Thank you for using our service!

Interlibrary Services
The Ohio State University Libraries
(614)292-9077
osuill@osu.edu

Article Express documents are delivered 24/7 directly to your ILLiad account from scanning libraries around the world. If there is a problem with a PDF you receive, please contact our office so we might report it to the scanning location for resolution.

NOTICE
WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use,” that user may be liable for copyright infringement.

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.
Distributional learning is error-driven: the role of surprise in the acquisition of phonetic categories

Paul Olejarczuk / Vsevolod Kapatsinski / R. Harald Baayen

1 University of Oregon – Linguistics, 1290 University of Oregon, Eugene, OR 97403, USA, E-mail: vkapatsi@uoregon.edu
2 University of Tuebingen, Tuebingen, Germany

Abstract:
Much previous research on distributional learning and phonetic categorization assumes that categories are either faithful reproductions or parametric summaries of experienced frequency distributions, acquired through a Hebbian learning process in which every experience contributes equally to the category representation. We suggest that category representations may instead be formed via error-driven predictive learning. Rather than passively storing tagged category exemplars or updating parametric summaries of token counts, learners actively anticipate upcoming events and update their beliefs in proportion to how surprising/unexpected these events turn out to be. As a result, rare category members exert a disproportionate influence on the category representation. We present evidence for this hypothesis from a distributional learning experiment on acquiring a novel phonetic category, and show that the results are well described by a classic error-driven learning model (Rescorla, R. A. & A. R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (eds.), Classical conditioning II: Current research and theory, 64–99. New York, NY: Appleton-Century-Crofts).

Keywords: predictability, categorization, phonetic categories, distributional learning, error-driven learning, Rescorla-Wagner model

DOI: 10.1515/lingvan-2017-0020

Received: May 11, 2017; Accepted: August 18, 2017

1 Introduction

In modern linguistic theory, phonetic categories refer to structured mental representations of the contrast-bearing speech sounds in one’s language. For instance, English speakers are said to represent a /p/ category separately from a /b/ category. Like other categories, phonetic categories have graded internal structure: some instances of /p/ and /b/ are relatively typical while others are marginal (Kuhl 1991; McMurray et al. 2002; Miller 1994; Pisoni and Tash 1974). An individual speech sound can be described as a point in a multidimensional perceptual space. For example, any member of the /p/ category is defined by a set of values along several continuous phonetic parameters, including VOT, burst spectrum, formant transitions, and the F0 of neighboring vowels. When constructing their internal representations, both infants and second language learners appear to respond to the way in which these points cluster – or distribute – along the relevant dimensions. The bottom-up analysis of input structure hypothesized to be involved in phonetic category acquisition has accordingly been named distributional learning.

In distributional learning, learners are said to engage in a sort of unsupervised cluster analysis: densely populated pockets of perceptual space are grouped together while sparsely populated regions are treated as boundaries between the groups. Evidence for sensitivity to distributional information comes mainly from two behavioral phenomena. The first is perceptual warping: two stops are more likely to sound different if they were first presented as part of two different clusters of VOT values than if exposure consisted of a single cluster spanning both values (Maye and Gerken 2000; Maye et al. 2002; Maye et al. 2008). The second phenomenon is generalization behavior: categories defined by highly variable input distributions are more readily extended to novel tokens (Zhao 2010).

While these results support the idea that input distributions matter, they provide only a rudimentary understanding of the mapping between these distributions and the acquired category representations. The most we can conclude from them is that human learners are sensitive to the modality and variance of a distribution. This leaves many questions unanswered. Consider the role of frequency: since input distributions are defined in terms of relative frequencies, the very notion of distributional learning entails that each perceived instance
of a speech sound has some bearing on the category. But does each instance contribute equally to the category, or is the magnitude of its influence contingent on some factor?

The most influential theories of categorization assume that each perceptual experience contributes equally to the strength of a representation. This assumption is made explicit in exemplar models, where the category is a set of memory traces of individual experiences (e.g. Nosofsky 1986; Ashby and Waldron 1999; see Johnson 1997; Pierrehumbert 2001; for sound categories). It is also implicit in prototype theories, where categories are represented as parametric summaries of experienced distributions (e.g. Flannagan et al. 1986; Feldman et al. 2009), because each instance contributes equally to the parameter estimates. This ‘one exemplar, one vote’ view amounts to a linear mapping between input frequency and category structure, where representations are derived from raw counts. Another way to put this is to say that category learning is essentially Hebbian in nature: the strength of the association between a region of perceptual space and a category is directly proportional to the number of times this region is experienced and interpreted as belonging to that category (Hebb 1949).

