



## Representing absence of evidence: why algorithms *and* representations matter in models of language and cognition

Franziska Bröker & Michael Ramscar

To cite this article: Franziska Bröker & Michael Ramscar (2020): Representing absence of evidence: why algorithms *and* representations matter in models of language and cognition, Language, Cognition and Neuroscience, DOI: [10.1080/23273798.2020.1862257](https://doi.org/10.1080/23273798.2020.1862257)

To link to this article: <https://doi.org/10.1080/23273798.2020.1862257>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 24 Dec 2020.



[Submit your article to this journal](#)



Article views: 790



[View related articles](#)





[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

## Representing absence of evidence: why algorithms *and* representations matter in models of language and cognition

Franziska Bröker <sup>a,b</sup> and Michael Ramscar <sup>c</sup>

<sup>a</sup>Department for Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany; <sup>b</sup>Gatsby Computational Neuroscience Unit, University College London, London, UK; <sup>c</sup>Quantitative Linguistics, University of Tübingen, Tübingen, Germany

### ABSTRACT

Theories of language and cognition develop iteratively from ideas, experiments and models. The abstract nature of “cognitive processes” means that computational models play a critical role in this, yet bridging the gaps between models, data, and interpretations is challenging. While the *how* and *why* computations are performed is often the primary research focus, the conclusions drawn from models can be compromised by the representations chosen for them. To illustrate this point, we revisit a set of empirical studies of language acquisition that appear to support different models of learning from implicit negative evidence. We examine the degree to which these conclusions were influenced by the representations chosen and show how a plausible single mechanism account of the data can be formulated for representations that faithfully capture the task design. The need for input representations to be incorporated into model conceptualisations, evaluations, and comparisons is discussed.

### ARTICLE HISTORY

Received 1 June 2020  
Accepted 24 November 2020

### KEYWORDS

Computational modelling; representations; error-driven learning; language acquisition; negative evidence

### Introduction

Our understanding of the mind and brain advances through an iterative process of developing theories and then empirically testing their predictions. Models play an integral part in this process, since (in principle at least) they allow theoretical constructs to be formalised and quantified. Accordingly, models ideally serve to increase the clarity and specificity of the predictions made by theories. The devil, however, is in the details. As Goodman (1976) noted,<sup>1</sup> models vary widely in their specificity and purpose.

A great deal of thought has gone into the development of models simulating *how* the brain computes whatever it computes (algorithmically) as it solves whatever problems it solves (at what is sometimes called the computational level). However, input representations – specifically, the data structures that co-determine the output or performance of computational mechanisms/algorithms that will be the focus of this work – have generally received less attention. This is reflected in the nature of the debates between cognitive modellers themselves, which have tended to argue at length about the merits of viewing cognitive processes as say, generative grammars, connectionist architectures, or forms of Bayesian inference, whilst implicitly agreeing on the nature of inputs to these processes (see e.g.

Griffiths et al., 2010; McClelland et al., 2010). On one hand, this is curious because all model predictions depend to some extent on the representations that are input into an algorithm (in the extreme case, any model can be broken by choosing a “bad” input representation, or “fixed” by hand tailoring a “good” input representation), however detailed, empirically grounded theories of how the brain encodes complex inputs are rare in the literature. On the other hand, this is hardly surprising when one considers that, for example, visual inputs are first massively compressed as they pass from retinal cells through to the lateral geniculate nucleus and then massively expanded when they are processed in V1 (DiCarlo et al., 2012; Stevens, 2001), such that it seems that the actual input representations to any neural or cognitive model (apart, perhaps, from those dealing with sensory receptors) can never be precisely inferred by simply observing the physical properties of the environment. Accordingly, it follows that the actual input representations implemented in all such models are currently somewhat massively unconstrained and that from both a theoretical and empirical perspective, the actual nature of neural and cognitive representations must always be something of a “black box”. These considerations raise in turn the question of the degree to which the behaviour of any model

**CONTACT** Franziska Bröker  franziska.broeker.15@ucl.ac.uk

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group  
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

developed to emulate a putative cognitive or neural mechanism is in fact limited, or even determined, by the intuitive and often unanalysed input representations specified for it. Is it even possible to make claims about empirical validity of the algorithmic or conceptual properties of a model, if any predictions derived from them can be changed by the choice of input representation?

In this paper, we suggest that while this problem has sometimes been acknowledged, the theoretical impact of evaluating and comparing algorithms while ignoring input representations (i.e. the data structures/knowledge representations that serve as inputs to algorithms/inference mechanisms) may be greater than is generally appreciated. To illustrate how interpretations of mechanisms interact with the choice of representation in a model, we will revisit a set of studies investigating language acquisition in the face of implicit negative evidence (Hsu & Griffiths, 2009, 2016). While the problem of selecting input representation is particularly relevant to the domain of language acquisition, we suggest that the issues highlighted by these analyses apply more broadly across the brain and cognitive sciences.

### ***Negative evidence in language learning: a case study***

As they learn their native languages, children somehow master complex alternation patterns (e.g. of regular and irregular inflection, or different verb argument structures) in spite of the fact that they never receive any explicit instruction on their use, or even any explicit feedback on the errors that they inevitably make in the course of the learning process. Accordingly, explaining how children manage to master the complexities of language, and accounting for the specific developmental patterns that tend to accompany this, has emerged as a central question in cognitive science, and also as an area of considerable debate. For example, it has been argued that given the evidence they have available to them, children could never learn to make the transition to adult language based on experience alone, and that learning even simple aspects of grammar is logically impossible in the absence of innate constraints on what is learned (see e.g. Johnson, 2004; Pinker, 1984, 1989). By contrast, although the information offered by violations of expectation was often marginalised or ignored in early discussions of language learning (Brown & Hanlon, 1970; Pinker, 1984, 1989), the development of computational models of learning (and in particular, error-driven models of learning) has led to a rebirth of interest in the role that this kind of “indirect” (or implicit) negative evidence plays in the process (see Ramscar et al., 2013, for a review). To this end,

Hsu and Griffiths (2009, 2016) proposed that the assumptions that children make about the way that their linguistic experiences are sampled offer a source of implicit evidence that can in turn help restrict the kind of generalisation that they make. Hsu and Griffiths suggest that learners might implicitly adopt one of two approaches towards the samples that they are exposed to. In the first of these approaches, they propose that a learner might classify observed utterances as being either grammatically correct or incorrect, which involves learning a mapping from sentences to grammaticality. (We note that the questions of whether grammaticality is actually binary in this way, see e.g. Gibson et al., 2013; Mahowald et al., 2016, or whether explicit feedback is actually relevant to children’s language learning, see e.g. Ramscar and Yarlett (2007), are both subject to much debate; while acknowledging these debates, for current purposes we will take this starting assumption from Hsu and Griffiths as given.) Since this does not involve any assumptions about the distribution from which utterances are sampled (which Hsu and Griffiths describe as a discriminative or weak sampling approach to learning), the absence of sentences from a language does not provide information about their grammaticality. In the alternative approach, they propose that a learner might acquire the probability distribution over valid utterances assuming linguistic input is sampled from the true language distribution (which Hsu and Griffiths describe as a generative or strong sampling approach to learning; these different learning/sampling strategies will be discussed in greater detail below). Hsu and Griffiths suggest that it is only this latter case that allows implicit evidence to be harnessed in support of learning.

To explore this proposal, the authors present evidence from a series of computational simulations, and a set of studies that test model predictions against empirical data from an artificial grammar learning task. Crucially, the experiments manipulated training procedures to examine the impact that the order in which information was presented had on participants’ ability to make use of implicit negative evidence. The authors hypothesised that these different training procedures would prompt subjects to adopt different learning approaches that would either be sensitive to the distribution of grammatical sentences or not. The results appeared to confirm their predictions, showing that the two experimental groups did indeed seem to respond differently to the absence of observations in both learning conditions, with their behaviour matching the predictions of the two models. The authors conclude that the learnability of correct linguistic generalisations/exceptions may thus depend on the sampling

assumptions made by learners. They argue that, if a strong sampling assumption is made, implicit negative evidence can be exploited, which could prove critical to language learnability. Notably, these findings also complement results from studies of category learning where similar distinctions between discriminative and generative models have been explored in relation to different training procedures which are often referred to as classification versus inference/observation learning (e.g. Hsu & Griffiths, 2010; Levering & Kurtz, 2015; Love et al., 2015).

In what follows we revisit the studies reported by Hsu and Griffiths in order to provide a case study of the way that the choice of input representations can be of critical relevance when it comes to the interpretation of model performance, and in turn to establishing the degree to which empirical results can be taken as support for the formal predictions of models. In particular, we shall examine whether the results observed necessarily stem from the sampling assumptions associated with different learning strategies/mechanisms or whether the representations chosen determined the predictions that the models made.

In the first part of this case study, we shall analyse the models proposed by Hsu and Griffiths (2009, 2016). We will highlight the way that the representations employed in developing these models embody assumptions about the representations available to learners. When made explicit, these assumptions bring into question the degree to which these models can be taken to support the idea that distinct learning mechanisms are required in this instance. We then show how by starting with a simple algorithm and an alternative input representation that can be reasonably derived from the task structure one can derive a straight-forward model based on general learning principles that offers a very different account of the experimental data. We demonstrate that the behavioural differences between the experimental conditions can be modelled as emerging from a single learning mechanism given task-informed representations as opposed to a dual strategy perspective.

For present purposes, it is important to note that the issue here is not one of establishing that this single mechanism model is right or that the models put forward by Hsu and Griffiths (2009, 2016) are wrong. Rather, the comparisons and contrasts of all of the models reported in the case study below are intended to emphasise that *how* a model computes and *what* it computes over are of equal importance, such that an algorithm cannot be evaluated independently of the representations it is provided with. Accordingly, any inferences about *how* and *what* an empirical system, like the brain, computes must necessarily be constrained

by similar factors. In the light of this, we then discuss how a more rigorous consideration of both mechanisms and input representations to models can help improve their contribution to our understanding of the mind/brain and discuss how these issues can be addressed in future work.

### Case study

We begin our case study by describing in more detail the experimental design employed in Hsu & Griffiths' work, before introducing the models that were intended to predict subjects' performance in these experiments. We then summarise the empirical and simulation results. Finally, we analyse the representational choices in these models and discuss how such choices can serve to influence the behaviour of, and resulting predictions from, models (and indeed the structure of a model itself), and examine the degree to which these representational choices can in turn serve to influence subsequent interpretations of behavioural data.

### Experimental design

The work described in our case study sought to test the hypothesis that the assumptions made by learners regarding the way in which linguistic observations are sampled can yield dissociable end states when it comes to language learning (Hsu & Griffiths, 2009, 2016). To empirically test this idea, three experiments were conducted in which students learned to classify sentences from artificial languages as being grammatical or ungrammatical. Whether a sentence was correct or incorrect depended on properties of its structure that were initially unknown to subjects and thus had to be learned.

In Experiments 1 and 2, the training materials that were presented to subjects in order to facilitate this learning comprised a set of three-word sentences, in which some sentences would be grammatical and some ungrammatical. Each sentence consisted of two nouns and one verb and expressed a directed action between subject (first noun occurring) and object (second noun occurring) of the sentence. Nouns were drawn from a set of three pseudo-words while the set of verbs comprised four pseudo-words (Experiment 1) or five pseudo-words (Experiment 2). The grammaticality of each possible sequencing of nouns and verbs (noun-noun-verb, noun-verb-noun, verb-noun-noun) depended solely on the combination of individual verbs and their position in the sentence (i.e. one verb might only be grammatically correct if used in the beginning of a sentence, whereas another might be used in the beginning as well as the end of a sentence).

Experiment 3 was designed to provide an example of learning to contract nouns and subsequent modifiers. Grammaticality of a contraction depended on the specific modifier and its positioning after either the subject or object of the sentence. Grammaticality and number of training sentences presented to models and subjects are displayed in Table 1.

