5.
Norton, J., 2008. Must evidence underdetermine theory? In: Carrier, M., Howard, D., Kourany, J. (Eds.), The challenge of the social and the pressure of practice: Science and values revisited. University of Pittsburgh Press, Pittsburgh, PA, pp. 17–44.
Page, M., 2000. Connectionist modeling in psychology: A localist manifesto. Behav. Brain Sci. 23, 443–512.
Paradiso, M.A., 1988. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. Biol. Cybern. 58 (1), 35–49.
Pavlov, I.P., 1927. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxford University Press, London.
Pearce, J.M., 1987. A model for stimulus generalization in Pavlovian conditioning. Psychol. Rev. 94 (1), 61–73.
Pearce, J.M., 1994. Similarity and discrimination: A selective review and a connectionist model. Psychol. Rev. 101 (4), 587–607.
Pearce, J.M., 2002. Evaluation and development of a connectionist theory of configural learning. Anim. Learn. Behav. 30 (2), 73–95.
Perez, O.D., San Martin, R., Soto, F.A., 2018. Exploring the effect of stimulus similarity on the summation effect in human causal learning. Exp. Psychol. 65 (4), 183–200.
Popov, V., Ostarek, M., Tenison, C., 2018. Practices and pitfalls in inferring neural representations. Neuroimage 174, 340–351.
Pouget, A., Dayan, P., Zemel, R.S., 2003. Inference and computation with population codes. Annu. Rev. Neurosci. 26 (1), 381–410 URL http://www.annualreviews.org/doi/pdf/10.1146/annurev.neuro.26.041002.131112.
Rescorla, R.A., 1997 May. Summation: Assessment of a configural theory. Anim. Learn. Behav. 25 (2), 200–209.
Rescorla, R.A., Coldwell, S.E., 1995. Summation in autoshaping. Anim. Learn. Behav. 23 (3), 314–326.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical conditioning II: Current theory and research. Appleton-Century-Crofts, New York, pp. 64–99.
Richler, J.J., Gauthier, I., Wenger, M.J., Palmeri, T.J., 2008. Holistic processing of faces: Perceptual and decisional components. J. Exp. Psychol. Learn. Mem. Cognit. 34 (2), 328–342.
Rumelhart, D.E., McClelland, J.L. (Eds.), 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.
Schmajuk, N.A., 1997. Animal Learning and Cognition: A Neural Network Approach. University Press, Cambridge, MA.
Seger, C.A., 2008. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. Neurosci. Biobehav. Rev. 32 (2), 265–278.
Series, P., Stocker, A.A., Simoncelli, E.P., 2009. Is the homunculus “aware” of sensory adaptation? Neural Comput. 21 (12), 3271–3304.
Seung, H.S., Sompolinsky, H., 1993 Nov. Simple models for reading neuronal population codes. Proc. Natl. Acad. Sci. 90 (22), 10749–10753 URL http://www.pnas.org/content/90/22/10749.
Shanks, D.R., 1985. Forward and backward blocking in human contingency judgement. Q. J. Exp. Psychol. 37B (1), 1–21.
Shohamy, D., Myers, C., Kalanithi, J., Gluck, M., 2008. Basal ganglia and dopamine contributions to probabilistic category learning. Neurosci. Biobehav. Rev. 32 (2), 219–236.
Siegel, S., Allan, L.G., Hannah, S.D., Crump, M.J.C., 2009. Applying signal detection theory to contingency assessment. Compat. Cognit. Behav. Rev. 4, 116–134.
Silbert, N.H., Thomas, R., 2013. Decisional separability, model identification, and statistical inference in the general recognition theory framework. Psychon. Bull. Rev. 20 (1), 1–20.
Silbert, N.H., Thomas, R.D., 2017. Identifiability and testability in GRT with individual differences. J. Math. Psychol. 77, 187–196.
Skinner, B.F., 1938. The behavior of organisms: an experimental analysis. Appleton-Century-Crofts, Oxford, England.