All of the behavioural experiments in our case study started with a pre-training session in which subjects acquired the meaning of the pseudo-words (e.g. *blergen* is lion, *semz* is explode) by means of visual and auditory presentation of the stimuli. In the subsequent training session, subjects were told that they would be required to learn the grammaticality of sentences presented to them. During training subjects were exposed to both grammatical and ungrammatical sentences. After the training session, participants were asked to produce grammaticality ratings and then complete a sentence production task. Crucially, one critical test sentence was always withheld during training.

To predict subjects' behaviour in the experiments reported in our case study, two computational models were presented (described in more detail below). These models made different predictions about the grammaticality of the withheld test sentence as a function of the way that they learned. According to Hsu and Griffiths, these differences resulted from the fact that one model was sensitive to implicit negative evidence while the other model was not. To empirically capture the differences in the way that the two models learned and test these predictions, two training procedures were devised. In the behavioural experiments, two groups of subjects were presented with exactly the same sentences and given exactly the same information about their grammaticality during the experiment. Where the two behavioural conditions differed was in the temporal order in which sentence and grammaticality information were presented. The order in

which this information was presented was reversed between the two conditions.

The *sentence first* group was trained as follows: on each trial subjects were presented with a visual scene and an accompanying sentence on the screen, which was also read out by an adult voice. Subsequently, subjects were asked to guess the sentence's grammaticality and received immediate feedback on their response. In contrast, the *grammaticality first* group was informed prior to training that sentences produced by an adult's voice would always be grammatical, whereas those produced in a child's voice would always be ungrammatical. Training in the *grammaticality first* group thus differed from that in the *sentence first* group in that the sentences were spoken either by adults or children, no explicit responses were required, and the ordering of grammaticality information and sentence presentation was reversed.

An important point to note about these different procedures is that the presentation of sentence and grammaticality information is different. The *sentence first* group received information about grammaticality after encountering the entire sentence whereas the *grammaticality first* group was provided with information about the grammaticality before and while encountering the sentence, i.e. the temporal structure of grammaticality and sentence was reversed. Hsu and Griffiths (2009, 2016) assumed that these two training procedures would lead subjects to adopt different assumptions regarding the way that the sentences were sampled: Either subjects could assume that sentences were sampled from the true language distribution which is beneficial to learning from the absence of observations; Or they could make no such assumption and would then be unable to harness such information. Accordingly, in the behavioural experiment, the critical test sentences served to test whether subjects were able to learn from implicit negative evidence or not.

**Table 1.** The grammaticality of sentences in Experiments 1–3.

(a)	Sentence structure			(b)	Sentence structure			(c)	Position	
	Verb	C1	C2		C3	Verb	C1		C2	C3
V1	+ (9)	+ (9)	– (6)	V1	+ (12)	+ (12)	– (8)	M1	+ (16)	+ (16)
V2	– (3)	+ (18)	– (3)	V2	+ (12)	+ (12)	– (8)	M2	– (16)	+ (16)
V3	+ (18)	– (3)	– (3)	V3	– (4)	+ (24)	– (4)	M3	+ (16)	– (16)
V4	+ (18)	? (0)	– (6)	V4	+ (24)	– (4)	– (4)	M4	+ (32)	? (0)
				V5	+ (24)	? (0)	– (8)			

Notes: + and – indicate grammatical and ungrammatical usage, respectively. The number in parentheses denotes the number of times with which models were trained on a sentence, and the number of times with which subjects were exposed to a sentence during the experiment, respectively. ? indicates that the grammaticality is not determined by the data, i.e. the respective sentence was not presented to models and subjects. These critical test sentences were unseen in training and used to evaluate model predictions by comparing the model's grammaticality judgements to subjects' actual responses. (a) Grammaticality of the four verbs used in Experiment 1 depended on the context in which they appeared (C1: noun-noun-verb, C2: noun-verb-noun, C3: verb-noun-noun). Verb V4 was never shown in sentence structure C2. (b) Grammaticality of the five verbs used in experiment 2. Verb V5 was never shown in sentence structure C2. (c) Grammaticality of contractions in Experiment 3 depended on the four modifiers and their position after the subject (P1) or object (P2) of the sentence. Modifier M4 was never presented after the object of a sentence (P2).

### Modelling sampling assumptions

To simulate and predict subjects' behaviour on this task, Hsu and Griffiths (2009) developed two models that were in turn designed to exemplify two different approaches to classification that have long been distinguished in machine learning (usually referred to as *discriminative* and *generative* classifiers; Ng & Jordan, 2002). Discriminative classifiers learn a direct mapping between some input  $X$  and a set of outputs (or labels)  $Y$ , i.e.  $p(Y|X)$ . From this perspective, discriminative models assume that learners acquire the distribution of labels given an input (in this context, the probability of grammaticality given a particular sentence). By contrast, generative classifiers take a more indirect approach, in which the joint probability of  $p(X, Y)$  becomes available through learning, and then Bayesian inference is used to calculate the posterior  $p(Y|X)$ . From a model capturing the joint distribution, observations can be sampled (in this context, sampling grammatical and ungrammatical sentences) (Figure 1). This formal engineering distinction has prompted various researchers to speculate as to whether a similar dichotomy might apply in human learning, and under which conditions different learning strategies that parallel the discriminative and generative distinction might be observable (Levering & Kurtz, 2015; Love et al., 2015).

#### Sentence first model (discriminative, weak sampling model)

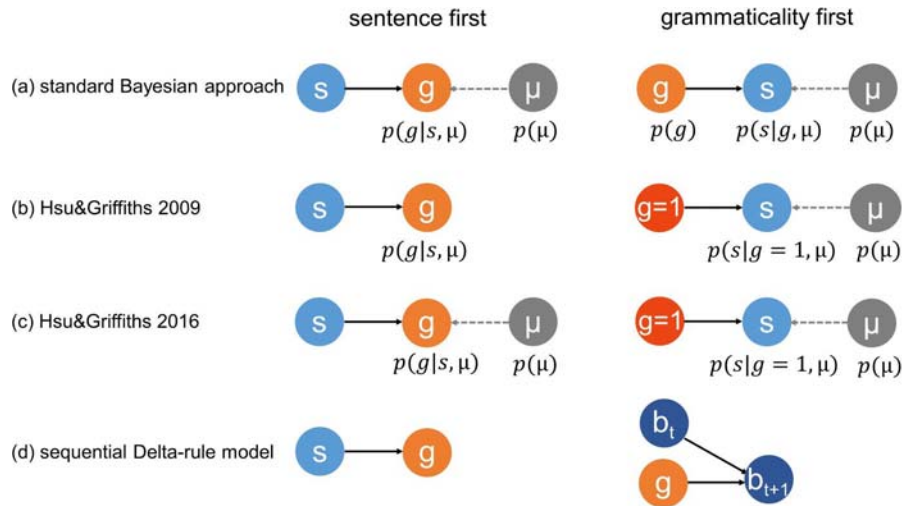
For the discriminative model, Hsu and Griffiths (2009) employed a standard logistic regression model. This model learned the probability of a given sentence being grammatical or ungrammatical. The sentences presented to subjects served as inputs and grammaticality as output to the model. Sentences were represented by a set of binary variables encoding (a) the identity of the verb, (b) its position in the sentence, and (c) the interaction of both (e.g. in Experiment 1, V1-noun-noun was encoded as 1000|100|100000000000, and noun-noun-V2 was encoded as 0100|001|000001000000; Figure 2). These variables predicted the binary outcome of grammaticality. After fitting the model on the sentences shown to subjects in the experiment, its predictions on all possible sentences were used to predict the grammaticality judgements of the *sentence first* group during test. In a series of follow-up studies (Hsu & Griffiths, 2016), the logistic regression model was reformulated within the Bayesian framework by placing priors over the regression coefficients, i.e. the model learned a joint distribution over grammaticality and model parameters conditioned on sentences (Figure 1). In these later studies, Hsu and Griffiths refer to this form of learning as “weak sampling”

(i.e. weak sampling is used here to refer to models that learn to classify observations by assuming that the process that generated observations – in this case sentences – is independent of their classification labels – in this case grammaticality; Xu and Tenenbaum (2007)). However, in all important respects, the discriminative and weak sampling models are the same.

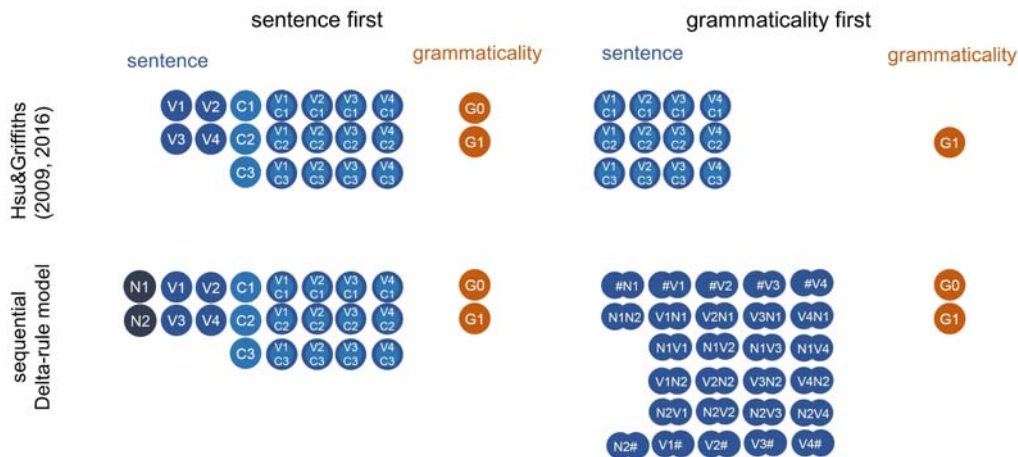
#### Grammaticality first model (generative, strong sampling model)

To account for the behaviour of the *grammaticality first* group, Hsu and Griffiths (2009) employed a hierarchical Bayesian model, the Dirichlet-Multinomial model. The model was trained on vectors of counts representing how often each verb was encountered in a grammatically correct sentence structure (e.g. the information handed to the model in experiment 1 was (9 9 0) for V1, and (0 18 0) for V2, Figure 2). Note that in contrast to the discriminative model, this model was only trained on the grammatical sentences; this is an important representational choice and will be discussed in greater detail below. Accordingly, the model learned a joint distribution over the grammatical sentences and model parameters (not a joint distribution over grammaticality and sentences as may be expected). After training, a threshold was applied to model predictions which converted the probabilities of verb occurrence into a one or zero prediction of grammaticality. These values were used to account for the grammaticality judgements of the *grammaticality first* group. In the later studies, Hsu and Griffiths re-describe this as a “strong sampling” model (i.e. strong sampling is thus used to refer to models that learn to classify observations by assuming that observations – in this case sentences – are generated from a distribution associated with the label – in this case grammaticality – being learned).

These two models thus make different predictions about the way that implicit negative evidence will affect learning: Since the discriminative model is indifferent to the distribution of observed sentences, it cannot exploit information provided by their absence. The absence of a sentence provides no added value to the model. By contrast, learning in the generative model involves estimating the distribution of grammatical sentences. In this case, the absence of a sentence provides weak evidence against its grammaticality, because the generative model implicitly assumes that all sentences that are part of the language will be sampled in the long run. Hence, the two models make diverging predictions about the evidence emerging from absent observations.



**Figure 1.** Schematic illustration of the models presented in Hsu and Griffiths (2009, 2016) (a–c), and the sequential Delta-rule model developed here (d). (a) In a standard Bayesian approach to modelling the task, a discriminative (or weak sampling) and generative (or strong sampling) model would operate on variables representing sentences  $s$  and grammaticality  $g$ . To model the *sentence first* group, a discriminative/weak sampling model would only learn the conditional distribution over  $g$ . In contrast, to model the *grammaticality first* group, at least the conditional distribution over  $s$  given  $g$  would be learned (strong sampling model) or even the full joint distribution (generative model). Note that in either case, the standard approach would capture both grammatical and ungrammatical utterances in the probabilistic model. Models could potentially have priors over their parameters  $\mu$ , however, this factor is not usually considered to be relevant to whether a given model is classified as generative or discriminative. (b) The discriminative (*sentence first*) and generative (*grammaticality first*) models implemented by Hsu and Griffiths (2009). It is important to note that Hsu and Griffiths’ generative model was not modelling the joint distribution of  $s$  and  $g$ , but rather this model was only trained on grammatical sentences. (c) Hsu and Griffiths (2016) modified the discriminative model by incorporating a prior over  $\mu$  (and now referred to the *sentence first* model as a weak sampling model and the *grammaticality first* as a strong sampling model). (d) The sequential Delta-rule model for the *sentence first* condition predicts  $g$  from  $s$ , while the *grammaticality first* model predicts the bigram at time step  $t+1$  ( $b_{t+1}$ ) from  $g$  and  $b_t$ .



**Figure 2.** Representations of the sentence and grammaticality variables in Experiment 1. Importantly, the representation of the sentence variable differed between the two models. In Hsu and Griffiths’ *sentence first* models the presence of a particular verb (e.g.  $V1$ ), sentence structure (e.g.  $C2$ ) and their interaction (e.g.  $V1C2$ ) were represented by binary variables. In contrast, the *grammaticality first* models represented sentences as counts of interactions between verb and sentence structure only. Because the *grammaticality first* model was only trained on grammatical sentences, it followed that grammatical and ungrammatical information were only represented in the *sentence first* model. The sequential Delta-rule *sentence first* model was trained on a sentence representation similar to that in Hsu and Griffiths’ *sentence first* model, but with additional input cues representing the nouns in the sentences. The sequential Delta-rule *grammaticality first* model represented sentences as sets of bigrams. Both grammaticality and ungrammaticality information was represented in the sequential Delta-rule models. Representations for the other experiments were chosen in an analogous fashion.

## Results

As we described earlier, the human subjects in our case study learned about the grammaticality of the sentences in the artificial language, with their training differing in whether they were presented with a sentence before receiving information about its grammaticality, or vice versa. The results of these studies showed that the two experimental groups did indeed judge the grammaticality of novel sentence structures differently. The group that received feedback after presentation of sentences was more likely to judge the critical test sentences to be grammatical than the group that was presented the sentences by adults and children (Figures 3(a), 4(a), 5(a)). Crucially, these differences were consistent with the qualitative predictions of Hsu & Griffiths' models. Accordingly, the authors concluded that the way that participants had learned in the experiments was in turn consistent with the way that the two models learned. That is, that the different training procedures had led subjects to adopt either a strong or weak sampling assumption about the sentences in the experiment and that this had then made learning from implicit negative evidence either possible or impossible. Accordingly, the authors conclude that these results support their hypothesis (i.e. that humans use information about the sampling of their linguistic observations to guide their learning).

### **How does the choice of representations affect model outcome?**

Having described the models, the behavioural task, and the subjects' subsequent performance, we now turn our attention to the way that representational choices made at the modelling stage might influence the degree to which the former can be taken as predictors of the latter. Because the two models made different predictions from one another, and because the two groups of subjects performed differently from one another in much the same way, Hsu & Griffiths' concluded that their subjects' performance reflected the different learning strategies that they took their models to embody. The reasoning underlying this conclusion neatly exemplifies the way in which the theoretical output of computational research ultimately relies on analogies between the workings of models and the processes modelled. In this case, because the behavioural differences observed appear to parallel the different predictions of the models, it seems only natural to conclude by analogy that learning in the models and participants was subject to the same underlying constraints.

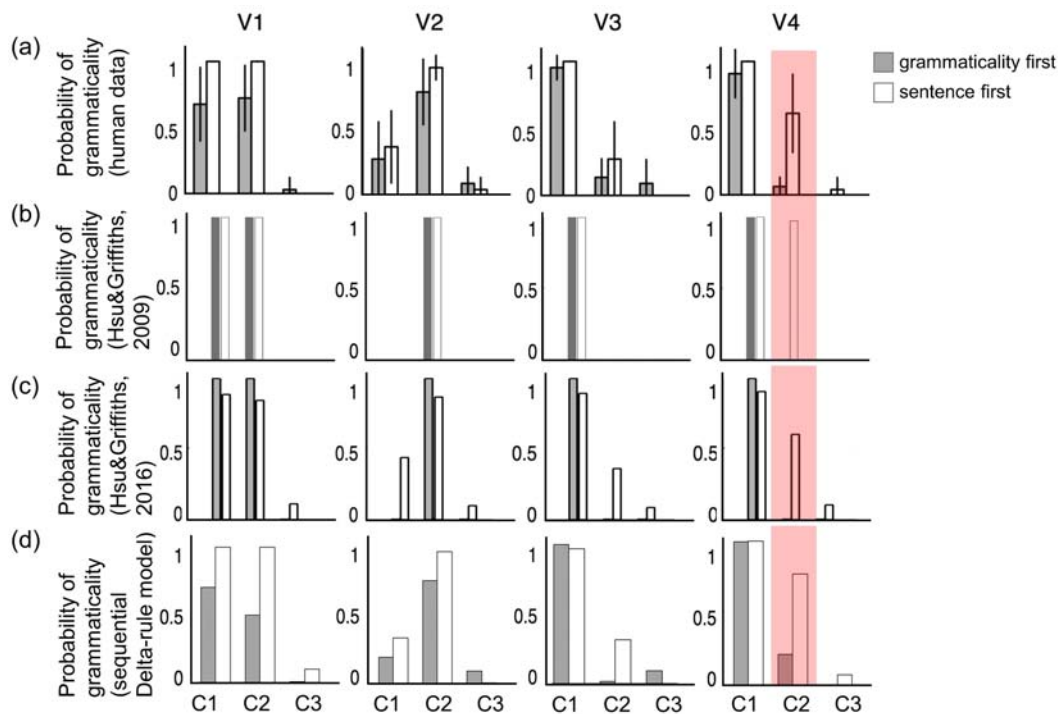
However, this raises a question. Since this analogy is driven by the performance of the subjects and its relationship to the performance of the models, it seems worth asking what, exactly, is driving the performance of the models. Because algorithms and input representations always interact in computational processes it follows that a model's performance (and hence any predictions drawn from it) can not be solely attributed to one or the other of them. It thus follows that when two models make different predictions, while these *might* reflect differences in their underlying computations, they might also reflect differences their representations. Since it is unclear precisely how the brain represents information in any particular context, such as an experimental task, the choice of a particular input representation is potentially a strong determinant of a model's performance.

With regards to our case study, the two models that were supposed to predict empirical performance in the behavioural studies were trained on two rather different input representations, whereas it seems unclear that any analogous differences were actually present in the training given to subjects in the behavioural studies themselves (Figure 4). With regards to the models, the *grammaticality first* model was trained only on grammatical sentences, whereas the *sentence first* model was also trained on ungrammatical sentences. Moreover, the *grammaticality first* model was operating on a different representation of sentences. By contrast, as we noted above, in the empirical study, all of the subjects were exposed to identical sets of sentences and information about their grammaticality. Accordingly, it follows that the differences in the models' predictions on the critical test sentences might actually have resulted from the different representations chosen (in which case any conclusions about different sampling assumptions resulting in the differences in their subjects' behaviour may be unwarranted). These considerations thus raise two questions: How did training the *grammaticality first* model on only a subset of the data (the grammatical sentences) impact its predictions? How did the different representation of sentences impact model predictions? We next address these questions in turn.

### **Training on different subsets of data**

As we highlighted above, while the subjects in our case study received the same information during training, the models used to predict their performance did not. The *sentence first* model was trained on grammatical and ungrammatical sentences, while the *grammaticality first* model was trained on

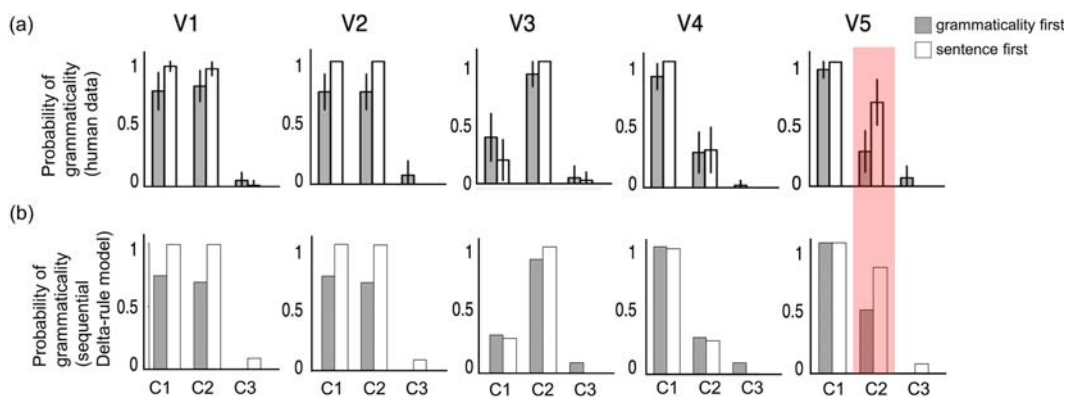




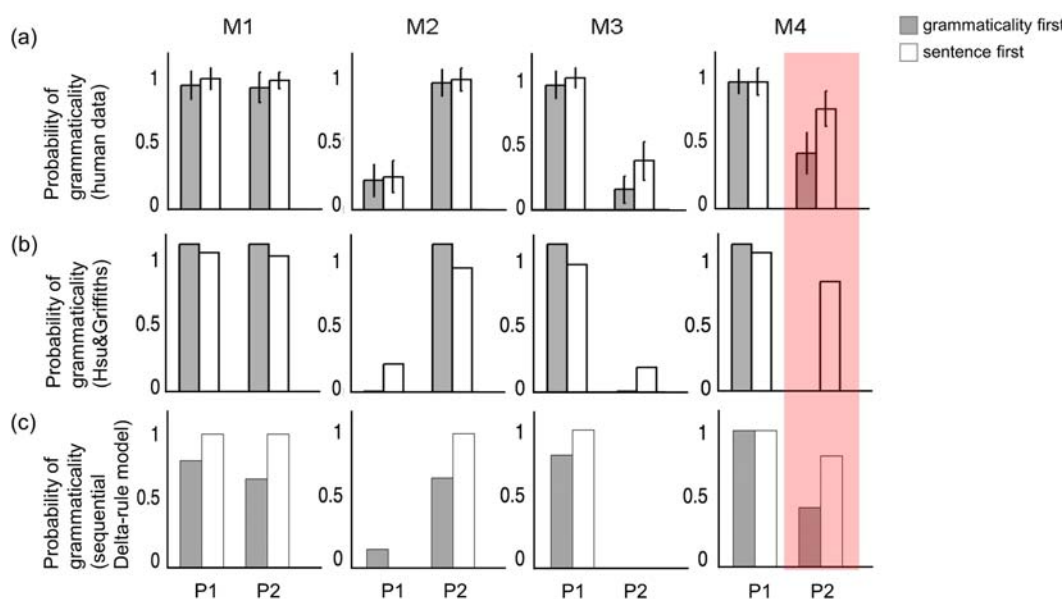
**Figure 3.** Results for Experiment 1. (a) Human judgement of grammaticality. (b) Model predictions by Hsu and Griffiths (2009). (c) Model predictions by Hsu and Griffiths (2016). (d) Predictions from the sequential Delta-rule model. (Figures (a) and (c) were adapted from Hsu and Griffiths (2016); figure (b) adapted from Hsu and Griffiths (2009))

grammatical sentences alone. Hsu and Griffiths justify these representational choices by noting that because some of the subjects trained in this condition later reported that they had ignored the ungrammatical utterances from children, it made sense to train the *grammaticality first* model on only the grammatical sentences. In this regard, it is important to note for current purposes that Hsu and Griffiths (2016) conducted a follow-up experiment where the *grammaticality first* training condition contained only grammatical sentences. In this follow-up experiment, the significant differences that were observed when

the *grammaticality first* training condition did contain ungrammatical sentences *did not* replicate (see Hsu & Griffiths, 2016, supplementary material). Given this, it seems reasonable to conclude that the learning of the subjects in both conditions of the experimental study in our test case was in fact influenced by the ungrammatical sentences. Accordingly, it seems far from clear whether the model of the *grammaticality first* training proposed in this case does in fact adequately reflect the information presented to subjects in the task, or the way it appears to have influenced their behaviour.