Soto, F.A., 2018. Contemporary associative learning theory predicts failures to obtain blocking: Comment on Maes et al. (2016). J. Exp. Psychol. Gen. 147 (4), 597–602.
Soto, F.A., Gershman, S.J., Niv, Y., 2014. Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. Psychol. Rev. 121 (3), 526–558.
Soto, F.A., Quintana, G.R., Pérez-Acosta, A.M., Ponce, F.P., Vogel, E.H., 2015a. Why are some dimensions integral? Testing two hypotheses through causal learning experiments. Cognition 143, 163–177.
Soto, F.A., Siow, J.Y.M., Wasserman, E.A., 2012. View-invariance learning in object recognition by pigeons depends on error-driven associative learning processes. Vision Res. 62, 148–161.
Soto, F.A., Vogel, E.H., Castillo, R.D., Wagner, A.R., 2009. Generality of the summation effect in human causal learning. Q. J. Exp. Psychol. 62 (5), 877–889.
Soto, F.A., Vucovich, L., Musgrave, R., Ashby, F.G., 2015b. General recognition theory with individual differences: A new method for examining perceptual and decisional interactions with an application to face perception. Psychon. Bull. Rev. 22 (1), 88–111.
Soto, F.A., Vucovich, L.E., Ashby, F.G., 2018. Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. PLOS Comput. Biol. 14 (10), e1006470.
Soto, F.A., Wasserman, E.A., 2010a. Comparative vision science: Seeing eye to eye? Compat. Cognit. Behav. Rev. 5, 148–154.
Soto, F.A., Wasserman, E.A., 2010b. Error-driven learning in visual categorization and object recognition: A common elements model. Psychol. Rev. 117 (2), 349–381.
Soto, F.A., Wasserman, E.A., 2011. Asymmetrical interactions in the perception of face identity and emotional expression are not unique to the primate visual system. J. Vision. 11 (324), 1–18.
Soto, F.A., Wasserman, E.A., 2012a. A category-overshadowing effect in pigeons: Support for the Common Elements Model of object categorization learning. J. Exp. Psychol. Anim. Behav. Process. 38 (3), 322–328.
Soto, F.A., Wasserman, E.A., 2012b. Visual object categorization in birds and primates: Integrating behavioral, neurobiological, and computational evidence within a “general process” framework. Cognit. Affect. Behav. Neurosci. 12 (1), 220–240.
Soto, F.A., Wasserman, E.A., 2014. Mechanisms of object recognition: what we have learned from pigeons. Front. Neural Circuits 8, 122.
Soto, F.A., Zheng, E., Fonseca, J., Ashby, F.G., 2017. Testing separability and independence of perceptual dimensions with general recognition theory: a tutorial and new R package (grtools). Front. Psychol. 8, 696.
Stanford, K., 2017. Underdetermination of Scientific Theory. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy, winter 2017 Edition. Metaphysics Research Lab, Stanford University URL https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/.
Stanford, P.K., 2001. Refusing the devil's bargain: What kind of underdetermination should we take seriously? Philos. Sci. 68 (S3), S1–S12.
Suri, R.E., 2002. TD models of reward predictive responses in dopamine neurons. Neural Netw. 15 (4-6), 523–533 URL http://www.sciencedirect.com/science/article/B6T08-461XHD9-7/2/8db6b7fa3e11e78a3df8950d45598f84.
Tenenbaum, J.B., Griffiths, T.L., 2001. Structure learning in human causal induction. Adv. Neural Inform. Process. Syst. 13.
Terman, M., Terman, J.S., 1972. Concurrent variation of response bias and sensitivity in an operant-psychophysical test. Percept. Psychophys. 11 (6), 428–432.
Thorwart, A., Livesey, E., Harris, J., 2012. Normalization between stimulus elements in a model of Pavlovian conditioning: Showjumping on an elemental horse. Learn Behav. 40 (3), 334–346.
Townsend, J.T., 1990. Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. Psychol. Sci. 1 (1), 46–54.