**Figure 4.** Results for Experiment 2. (a) Human judgement of grammaticality (figure was adapted from Hsu and Griffiths (2016), model predictions were not provided by the authors). (b) Predictions from the sequential Delta-rule model.



**Figure 5.** Results for Experiment 3. (a) Human judgement of grammaticality. (b) Model predictions by Hsu and Griffiths (2016). (c) Predictions from the sequential Delta-rule model. (Figures (a) and (b) were adapted from Hsu and Griffiths (2016)).

A point that seems particularly relevant in this regard is the role of implicit evidence in learning. Depending on the structure of the learning task, the absence of observations can not only be seen to offer implicit negative evidence (as it seems the authors of our case study intended) but also implicit *positive* evidence. To return to the case study, the fact that the “grammatical” adult speaker never produced the critical test sentence during training provided subjects with weak evidence in favour of its being ungrammatical (negative evidence). However, as a necessary corollary, the absence of the critical test sentence from the utterances of the “ungrammatical” child speaker provided weak evidence in favour of its being grammatical (positive evidence). Given that the investigation in our case study was intended to examine the effects of implicit evidence on language learning, it seems reasonable to suppose that an ideal model of the learning task under consideration ought to have included all of the implicit information available to the learner: positive and negative. Since it seems in this case that the information available to subjects was not fully represented in the *grammaticality first* model, it also seems to follow that the models implemented in our case study fall short of this ideal. Understanding the impact of representational choices on a model’s performance is a critical part of understanding any predictions derived from its behaviour, and in turn its relationship to the empirical behaviour observed in studies testing these predictions. Accordingly, in our case study, if the predictions of the models on the critical test sentences turn out to be

less distinct when both of the models are trained on the same data, this would clearly raise questions about the degree to which they actually shed light on subjects’ learning in this instance. Accordingly, we next turn our attention to the relationship between the workings of the *grammaticality first* model and its input data.

As we have sought to highlight so far, the implementation of the *grammaticality first* model in our case study only allocated significant probability mass to observed grammatical observations. It thus follows that in this model the absence of a sentence implies none, or minimal, probability mass such that it is deemed ungrammatical (i.e. the model predicts that human subjects will rate unobserved sentences as being ungrammatical). However, when a full generative model over sentences and grammaticality is employed this prediction will change significantly.<sup>2</sup> While the probability of the critical test sentence being grammatical would remain low under the full model, the probability of the critical test sentence being *ungrammatical* would now also be low under the full model. In a case where the same number of grammatical and ungrammatical sentences were seen in training, the model would judge the test sentence to be equally grammatical or ungrammatical, and thus would predict that subjects would judge its grammaticality around chance level on average. This is because in the full model, implicit positive and negative evidence would compete. Accordingly and critically, it turns out that in this regard, the predictions of the strong sampling model (i.e. the full *grammaticality first* model trained on all of the data) would now

be very similar to those made by the weak sampling model (i.e. the full *sentence first* model trained on all of the data). This is because in the full *grammaticality first* model implicit negative evidence will only overpower implicit positive evidence if grammatical observations are more frequent in training. To return to the case study, while subjects were exposed to more grammatical than ungrammatical sentences in the behavioural experiments, it appears that the differences predicted by a full generative model would in fact be far less pronounced than those made by the model presented, that was trained on only a subset of the data.

To summarise, while the *grammaticality first* model in our case study was trained on only grammatical sentences, this representational choice was in fact unsupported given other data reported. Further, as we have explained above, the degree to which the two models make different predictions is highly sensitive to the representation of the task and the training set. While it remains the case that only the generative model could learn from implicit evidence from sentences, the factors we have highlighted clearly raise questions about the degree to which the model does indeed predict the behaviour observed in the empirical study, or indeed supports the idea that this behaviour reflects a different sampling assumption that is made by subjects during their training.

### **Training on different sentence representations**

So far we have concentrated on the training sets that were used to model the learning of human subjects in the experiment. As we have sought to highlight, not only were the models presented with different training data, but also sentences themselves were represented differently in the two training sets. On one hand, the *sentence first* model represented sentences as a set of binary variables encoding verb, verb position and their interaction. On the other hand, the *grammaticality first* model only received information about verb position interactions. This representational choice further complicates any interpretation of the relationship between the model predictions and the actual behaviour observed, since it potentially offers an alternative reason for the models' different predictions. To help make this point clear, we next focus on how it was that the *sentence first* model came to make its predictions in more detail.

In the logistic regression model, the contribution of the verb and position variables are determined by whether they were presented on more grammatical or more ungrammatical trials (e.g. since V4 and C2 in Experiment 1 were part of more grammatical than ungrammatical sentences, V4 and C2, by themselves, predict a

sentence to be grammatical, while C3 alone predicts ungrammaticality). The contributions of the interaction variables are then added on top of these individual predictions. It is important to note here that the contribution of the interaction variable representing the verb-sentence combination in the critical test sentence is (near) zero which means that in the critical test sentence this variable barely contributes to the prediction at all.

Accordingly, when the *sentence first* model predicts that the critical test sentences are grammatical above chance level, this prediction is driven by the verb and its position. These two variables make positive contributions towards grammaticality predictions, while the interaction variable has very little impact on the model's behaviour and hence its predictions. By contrast, the *grammaticality first* model is only trained on these interaction terms (represented as counts of which verbs occur in which sentence structures during training). The representation that this model is trained on does not directly encode how often a particular verb, or its position, is grammatical.

Given these representational choices, it was impossible for the *grammaticality first* model to account for evidence that some verbs were more likely to occur in grammatical sentences than others, and the model only indirectly captured that some positions of verbs were more likely to be grammatical than others through the estimation of hyperparameters. Accordingly, when it comes to the critical test sentences, the model actually learned very little about their grammaticality. This is because (a) the critical verb-sentence interaction did not occur in training, meaning that the model assigns near zero probability mass to it, and (b) verb and position were given no, or little weight, in the model, which means that their individual contributions in favour of grammaticality were not fully exploited. It is these factors that ultimately cause the model to classify the critical test sentences as ungrammatical.

Had the evidence from the verbs and their positions been incorporated more explicitly in the representations of the *grammaticality first* model, it might have had a considerable impact on its predictions. Indeed, it appears likely that the incorporation of this evidence would weaken and even overwhelm the influence of implicit negative evidence on the predictions made by the model. As a consequence, it seems that training the *grammaticality first* and *sentence first* on the same sentence representations could result in their making similar predictions about the critical test sentences. These considerations raise further questions about the degree to which the different predictions of the models support the conclusion that the subjects in the

behavioural experiment actually employed different learning strategies. Accordingly, we next turn our attention to this last question, and examine whether the differences observed in the behavioural data can be captured by a single mechanism learning from representations that more faithfully capture the task structure of the empirical studies.

We begin by introducing a simple, widely used learning algorithm, which we train on representations directly derived from the temporal structure of the task. We show that providing an accurate account of the behavioural results described in the case study does not require different learning strategies but rather can be given in terms of a single algorithm applied to different representations that capture the order of information presented to subjects. This in turn suggests that these factors alone may have given rise to the differences observed in the empirical studies. The goal of this exercise is not to provide a solution to the problem of language acquisition, nor is it to advocate that the model we describe is the “correct” one for this task. Rather, its purpose will be to illustrate and explore the importance of representational choice to computational modelling and to underscore how representations are as important as mechanisms when it comes to model interpretation.

### Alternative modelling approach: sequential delta-rule model

It has long been established that the temporal structure of learning tasks can influence what is learned, as well as how quickly learning progresses (e.g. Anderson et al., 2002; Ashby et al., 2002; Levering & Kurtz, 2015; Ramscar et al., 2010; Reips & Waldmann, 2008; Yamauchi & Markman, 1998, also referred to as classification versus inference/observation learning in categorisation tasks). For instance, Ramscar et al. (2010) showed that, as a consequence of the information structure of the task, children are better at learning colour words if the colour word follows the noun it describes, as opposed to preceding it. This finding, and many others like it, indicate that learning outcomes can often depend on the temporal order in which learners encounter information.

It has also long been known that simple error-driven learning models are particularly sensitive to these kinds of temporal order effects (see e.g. Rescorla, 1988; Widrow & Hoff, 1960). In error-driven learning, inputs (typically taken to represent cues or features in the environment) are forced to compete for predictive value over a set of outputs (typically taken to represent outcomes to be predicted, e.g. events in the environment), with the value of individual inputs being

reinforced when they contribute to successful predictions, and decremented when they contribute to prediction errors. As a result of this process of competition, these models learn to assign high weights to diagnostic inputs, and low or even negative weights to non-diagnostic inputs. This process – which causes inputs to compete for value – is, however, sensitive to the order and the structure of sets of inputs and outcomes. If, for example, the outcomes and inputs in a labelling model are reversed, so that only a single input (e.g. a label) now predicts a multitude of outcomes (e.g. features), no competition can take place, and rather than learning about any diagnostic features, a model configured like this will now simply learn the correlations between the label and features (Ramscar et al., 2011). Importantly, these different predictions about how learning unfolds are fully characterised by the prediction errors and the input output sequence, i.e. the model makes no explicit distributional assumptions about the data.

The simplest error-driven learning rule is the Delta rule (Widrow & Hoff, 1960) which implements gradient decent learning, solving a multiple linear regression problem (Stone, 1986). This learning rule has been widely employed to model performance in human learning tasks, such as in artificial grammar learning (Dienes, 1992) and it has often been used to model children’s use of implicit negative evidence in language learning in the past (e.g. Ramscar et al., 2013; Ramscar & Yarlett, 2007).

Given that Delta rule learning models can account (at least in principle) for some of the important characteristics associated with the effects of sequencing on training, we sought to re-evaluate the task demands associated with the empirical results of the case study in order to examine whether the different behaviour observed could be captured by this single, well-established mechanism (which we will refer to as the *sequential Delta-rule model*).

The sequential Delta-rule model of the *sentence first* group in Experiment 1 and 2 employs a representation analogous to the one chosen by Hsu and Griffiths (2016). This included binary inputs encoding the verb, its position, and the interaction of both. Additionally, two invariant and constantly present inputs were used to represent the presence of nouns in each sentence (e.g. input set (N1,N2,V1,C1,V1C1) for sentence noun-noun-V1, Figure 2). The nouns in the model function as a bias term, allowing it to account for the fact that more trials contained grammatical than ungrammatical sentences, such that subjects could be expected to have a bias towards judging sentences grammatically correct. In this model, grammaticality at the output was represented as binary variables. This choice of the

input and outcome is very similar to that of Hsu & Griffiths' *sentence first* model and straightforwardly captures the temporal sequencing of the information in the task. The sentence representations used to simulate Experiment 3 were analogous to the ones used to model the training set in Experiments 1 and 2, except that instead of verbs they encoded modifiers, their positions and interactions as binary inputs with additional inputs for nouns, in keeping with the stimuli employed in the empirical study.

In these empirical studies, the subjects in the *grammaticality first* group were presented with a very different learning problem to those in the *sentence first* group. Grammaticality information was present from the beginning of each trial, which enabled subjects in this condition to continually predict the next words in the sentence and receive error feedback sequentially, in real time, as the speaker uttered each of the words in the sentence. To capture the availability of this evidence to learners in the task structure, we divided sentences into word bigrams, each sentence starting and ending with a null symbol (e.g. noun-V1-noun was split into a set of inputs (#N1,N1V1,V1N2,N2#), Figure 2). Each trial was then modelled by multiple model updates for all individual bigrams. In each update one bigram cue and the cue indicating the grammaticality of the sentence predicted the next bigram (Figure 1).

We optimised a single learning rate  $\alpha$  on qualitative fit for each of the two groups for both the training and the critical test sentences (sentence-first:  $\alpha = 0.2$ ; grammaticality-first:  $\alpha = 0.03$ ) which was identical across all three experiments. All input-output pairs were randomised during training and 1000 simulations run to obtain average predictions. The models were trained using a version of the Delta rule implemented in the *ndl* package in R (Shaoul et al., 2013).

To derive predictions about subjects' response propensities, the differences in activation measures between grammaticality and ungrammaticality after training were calculated. To simulate learning in the *sentence first* group sentences were presented to the model as inputs, and then the activation difference between grammatical and ungrammatical outputs were retrieved and averaged across the simulations. These averages across sentences were then scaled between 0 and 1 to predict the probability of subjects' grammaticality judgment. To simulate learning in the *grammaticality first* group the model was presented with all of the sentence bigrams in sequence, together with the relevant grammaticality input. The activation of all the predicted bigrams was then summed at the output. This served as a measure of sentence predictability under either the assumption of it being grammatically correct or

incorrect. Again the differences between these activations were averaged over all simulations and scaled.

Since we employed the Delta rule with an additional step to map predictions of this model to probability estimates, our models for the two groups can be seen as an approximation to two different (multinomial) logistic regression models (Figure 1): the *sentence first* model that predicts grammaticality from sentences (like the logistic regression model employed by Hsu & Griffiths), and the *grammaticality first* model that predicts the next bigram from the current bigram and grammaticality.

As we noted at the outset, the goal of these simulations was to provide a formal illustration of the importance of considering *all* of the components of a model's implementation in relating it to an underlying theory. We next compare the predictions of our models against the data, and show that they also provide a plausible and accurate account of the empirical results observed in the case study, and thus can be seen as serving to underline the importance of this point.

## Model comparison

A comparison of the grammaticality judgements made by subjects in the case study to the predictions made by the sequential Delta-rule model revealed a close qualitative match across the three experiments (Figures 3–5). The models successfully accounted for the differences seen in the response propensities on unobserved sentences between the two experimental groups, with the magnitudes of the predicted response differences matching those in the different experiments. Thus, the models also appear to successfully capture the effects that emerge from the interplay of explicit, as well as implicit, positive and negative evidence under the different temporal task structures.

In the case study, Hsu & Griffiths focussed their discussion on accounting for the difference of the critical test sentences, however as a further test of our model we examined its ability to fit the empirical data Hsu and Griffiths collected from subjects grammaticality judgements on the sentences that were observed during training. This allowed us to make a fuller evaluation of the performance of the various models described above and provided a richer test set for comparing between them, since obviously an evaluation of these models on a single data point (the critical test sentence) will, by necessity, be very limited in its scope. Evaluation of the models' performance on this data revealed that, the sequential Delta-rule model and the logistic regression model fitted subject performance on the trained sentences equally well (as might be expected

given the similarity between the implementations of the two *sentence first* models).

With regards to the *grammaticality first* group, the sequential Delta-rule model provided a better account of the qualitative features of subjects' performance than the strong sampling model in the case study. For example, the proportions of grammaticality judgements for sentences that were explicitly taught as grammatical ranges more strongly in the *grammaticality first* group than in the *sentence first* group. This pattern can also be captured by the sequential Delta-rule model which mimics the tendency for *grammaticality first* training to result in sentences being classified as grammatical at a higher probability than after *sentence first* training. By contrast, the strong sampling model predicts perfect performance under the same circumstances (this is because in this model, all probability values above a specified threshold were taken to predict perfect performance).

Even though the sequential Delta-rule model is more successful at capturing the overall patterns of the behaviour of the subjects in the case study, some mismatches can still be observed (while these discrepancies may reflect deficiencies in these models as well, it is also of course possible that they reflect deficiencies in the degree to which the experimental design actually succeeded in testing the hypotheses it was supposed to test). A formal model comparison would allow a more detailed analysis of the models than the one presented thus far. However, because we were unable to access Hsu and Griffiths' original data, and because our goal here is not to advocate that the sequential Delta-rule model is the correct model for this task, or even to enter into a debate about whether the idea of there being "correct" models is a good one in the first place, we will not concern ourselves with this possibility here.

Rather, since our aim is to examine how and why it is that the representations and algorithms/mechanisms chosen by modellers can result in different models that provide equally plausible fits of the same data, and thereby support very different conclusions about any underlying processes, we shall focus on these points. For this purpose, it is sufficient that Hsu & Griffiths' models and the sequential Delta-rule models can all account for the main qualitative aspects of the data, such that although it is clear that the two sets of models make use of input representations that ultimately make different assumptions about the information necessary to capturing the constraints on learning that were relevant to this task, it could be argued that either serves to offer a plausible explanation of the data when considered in isolation.

Given that the sequential Delta-rule models appear to offer a different perspective into human processing than those proposed in the case study, it is worth examining in more detail how these models differ from Hsu & Griffiths' models. In doing so, we shall seek to better understand exactly why it is that they support different conclusions about the mechanisms that gave rise to the data that they are fitting, and to highlight the critical role that the conceptualisation and treatment of representational entities in experimental tasks can play in the modelling of behaviour.

### **Model interpretation**

We begin our detailed discussion of the way that the representational and algorithmic choices embodied in the different models under discussion served to shape the conclusions about learning processes drawn from them by first comparing the workings of the sequential Delta-rule model to those of Hsu & Griffiths' models. We then discuss how Hsu & Griffiths' conceptualisation of the task variables generates different hypotheses about the underlying learning strategies. We then explain why it is that the sequential Delta-rule model indicates that a single mechanism is sufficient to explain the data.

### **Dual mechanism interpretation**

In their models, Hsu and Griffiths make a clear conceptual distinction between what they consider to be the output that is to be predicted (grammaticality) and what they consider to be the input (sentences). This classification of what counts as output and input is fundamental to the way that the models are then classified as either being strong sampling (generative) or weak sampling (discriminative). Although both models learned probability distributions over some variables, the question of whether or not they captured a distribution over input (sentences) determined their classification into the two model classes.

To further explain this point, it is important to note that technically the logistic regression (*sentence first*) model is as much a strong sampling model as is the Dirichlet-Multinomial (*grammaticality first*) model. The former serves to capture a distribution over grammaticality (strong sampling assumption on grammaticality) while the latter serves to capture a distribution over sentences (strong sampling assumption on sentences). Hsu and Griffiths' classification of these two approaches into weak and strong sampling models only emerges under a particular conceptualisation – and hence representation – of the task. It was because the sentences were always

conceptualised as input that the models ended up as being described as either weak sampling or strong sampling, simply because the classification of the models was determined by the way that they captured a distribution over the sentence variable. However, there is in principle no reason why both models should not be described as strong sampling models, albeit expressing strong sampling assumptions about different variables.

These considerations appear to shed some light on the origins of the hypotheses that were originally examined in our case study. Hsu and Griffiths' original hypotheses seem to have been inspired by the distinction between generative and discriminative models that is commonly made by engineers (Ng & Jordan, 2002), which has also prompted other researchers to examine whether analogous differences in learning can be found in humans. In this vein, empirical studies have offered evidence that seems to suggest that discriminative models (i.e. models with weak sampling assumptions on training items) might better fit training where an explicit label or response feedback follows presentation of an item, and that generative models (i.e. models with strong sampling assumptions on items) better fit training where a label precedes or accompanies the presentations of an item (e.g. Levering & Kurtz, 2015). By analogy, Hsu & Griffiths assumed that their two experimental training procedures would have the same effect on the sampling assumptions made by the learners in their task. Accordingly, in their design, in one condition, subjects observed an item, made a response, and received feedback in terms of the item being labelled as grammatical or not. In the other condition, subjects were first given a grammaticality label and where then presented with an item. However, it is far from clear that this simple manipulation is sufficient to cause learners to make different assumptions about the way that the input they are exposed to is being sampled. Indeed, it is equally unclear why this manipulation should preclude learners from building a generative model over the sentences in both conditions, just as it is equally unclear exactly what kind of manipulation would prompt subjects to employ one or the other of these statistical sampling assumptions in a task.

In other words, notwithstanding the analogy to the distinction between generative (strong sampling) and discriminative (weak sampling) models that is routinely made in machine classification, *a priori* there is no principled distinction to be drawn between the structure of the models employed by Hsu and Griffiths and no principled link to be drawn between the manipulation in the task structure of their training experiment and the distinction between the sampling assumptions that it was

supposed to invoke. That is, it is unclear that the task design explains the structure of the models, just as it is unclear that the structure of the models explains the task design. Instead, it seems that the technical differences in the models are offered as a candidate explanation of what gave rise to subjects' different behaviour as a kind of analogy.

Thus to summarise, both of the models proposed by Hsu and Griffiths (2016) can be validly described as strong sampling models with respect to some of their variables. It is only in relation to the way that the sentences used in training are represented in the models – in both of the models the sentences were invariably conceptualised as “inputs” – that the models fall into the classes of weak sampling (discriminative) versus strong sampling (generative). Accordingly, to the extent that these models add support to the idea that subjects' in the experiments were using different learning strategies, this support relies on an analogy between what the models learned and what subjects' learned. Yet the distinction that this analogy hinges on results entirely from the way that conceptual choices made by the modellers caused the models to be classified.

### **Single mechanism interpretation**

In building the sequential Delta-rule model, we sought to take a less top-down approach to the conceptualisation – and representation – of the task structure, and in particular, to take advantage of differences that were clearly – and objectively – present in the two training regimes. Rather than predetermining what counted as input or output in the representation of the training set, the sequential Delta-rule model represents task variables as either inputs or outputs depending on when, exactly, the sentential or grammatical information conveyed by the various stimuli occurred in time. Given that the models in the case study embodied the assumption that two learning strategies were required to account for the empirical data, we were particularly interested to examine whether this data could be accounted for by a single learning mechanism trained on representations derived from the temporal structure of the task. By linking the model structure more tightly to the task structure, we sought to examine whether the different ordering of inputs and outputs in the two experimental conditions was, by itself, sufficient to provide a plausible account of the data.

Because subjects were presented with grammaticality information after the sentences were presented in the *sentence first* condition, it follows that this information

could not have been used to inform subjects' learning as the sentences unfolded in time. By contrast, learners in the *grammaticality first* condition were informed in advance as to whether the sentence they went on to hear were grammatical or not. Accordingly, learners in this condition would have been able to take advantage of any sequential differences that exist between grammatical and non-grammatical sentences as they occurred in time. Thus, while the representation of the sentence variable in the *sentence first* condition in the sequential Delta-rule model largely corresponds to that employed by Hsu and Griffiths in the case study, the representation of the *grammaticality first* condition was reconstructed so as to better capture the temporal impact of the grammaticality variable in this training condition.

Thus, in this model sentences were re-represented as sequences of bigrams that served as outcomes of preceding bigrams and then as inputs for following bigrams. It is perhaps worth pointing out that these differences in the representations of the two *grammaticality first* models further serve to illustrate the challenge modellers face when translating experimental variables into model variables because representational choices will rarely be either obvious or clearly objective.

It is further worth noting that the kind of representation employed in the sequential Delta-rule model could easily be transferred to a probabilistic model of the kind favoured by Hsu and Griffiths, where it could be expected to yield similar predictions. From this perspective, one could think of the sequential Delta-rule model as a non-probabilistic approximation to two conditional distributions, for which the nature of the distributions now entirely emerges from the characterisation of the time structure of the problem, rather than being an ad-hoc decision of the experimenter as to what is an output and what is an input (Figure 1).

Another point worth noting that emerges from these considerations is that explicitly probabilistic models are not the only learning algorithms capable of exploiting implicit evidence. That is, the explicit sampling assumptions embodied in generative models may be sufficient for learning from the absence of observations, but they are not necessary for this purpose. Any algorithm that represents expectations of the presence of observations (for example, a Delta rule model) is capable of learning from their absence, which means that it is also capable of learning from implicit negative evidence (see also Ramscar et al., 2013).

In summary, we have described two sets of models which make very different assumptions, both about the learning process and, perhaps as a consequence,

the information that was considered relevant to modelling it in the conditions reported in the case study, and the information that was included in the input representations to the models. Nevertheless, as formulated, both of these sets of models appear to be able to provide plausible accounts of the empirical results observed (at least insofar as grammaticality judgments can be taken as evidence of the subjects' actual learning in the experiment). We have also shown in detail how ultimately the performance of these models depends not only on their different learning assumption but also on the way that the task structure and task information was represented within them.