Townsend, J.T., 2008 Oct. Mathematical psychology: Prospects for the 21st century: A guest editorial. J. Math. Psychol. 52 (5), 269–280 URL http://www.sciencedirect.com/science/article/B6WK3-4T13JGH-1/2/5e61db91a36d147-f239e23b72cd6d6b2.
Tsunoda, K., Yamane, Y., Nishizaki, M., Tanifuji, M., 2001. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat. Neurosci. 4 (8), 832–838. https://doi.org/10.1038/90547.
Turnbull, M.G., 2018. Underdetermination in science: What it is and why we should care. Philosophy Compass 13 (2), e12475.
Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J., Van Maanen, L., 2017. Approaches to analysis in model-based cognitive neuroscience. J. Math. Psychol. 76, 65–79.
Turner, D.C., Aitken, M.R.F., Shanks, D.R., Sahakian, B.J., Robbins, T.W., Schwarzbauer, C., Fletcher, P.C., 2004. The role of the lateral frontal cortex in causal associative learning: exploring preventative and super-learning. Cereb. Cortex 14 (8), 872–880.
van Gerven, M.A.J., 2017. A primer on encoding models in sensory neuroscience. J. Math. Psychol. 76, 172–183.
Van Hamme, L.J., Wasserman, E.A., 1994. Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. Learn. Motiv. 25 (2), 127–151.
Van Rossum, M.C.W., Roth, A., 2010. Modeling synapses. In: De Schutter, E. (Ed.), Computational modeling methods for neuroscientists. MIT Press, Cambridge, MA, pp. 139–160.
Van Zandt, T., Ratcliff, R., 1995. Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. Psychon. Bull. Rev. 2 (1), 20–54.
Villagrasa, F., Baladron, J., Vitay, J., Schroll, H., Antzoulatos, E.G., Miller, E.K., Hamker, F.H., 2018 Oct. On the role of cortex-basal ganglia interactions for category learning: A neurocomputational approach. J. Neurosci. 38 (44), 9551–9562 URL http://www.jneurosci.org/content/38/44/9551.
Vogel, E.H., Castro, M.E., Saavedra, M.A., 2004 Apr. Quantitative models of Pavlovian conditioning. Brain Res. Bull. 63 (3), 173–202.
Wagner, A.R., 2003. Context-sensitive elemental theory. Q. J. Exp. Psychol. 56B (1), 7–29.
Wagner, A.R., 2007. Evolution of an elemental theory of Pavlovian conditioning. Learn Behav. 36 (3), 253–265.
Wagner, A.R., Brandon, S.E., 2001. A componential theory of Pavlovian conditioning. In: Mowrer, R.R., Klein, S.B. (Eds.), Handbook of Contemporary Learning Theories. Erlbaum, Mahwah, NJ, pp. 23–64.
Wagner, A.R., Rescorla, R.A., 1972. Inhibition in Pavlovian conditioning: Application of a theory. In: Boakes, R.A., Haliday, M.S. (Eds.), Inhibition and Learning. Academic Press, New York, pp. 301–336.
Wenger, M.J., Ingvalson, E.M., 2002. A decisional component of holistic encoding. J. Exp. Psychol. Learn. Mem. Cognit. 28 (5), 872.
White, C.N., Poldrack, R.A., 2013. Using fMRI to constrain theories of cognition. Perspect. Pyschol. Sci. 8 (1), 79–83.
Whitlow, J.W., Wagner, A.R., 1972. Negative patterning in classical conditioning: Summation of response tendencies to isolable and configural components. Psychon. Sci. 27, 299–301.
Young, M.E., 1995 Mar. On the origins of personal causal theories. Psychonomic Bulletin & Review 2 (1), 83–104.
Yuille, A., 2005. The Rescorla-Wagner algorithm and maximum likelihood estimation of causal parameters. Adv. Neural Inform. Process. Syst. 17.
Zhu, W., Shelley, M., Shapley, R., 2009 Apr. A neuronal network model of primary visual cortex explains spatial frequency selectivity. J. Comput. Neurosci. 26 (2), 271–287. https://doi.org/10.1007/s10827-008-0110-x.