The dual mechanism account emerges from a top-down conceptualisation of the variables in the task as being either input or output regardless of the actual structure of the training condition. In the case study, this then led to the classification of two models as either being weak or strong sampling, and by analogy this was then offered as an account of the empirical data. By contrast, we have shown that when the temporal order of information is allowed to determine the representations in models of these training conditions, then the behaviour that was interpreted as providing evidence for two distinct learning strategies can be seen to be the product of a single learning mechanism. These various models thus serve to reveal how different conceptualisations of a task (which given the intimate relationship between data structures and algorithms we noted at the outset, can be expected to influence almost every aspect of a model's implementation) can in turn lead to very different conceptions of the underlying representations and mechanisms that give rise to human behaviour. These considerations thus also raise questions about how models that operate on different representations and algorithms are to be compared. In the discussion that follows, we will not try to answer these questions but we will at least try to sketch some of the considerations that any answers will have to address.

## Discussion

The production of models of ever-increasing detail and specificity is a basic goal of the scientific process. While verbal models, for example, are remarkably flexible, they are also inevitably imprecise. From this perspective, the benefits provided by computational models are obvious: they force researchers to commit to concrete algorithms and input representations, and in return allow them to make quantitative, rather than merely qualitative, predictions and statements. However, although mathematical models offer the



promise of precision, delivering on that promise hinges to a large degree on our understanding of the models themselves, and the relations between models and the systems that they are taken to exemplify. In this regard, it is notable that although much attention in the language and brain sciences is paid to *how* and *why* the brain computes, answers to these questions cannot be developed independently of the representations that are the subject of these computations. This consideration is particularly important in relation to the concrete implementations of computational models, and the way that they are interpreted and compared.

For example, models of a learning process can make different assumptions about the distribution from which a target variable is being sampled. These assumptions may either be strong or weak and might roughly relate to the distinction made between generative and discriminative classifiers in the domain of machine learning, such that models from one domain can end up serving as analogical models in another domain, computational cognitive science. This seems to have been the strategy that Hsu and Griffiths (2009, 2016) adopted in modelling the artificial grammar learning task in the case study discussed above. However, as we have sought to show in our analysis of these models, the relationships between models and the phenomena they seek to capture at these various levels of abstraction inevitably hinge on analogies, such that caching out the promise of specificity in computational modelling inevitably hinges on the accuracy of these analogies. Do the correspondences and relations posited stand scrutiny and to what degree?

In analysing these models against the conclusions drawn from them, we have sought to illustrate how a degree of slippage will inevitably arise even in computational models because when it comes to modelling the “same” task using different paradigms, the nature of the relationship between processes and representations that serve as their inputs means that it is unlikely that two different models will represent the same task in its entirety in exactly the same way.

We illustrated this point by showing how two empirical patterns previously associated with different learning strategies, which might in turn be taken to reveal different learning mechanisms, could be equally plausibly explained within a single learning framework. Indeed, we might even suggest that this latter model offers a simpler and more economical explanation to that originally proposed in the case study, and also, because the algorithm we used in the sequential Delta-rule models appears to be broadly compatible with the neural structures involved in learning in the

brain (see e.g. O’Doherty et al., 2003; Schultz, 2006), it may provide a better analogy to the processes actually involved in language learning.

Having said this, however, it is worth reflecting on some of the limitations of all of the models described above in this regard. The models in the case study and the sequential Delta-rule models both treat “grammar learning” as a process of learning how to combine discrete form elements into larger messages, yet it is far from clear how well this conception of the task reflects what children actually do. While adult speech production and comprehension *can* be described in combinatorial terms, the processes that best characterise children’s communicative learning are *discriminative* (rather than combinatoric or compositional), because in order for a child to interpret speech as sequences of discrete forms (at various levels of description), they must first learn to discriminate those forms from a more or less continuous input in which the mappings between the forms and physical signals (i.e. sound waves) are far from straightforward (in listening to speech, language users not only routinely use context to infer – and “hear” – form elements that are not present in the signal, but also, again depending on context, routinely interpret the same physical signals as representing *different* forms; Ramscar and Port (2016)). It is thus far from clear that the alphabetic/combinatoric characterisation of language used in these models is at all a reflection of the true nature of the task faced by language learners.

What is important to note for current purposes is that there are other models one might conceive of here, such that there may in fact be better ways of modelling and interpreting the same empirical phenomena and data. The results from the sequential Delta-rule model serve to underline this point by showing how it is possible for a single learning mechanism to give rise to two different learning outcomes. In doing so, it illustrates an obvious yet easily overlooked fact: It does not follow that because a model can capture some of the observed performance that results from an underlying process that the model actually capture the underlying process itself.

Accordingly, we do not intend to suggest that our analyses offer proof of a single mechanism theory of learning in this task. Nor do we necessarily disagree with many of the abstract descriptions of the learning process that the original models described in our case study here sought to capture. For example, it is clear that expectations, and violations of these, in the form of implicit negative evidence, are very valuable aspects of learning. Indeed, as we noted above, in this regard – at an abstract level – the sequential Delta-rule model

is similar to the models originally proposed by Hsu and Griffiths (2009, 2016) in the case study.

Further, just as we have shown that the predictions of those original models depended to a large extent on the input representations they employed, the predictions of the sequential Delta-rule models are also sensitive to the representations chosen and would necessarily change if different choices were made. The important point here is that these choices need to be made explicit, they need to be justified, and the plausibility of these justifications needs to be examined. Because different models can behave similarly, while the interpretations of these models can differ widely, the process of drawing conclusions from models of cognitive/neural processes will inevitably depend on details that are devilishly difficult to ascertain, and this problem is exacerbated when these details are taken for granted or left implicit, as is often the case with input representations in cognitive models.

It is to these details that we will now turn our attention. It seems clear that theories of language and cognition need to pay as much attention to input representations as they do to model architectures. It also seems clear that great care should be taken when applying tools from machine learning to problems in biological learning, not least because machine learning models are typically applied to pre-defined representations which currently have no clear analogues in the brain. We will expand on these points below.

### **Representation, algorithm, and abstract model**

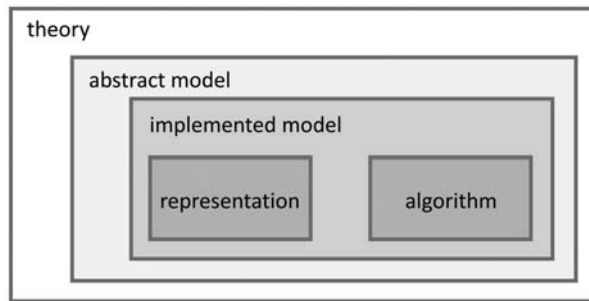
As we have noted throughout this paper, a model's predictions can never be arrived at independently of the input representations that are chosen for it. Our results provide one example of how the interactions that necessarily occur between representations and the mechanisms operating on them can lead to situations where the exact same behavioural data can be explained by very different models. In doing so we have sought to emphasise how important choice of representation is to modelling, and we have also sought to highlight how representational choices tend to be ignored when modellers of cognitive and neural processes theorise about the relative contributions of algorithms, computations, etc. to the more abstract processes they seek to understand.

One reason for this may be the many and varied uses to which the word *representation* is put in the brain and cognitive sciences. We have focussed here on just one of these uses: the strictly computational notion of a representation as an input to an algorithm. We noted a different conception of this word above when discussing

how the task of language learning is to be conceptualised: one might *represent* the task facing children in purely combinatoric terms, or one might also *represent* this task in a way that also incorporates the discriminations that lead to the elements that need to be combined in the first place. (Adding to the confusion, this use of *representation* to capture a high-level conceptualisation of the process under consideration is often referred to as characterising it at the *computational* level.) Then again, *representation* is also used as being the end of an investigation itself, when researchers seek to characterise the intermediate, internal representations that are thought to mediate between higher and lower-level cognitive processes. For example, in the categorisation literature, there has been much debate on whether 'concepts' are internally represented as prototypes or exemplars. What is interesting about this debate for current purposes is that advocates for theories on either side of it tend to assume the same input representations to their models of the categorisation process, such that from the current perspective, this debate is not so much about input representations as it is about the best algorithms for capturing human performance in categorisation tasks (Love, 2003). Given the promiscuous uses to which the word is put in the field, it is perhaps hardly surprising that as one switches between literatures, it is not always clear what is meant by *representation* or that different aspects of the ideas that relate to it are often lost in the mix.

To explain why the contributions of representations, algorithms, and computations will only rarely manifest themselves in fully independent ways (as suggested by Marr, 1982), it is important to recognise that in practice, models in the brain and cognitive sciences are typically presented in one of two distinct ways: either as *abstract* model descriptions, or as *implemented* models.<sup>3</sup> Abstract model descriptions typically comprise symbolic (i.e. verbal or algebraic) descriptions of the relationships between what are typically quite loosely defined quantities or entities. Accordingly, while abstract models can appear to be more or less "formal", they typically fail to fully specify representations (what exactly will be counted and in which format) and typically fail to fully specify the algorithms that will transform these representations into predictions (Figure 6). It is in fact only when these latter steps are made, and an abstract model is actually implemented, that it can be considered formal in any meaningful sense.

Further, because abstract models are almost always conceived of within theoretical frameworks that in the brain and cognitive sciences inevitably embrace a particular computational metaphor, and because these frameworks shape the way problems and solutions are



**Figure 6.** An implemented model can only make predictions from concrete representations and algorithms. The choice of these is shaped by the theory or framework in which the model is conceived of and interpreted, along with any experiments to test its predictions, etc. Critically, in practice, these choices are rarely fully specified by the theory or the abstract model. Understanding the resultant influence of these choices is critical to understanding the actual theoretical contribution of any implemented model.

conceptualised and interpreted, it follows that the modelling process can only really bring forth the clarity it promises if all the decisions embodied in an implemented model are properly described, analysed and discussed before any theoretical conclusions are drawn.

In this regard, it is notable that the problem of specifying input representations for models of neural/cognitive function appears to apply more generally to many different fields of research. For instance, recent work in category learning (Roark et al., 2020), perceptual learning (Zaman et al., 2020) and visual memory (Schurgin et al., 2020) supports a more general case that many well-established theories about underlying processes have made assumptions about the representation of physical stimulus properties as inputs that on examination are problematic. Firstly, it has been noted that assumptions made in one stimulus domain (e.g. visual) can fail to generalise to another (e.g. auditory), such that previous models fall short of explaining more general phenomena, and secondly, it has been noted that when representations that reflect parts of the brain’s transformation of the physical input are incorporated into models, be it on a group-level or even in individual subjects, this can drastically change their predictions and hence the interpretations about mechanism that they support.

This is a basic problem, and the difficulties it presents inevitably increase whenever there is disagreement about the actual function of cognitive mechanisms themselves. To return to the models in our case study and our reformulation of them, the important differences between them ultimately extend far beyond their technical implementations, because ultimately these models embody different metaphors of how

cognition works. All of the major theoretical frameworks in the brain and cognitive sciences embrace computational metaphors, and it follows from this that all abstract models will tend to be biased towards particular conceptualisations of problems that are shaped by particular algorithms and particular kinds of representational choices, which lead in turn to particular kinds of interpretations of findings (Figure 6). For example, Rescorla and Wagner introduced their take on the Delta rule as an elementary model of learning in which inputs and outcomes are associated. The representation of these inputs and outputs were defined by the experimenter based on their intuitive understanding of the features of the environment available to a learner. While this model could explain many learning phenomena in animals, there are results that the model cannot account for and more complex algorithms have been proposed to account for these in turn. For example, one finding that appeared to speak against the Rescorla-Wagner rule is retrospective revaluation (e.g. backward blocking). While other algorithms to account for this finding have been proposed within the elementary representational framework, Ghirlanda (2005) has shown that this problem can be resolved by a Delta rule model simply by changing the representation, and assuming that the brain never represents input cues as fully separate entities (a representational format that elemental models, by definition, tend to assume). In the same vein, Ramscar et al. (2010, 2013) have argued that because all Delta rule models implement a form of discriminative learning, it makes little sense to assume elemental representations in error-driven learning models, and that representation in these models should be conceptualised in abstract terms, based on the dimensions of the environment that a learner needs to discriminate, as opposed to the supposed “features” it contains.

In summary, implementing a model necessarily involves a commitment to a specific model architecture and a specific representation of the task. While these choices are influenced by the theory, they are rarely, if ever, governed by it. This inevitably leaves even implemented models open to interpretation, and when it comes to these interpretations any part of the model – input representation, algorithm, etc. – is as relevant as any other.

### ***Input versus label***

As we noted above, theorising in the brain and cognitive sciences leans heavily on computational metaphors. This is particularly apparent in relation to learning, where the study of biological learning increasingly borrows from

machine learning. The problems this can bring were illustrated in the original models discussed in our case study, where a commitment to characterising human learning in the generative/discriminative terms employed in machine classification appears to have led to the concomitant adoption of a somewhat restrictive view of what counted as being input or a label in the learning task, which inevitably led to a particular set of conclusions. In standard machine classification, this issue does not arise. The goal of machine classification is the labelling of inputs, such that what counts as an input and what counts as a label is never in question. As we have sought to show, once machine classification is related by analogy to human learning the question of what exactly is an input or an output becomes more complicated.

To return to the models of Hsu and Griffiths, it is far from clear exactly which events count as labels to the brain, i.e. which events are in fact targets to predict, just as it is far from clear which events offer support for these predictions, i.e. act as input. That is, as we noted above, there is no a priori reason to suppose the brain would represent the grammaticality or sentence information in the learning tasks in the case study by analogising them to the various parts of machine learning models. Accordingly, it follows that when engineering methods are applied by analogy to less well circumscribed domains like human learning, researchers must be careful to mind the gap.

This is not to say that probabilistic models and other techniques from machine learning are not useful tools to be applied to understanding language and language learning. Rather, it is to emphasise a point that we have sought to highlight throughout. As we noted above, all models are metaphors, and their scientific utility ultimately stands or falls in their value as analogues of the phenomena they sought to capture. In a similar vein, techniques from machine learning can only be applied to phenomena in human learning by analogy, and thus the value of these analogies ultimately depends on the degrees of correspondence that can be established between the two. When it comes to the representation of human learning in models, establishing these correspondences is a difficult and subtle task.

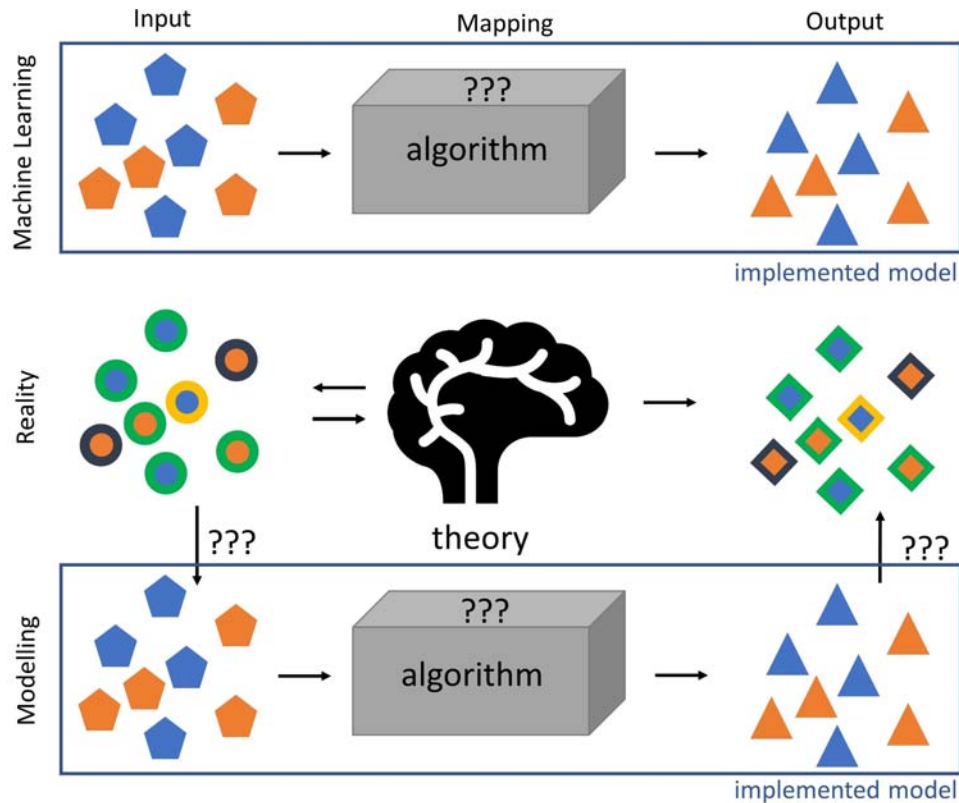
### ***Crafting representations***

These considerations point in turn to yet another difference between machine and human learning models. In machine learning, numerous techniques have been developed for efficiently analysing data and optimising the prediction of future data points. For the purposes of engineering, the usefulness of a model can be

defined straightforwardly. All that matters is its performance on a particular problem, and many engineering choices, such as the selection of an appropriate algorithm, can be made on a fairly objective basis, depending on the data and domain structure in question (Figure 7). For this reason, in machine learning, the pre-processing of data, hand-crafting of features, or automating feature selection in the algorithm itself, can all be entirely legitimate steps. This is because in engineering, if Model A can perform a well-defined task faster, more accurately, or more efficiently than Model B, then A is a better model than B.

By contrast, the goals of any cognitive/neural model are inevitably far less well defined. Rather, because the goal of these models is to act as analogues for poorly understood processes, it follows that in an ideal world every aspect of a model should stand on as sound a theoretical basis as is possible. That is, given the underspecified nature of the problem being fitted – human learning and behaviour in all its complexity – tweaking input representations and algorithms to improve the “performance” of a specific model without considering all of the myriad ways in which underspecification necessarily compromises cognitive modelling, might not only be unwarranted, it may risk derailing the whole enterprise altogether.

To return to the subject of our case study, language – and in particular the idea of what is to be considered an input in language learning – is a particularly interesting domain to think about in this regard. Much of linguistics is occupied with the characterisation of its structure and the nature of linguistic representations at various levels of abstraction. While linguistic abstractions can be described at syntactic, morphological, phonemic levels etc., it is far from clear that any of these notional representations exactly capture human communicative processing at an actual “functional” level. Rather, all of the information these abstractions seek to characterise seem to be processed by the brain in an almost holistic fashion, and it is far from clear that these traditional levels of linguistic description are as separable in reality as they might appear in theories. From a computational perspective, this poses enormous challenges. As we have noted throughout, representational choices are critical to the understanding of computational models yet when it comes to language, we are faced by an apparently hierarchically organised structure that simultaneously provides no obvious entry point from a representational perspective. For example, most modellers may be familiar with the idea that “words” and “meanings” are hard to define, but in fact this problem applies at every conceptual level of description, from high-level “grammatical regularities” to low-level



**Figure 7.** Machine learning versus cognitive modelling. In machine learning, the data arrives in a pre-defined representation. The challenge the field faces is that of finding a suitable abstract model and efficient algorithm to be able to make the best possible predictions on unseen data. Using these tools to account for learning in biological systems imposes additional challenges: (a) the actual nature of the input representations to the learning mechanisms employed in the brain at the various stages of processing are unknown (because sensory systems massively transform the raw input received from sensory receptors and there is no obvious representational entry point). Thus the mapping between input and output of the biological and artificial learning system are an essential part of the implemented model and typically influenced by the underlying theory, (b) the learning mechanism of the brain is unknown and thus the goal of abstract and implemented models is no longer optimal data prediction, but rather to serve as analogues for largely unobservable and underspecified processes in the brain.

“phonetic features” (Ramsar & Port, 2016; Samuel, 2020). Such is the nature of human communicative codes that it is impossible to fix upon a given representational level or to define representational units, without making simplifying assumptions that can do more to determine the contribution of a given model than any other subsequent, seemingly “formal”, choice (Figure 7). Moreover, as we have sought to illustrate in the concrete modelling examples discussed above, these are not just philosophical problems, but rather it is likely that these issues affect most modelling in the brain and cognitive sciences most of the time.

### Model comparison

To recapitulate: Representational choices can significantly alter the performance of a model, the predictions it makes and thus the way it is interpreted. Given that these choices are invariably underdetermined in cognitive/neural models this poses problems when it comes

to the evaluation and comparison of these particular models. The problem is fairly easily stated: because representational and algorithmic choices *are* invariably underdetermined, it follows by necessity that models can only be mapped onto theory by analogy, such that in modelling the problem of mapping by analogy is present at every formal level of description. As such, a formal comparison of two models will never in fact represent a formal comparison of the actual abstract models themselves, but rather a formal comparison of their implementations, and because the relationship of these implementations to underlying theory depends on analogies that could conceivably be better or worse at either the representational or algorithmic level, the theoretical implications of better or worse fits in model comparisons will inevitably be underdetermined. Is a better performing implementation a better model in this case? Ultimately, the answer to this question depends on one's belief in the underlying analogies supporting the various aspects of the implementation. For

example, consider a hypothetical case where a formal comparison of two implementations shows model A to be better than model B (e.g. in terms of generalisation performance). One can easily conceive of situations where the algorithm implemented in model B is far more biologically/psychologically plausible than that of A, but where the representation used as the input to the algorithm in model B was chosen poorly/specified inaccurately, such that what is driving the difference in performance of the models is not in fact the hypothesised mechanisms themselves but merely the inputs into them. Which is actually the better model? Given the complexity of exactly determining the contributions of different representations and algorithms, and of comparing these contributions between implementations, it is clear that providing a definitive answer to this question will be difficult, and may even be impossible. If it really is the case that cognitive/neural models are ultimately analogues, then this conclusion ought not to surprise us, since formal model comparison in this instance seems to equate to formal metaphor comparison, which seems something of a contradiction in turn.

Because no formal solution to this problem currently exists, or is even likely to exist, this points to an uncomfortable conclusion. Despite the apparent formality of cognitive/neural models, it appears that where processing or representational assumptions differ significantly (as for example, the two *grammaticality first* models in our case study), the best we can currently hope to achieve with model evaluations and comparisons will have a qualitative rather than quantitative flavour; and at worst, this might represent a necessary consequence of the methods currently employed in the field. If we accept this conclusion, and at present it seems we must, it follows that despite their apparent formality, cognitive/neural models should be presented in ways that facilitate and help maximise the effectiveness of qualitative comparisons. This means, for example, that many of the choices that modellers currently make and present implicitly need to be made more explicit. In particular, it helps highlight the importance of justifying representational choices, of considering alternate representations, and of considering the impact of specific representational choices on specific predictions. In short, it highlights the need for modellers to be more verbose about their representations when reporting on models. Bringing input representations into the focus of discussion will do more than merely help others better understand the workings of models. Rather, it is likely that it is only by including representations to their considerations of other aspects of their models that modellers will in fact be able to properly justify any conclusions that they draw from them. Formulating

these kinds of exhaustive model descriptions will, admittedly, be a challenging and time consuming task. However, given the analogical role that models play in this domain, eschewing this task will only result in confusingly mixed metaphors.

This is not to say that formal methods for model analysis and comparison have no part to play in this process. Far from it. Models' abilities when it comes to fitting and predicting empirical data are clearly important for the purposes of comparison and evaluation. Indeed, as we noted above, in many cases, models will share their basic computational and representational assumptions, and existing methods for model comparison will often be sufficient for these purposes. Indeed, given the way that science works in practice (most actual work is done in support of pre-existing theories and models, as supposed to proposing new theories and models), it likely follows that most models in the literature share either their algorithms or their representational assumptions (or both), with other models.

However, in what we suspect will be the majority of theoretically interesting examples of model comparison – those cases where two models *do not* share their basic computational and representational assumptions – traditional formal methods of analysis and comparison will fall short for the reasons we have described above. Other methods and approaches will be required. We have already stressed the need for making modelling and implementational decisions/descriptions (and the theories that inform them) more explicit when models are reported. One promising example of this approach is where empirical information/processes are explicitly used in the development of input representations themselves. For example, many models of spatial navigation assume that different aspects of the environment are represented by specialised cell types, a strategy that at least allows for some representational assumptions to be tested empirically (Barry et al., 2006; Hartley et al., 2000). Another approach that seems promising in this regard comes from machine learning, where tools are being developed that allow researchers to infer their input representations from and test them against large datasets. For example, tools can allow for representations to be inferred (e.g. from behavioural data, or the performance of DNNs) which can subsequently be correlated against further test sets of brain and behavioural data (Battleday et al., 2019; Houlsby et al., 2013; Hsu et al., 2019; Ma & Peters, 2020; Sanders & Nosofsky, 2020; Schatz et al., 2019; Yamins et al., 2014; Zheng et al., 2019). From the perspective described here, this appears to be a promising direction of research, since it offers the possibility of ultimately empirically constraining the search space for representations, and might even lead

to the development of tools for objectively testing some representational choices, in some domains at least.

## Conclusion

Computational models of cognitive and neural processes occupy a curious intellectual niche because they model unknown processes whose correspondences even at an input and output level are typically weakly constrained/largely underdetermined (how exactly do the “categories” in a model of categorisation correspond to anything in human thought processes?; what exactly is ‘language’ in a model of language learning?). As a consequence it follows that in most cases the correspondences between virtually every aspect of these models and the things that they stand for/represent can only be inferred at best. On the whole, these models serve their function more as metaphors than simulations, such that it is less the case that computational models are fitted to cognitive/neural processes than it is that these models shape our understanding of what these processes are. Accordingly, this means that these models are inevitably developed within larger computational frameworks which can shape and even determine the choices scientists make when selecting the actual representations and algorithms that they implement in their models. Although both of these choices can have a strong influence on a model’s performance and predictions, as compared to the attention that algorithms and broader frameworks have received in the literature, the representations that encode the inputs to these algorithms have been curiously neglected. Typically, the inputs to processes are simply taken for granted, such that the nature and influence of their representations, and the degrees of freedom associated with their choice, have been little analysed or discussed in the literature. This has often led to situations where algorithms have been either rejected wholesale as models for a given functional process at one extreme, with algorithmic features being reified unnecessarily at the other (see e.g. Fodor & Pylyshyn, 1988).

At the outset of this paper, we noted that our goal was not to try to formally establish whether or not one of Hsu and Griffiths’ models or our own was correct. As the forgoing hopefully makes clear, this is because to the best of our knowledge there is no formal way of establishing this, because the two sets of models embodied different representational and computational choices, and no formal methods for capturing and quantifying these differences exist. Instead, we have sought to use a qualitative comparison of these models to highlight the apparently unique problems that the interactions between input representations and algorithms

pose in the brain and cognitive sciences, to describe the problems that representational flexibility poses to model evaluation and comparison, and to give one example of what a more qualitative solution to this problem might look like. The adage “All models are wrong, but some are useful.” (Box, 1979) is well known to modellers. In a similar vein, when it comes to the mind and brain it is clear that all representations are also wrong, and that establishing that a representation is useful is a far harder task than many modellers appear to appreciate.

## Notes

1. “While scientists and philosophers have on the whole taken diagrams for granted, they have been forced to fret at some length about the nature and function of models. Few terms are used in popular scientific discourse more promiscuously than ‘model’. A model is something to be admired or emulated, a pattern, a case in point, a prototype, a specimen, a mock up, a mathematical description – almost anything from a naked blonde to a quadratic equation – and may bear to what it models almost any relation of symbolization.” (Goodman, 1976, p. 171)
2. Note that the concrete choices and implementations of the two models as exemplars to contrast generative and discriminative model predictions in general is somewhat misleading. This is because the models are imperfect counterparts, a result of the fact that the generative model was only conditioned on grammatical sentences. In a more standard approach to modelling this task one might think of a generative model over sentences and grammaticality rather than generative with respect to sentences and parameters (Figure 1). This point is further blurred because Hsu and Griffiths suggest at other times that it is only the difference in sampling assumptions on sentences that they interpret to be the cause of the different learning outcomes. These variations on their hypotheses notwithstanding, the concerns we raise about the representations chosen for the models apply regardless.
3. Although many researchers like to distinguish between the “algorithmic” and “computational” levels in modelling neural and cognitive processes, it is far from clear that this distinction can ever be perfect in practice. The fact that computations are inevitably defined as transformations of variables means that it is generally impossible to describe a computation without committing to variables. While domains with clearly defined variables may pose no problem in this regard (such as mathematics, where real numbers can serve as well-defined, discrete, variables such that a computation as for example “addition” can be analysed independently of the representation of the variables it transforms and the precise algorithm used, c.f. Marr), when it comes to the brain and cognitive sciences, the natures of the units, concepts, etc. being transformed are unknown. This means that in studies of the brain, the defining of “computations” invariably requires researchers to

commit to variables – representations – that are always empirically and theoretically underdetermined. Accordingly, the distinction between “algorithmic” and “computational” level only “works” insofar as the problem of defining variables – making representational choices – is ignored. To some degree, this problem can be fudged at an abstract level, but when computational models are used as analogues to cognitive/neural processes it will inevitably raise its head. All computational models require the definition of a specific set of variables for the task to be modelled, and also a definition of a precise encoding of these variables, and a precise definition of the algorithm transforming them. Given that any computational model is in fact a conjunction of an input representation and an algorithm, and given that the brain’s representations cannot be taken as given – meaning that all of the definitions just mentioned will be massively underdetermined empirically and theoretically – it is highly unlikely that the “computational” and “algorithmic” levels assumed by Marr can ever be decoupled in practice in the way that researchers often appear to assume. Further, it seems likely that the reason behind this important point being generally overlooked in the brain and cognitive sciences is the general myopia towards representational issues that we have sought to highlight here.

## Acknowledgments

The authors thank two anonymous reviewers for their helpful comments that have much improved the manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Gatsby Charitable Foundation and Max Planck Society (F. B.).

## ORCID

Franziska Bröker  <http://orcid.org/0000-0001-6707-984X>

Michael Ramscar  <http://orcid.org/0000-0003-1680-1112>

## References

- Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, 30(1), 119–128. <https://doi.org/10.3758/BF03195271>
- Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30(5), 666–677. <https://doi.org/10.3758/BF03196423>
- Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O’Keefe, J., Jeffery, K., & Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, 17(1–2), 71–97. <https://doi.org/10.1515/revneuro.2006.17.1-2.71>
- Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2019). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. Preprint arXiv:1904.12690.
- Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press. <https://doi.org/10.1016/B978-0-12-438150-6.50018-2>
- Brown, R., & Hanlon, C. (1970). Derivational complexity and order of development in speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). Wiley.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, 16(1), 41–79. [https://doi.org/10.1207/s15516709cog1601\\_2](https://doi.org/10.1207/s15516709cog1601_2)
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71. [https://doi.org/10.1016/0010-0277\(88\)90031-5](https://doi.org/10.1016/0010-0277(88)90031-5)
- Ghirlanda, S. (2005). Retrospective reevaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 31(1), 107–111. <https://doi.org/10.1037/0097-7403.31.1.107>
- Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sproule and Almeida (2013). *Language and Cognitive Processes*, 28(3), 229–240. <https://doi.org/10.1080/01690965.2012.704385>
- Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett Publishing.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364. <https://doi.org/10.1016/j.tics.2010.05.004>
- Hartley, T., Burgess, N., Lever, C., Cacucci, F., & O’Keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, 10(4), 369–379. [https://doi.org/10.1002/\(ISSN\)1098-1063](https://doi.org/10.1002/(ISSN)1098-1063)
- Houlsby, N. M., Huszár, F., Ghassemi, M. M., Orbán, G., D. M. Wolpert, & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, 23(21), 2169–2175. <https://doi.org/10.1016/j.cub.2013.09.012>
- Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 754–762). Curran Associates, Inc.
- Hsu, A., & Griffiths, T. L. (2010). Effects of generative and discriminative learning on use of category variability. In *32nd annual conference of the cognitive science society*, Portland, Oregon, USA.
- Hsu, A., & Griffiths, T. L. (2016). Sampling assumptions affect use of indirect negative evidence in language learning. *PLoS ONE*, 11(6), e0156597. <https://doi.org/10.1371/journal.pone.0156597>
- Hsu, A. S., Martin, J. B., Sanborn, A. N., & Griffiths, T. L. (2019). Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people.



- Behavior Research Methods*, 51(4), 1706–1716. <https://doi.org/10.3758/s13428-019-01201-9>
- Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, 71(4), 571–592. <https://doi.org/10.1086/423752>
- Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, 43(2), 266–282. <https://doi.org/10.3758/s13421-014-0458-2>
- Love, B. C. (2003). Concept learning. *The Encyclopedia of Cognitive Science*, 1, 646–652. <https://doi.org/10.1002/0470018860.s00499>
- Love, B. C., Ramscar, M., Griffiths, T. L., & Jones, M. (2015). Generative and discriminative models in cognitive science. In *Proceedings of the 37th annual meeting of the cognitive science society*, Pasadena, California, USA.
- Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior. Preprint arXiv:2005.02181.
- Mahowald, K., Graff, P., Hartman, J., & Gibson, E. (2016). SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3), 619–635. <https://doi.org/10.1353/lan.2016.0052>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., M. S. Seidenberg, & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356. <https://doi.org/10.1016/j.tics.2010.06.002>
- Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (pp. 841–848). MIT Press.
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2), 329–337. [https://doi.org/10.1016/S0896-6273\(03\)00169-7](https://doi.org/10.1016/S0896-6273(03)00169-7)
- Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT press.
- Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mice in adult speech. *Language*, 89(4), 760–793. <https://doi.org/10.1353/lan.2013.0068>
- Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE*, 6(7), e22501. <https://doi.org/10.1371/journal.pone.0022501>
- Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53, 58–74. <https://doi.org/10.1016/j.langsci.2015.08.002>
- Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6), 927–960. <https://doi.org/10.1080/03640210701703576>
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6), 909–957. [https://doi.org/10.1111/\(ISSN\)1551-6709](https://doi.org/10.1111/(ISSN)1551-6709)
- Reips, U. D., & Waldmann, M. R. (2008). When learning order affects sensitivity to base rates: Challenges for theories of causal learning. *Experimental Psychology*, 55(1), 9–22. <https://doi.org/10.1027/1618-3169.55.1.9>
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151–160. <https://doi.org/10.1037/0003-066X.43.3.151>
- Roark, C. L., Plaut, D. C., & Holt, L. L. (2020). A neural network model of the effect of prior experience with regularities on subsequent category learning. PsyArXiv.
- Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111, 104070. <https://doi.org/10.1016/j.jml.2019.104070>
- Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 3(3), 1–23 DOI: 10.1007/s42113-020-00073-z. <https://doi.org/10.1007/s42113-020-00073-z>
- Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., & Dupoux, E. (2019). Early phonetic learning without phonetic categories – Insights from machine learning. PsyArXiv.
- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, 57(1), 87–115. <https://doi.org/10.1146/annurev.psych.56.091103.070229>
- Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, 4(11), 1156–1172. <https://doi.org/10.1038/s41562-020-00938-0>
- Shaoul, C., Arppe, A., Hendrix, P., Milin, P., & Baayen, R. H. (2013). *NDL: naive discriminative learning* [computer software manual]. R package version 0.2.14. <http://CRAN.R-project.org/package=ndl>
- Stevens, C. F. (2001). An evolutionary scaling law for the primate visual system and its basis in cortical function. *Nature*, 411(6834), 193–195. <https://doi.org/10.1038/35075572>
- Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In *Parallel distributed processing: explorations in the microstructure of cognition* (Vol. 1, pp. 444–459). MIT Press.
- Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits*. Tech. Rep. Stanford Univ Ca Stanford Electronics Labs.
- Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3), 288–297. <https://doi.org/10.1111/desc.2007.10.issue-3>
- Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39(1), 124–148. <https://doi.org/10.1006/jmla.1998.2566>
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>
- Zaman, J., Chalkia, A., Zenses, A. K., Bilgin, A. S., Beckers, T., Vervliet, B., & Boddez, Y. (2020). Perceptual variability: Implications for learning and generalization. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-020-01780-1>
- Zheng, C. Y., Pereira, F., Baker, C. I., & Hebart, M. N. (2019). Revealing interpretable object representations from human behavior. Preprint arXiv:1901.02915.