These ‘one exemplar, one vote’ category models are at odds with psycholinguistic work in other domains. For instance, various measures of word recognition difficulty are a decelerating function of frequency, often modeled by the log transformation (e.g. Howes 1957; Kreuz 1987). Thus, Howes (1957) reports the drop in the intelligibility threshold to be a constant 4.5 decibel per logarithmic unit of word frequency. In a log-transformed frequency distribution, differences among relatively rare events are emphasized while differences among relatively frequent events are compressed. In other words, experience yields diminishing returns: the more frequent a word already is, the less it benefits from additional exposure (see also Forster and Davis 1984; Mandera et al. 2016). In fact, practice has been known to yield diminishing returns since at least Bryan and Harter (1899) work on telegraph operators at the end of the 19th century. This deceleration of learning is naturally captured by error-driven learning models. Unlike in simple Hebbian learning, the amount of learning engendered by an event in an error-driven model is proportional to the magnitude of surprise, or prediction error, that the event elicits (Rescorla and Wagner 1972). When an event unexpectedly occurs in a particular context, the learner associates features of that context with the event. As an event continues to be encountered in the same context, its occurrence becomes less and less surprising, and the weights are increased in progressively smaller increments.

In this article, we ask whether distributional learning of phonetic categories is essentially Hebbian, or essentially error-driven. If phonetic category learning is Hebbian, we might expect the structure of the internal representation to reflect the raw frequency distribution, whether as a fine-grained histogram assumed by exemplar models (e.g. Pierrehumbert 2001) or a set of parameters (e.g. mean and variance; Flannagan et al. 1986) summarizing the raw counts. On the other hand, if distributional learning is error-driven, the representation (whether histogram-like or parametric), will be based on the log transformation of the input distribution. On this account, infrequent exemplars that are associated with a category will exert a disproportionately large influence on the representation. This is because rare events are more surprising than frequent events and thus generate more prediction error, which leads to more learning per exposure.

We investigate the mapping between input frequency and category structure in a distributional learning experiment featuring a novel phonetic category. Two groups of adult participants were exposed to the same number of tokens distributed differently along a single phonetic dimension. After training, both groups were asked to rate the typicality of experienced exemplars, and to provide productions representative of the learned category. Both the ratings data and the production results suggest that the categories inferred by learners were based on log frequency rather than raw counts. This in turn suggests that distributional learning is error-driven rather than Hebbian in nature. To corroborate this interpretation, we model the ratings data with the Rescorla-Wagner model, a classic learning model that incorporates the effect of surprisal on learning rate (Rescorla and Wagner 1972; see Baayen et al. 2011; Arnold et al. 2017; Milin et al. 2017, for other recent applications). The model provided a good fit to the results, supporting the error-driven account of phonetic learning.

2 Method

Sixty-six University of Oregon undergraduates participated in the experiment in exchange for course credit. Each participant self-identified as a native, monolingual English speaker with no known hearing impairment and no appreciable exposure to any tone language.

The target category was a lexical tone consisting of a rise-fall (LHL) superimposed on the monosyllable /ka/. The carrier syllable, synthesized in MBROLA (Dutoit et al. 1996), was 715 ms in duration and its segmental properties remained constant across all tokens. The magnitude of the pitch excursion (the difference between L and H) varied along a continuum. The F0 manipulation was performed using the PSOLA algorithm as implemented in Praat (Boersma 2001). Following the pitch adjustment, all tokens were amplitude-normalized to 70 dB.
The participants were divided into two training groups of 33 learners. Each group was exposed to a different training distribution consisting of 550 /ka/ tokens. The Left distribution featured a negative skew, with most of the tokens containing relatively large pitch excursions; the mode was at 14ST. The Right distribution was its mirror image, with a mode at 8ST and a tail extending over the higher excursions. Stimuli ranged from 4ST to 17ST for the left-skewed group and from 5ST to 18ST for the right-skewed group, in 1ST increments. A pilot experiment indicated that perceptual resolution was constant across the continuum; participants reliably discriminated stimuli that were one or more semitone apart in magnitude of the pitch excursion. The two training distributions are illustrated in Figure 1, where panel (a) shows raw counts and panel (b) shows log-transformed frequency on the y-axis.

![Figure 1: The training distributions in raw frequency (a) and log frequency (b) space. The vertical lines represent distribution means. Note that the means overlap in the left panel but not in the right panel, where they are closer to the tails (see main text for discussion).](image)

The experiment was administered individually in a quiet behavioral laboratory using E-Prime software. Participants wore headphones and sat facing a computer screen. Training consisted of passive auditory exposure to the 550 tokens in random order. Participants were told that the stimuli were resynthesized from actual productions of several speakers of a fictitious tone language, and that all productions represented the same lexical item (i.e., all tokens belonged to the same category). They were instructed to listen without engaging in any other activity. The trials advanced automatically after a 500 ms interval; training lasted approximately 13 min.

The test phase immediately followed training. Participants were presented with /ka/ tokens, with pitch excursion (the difference between the L and H of the LHL) ranging between 1 and 21 semitones in 0.5ST increments, once each in random order (41 trials total), and instructed to rate the category goodness of each item using a serial response box. Typicality ratings are a common way to investigate graded category structure (e.g., Miller 1994), and correlate highly with perceptual magnet effects in category recognition (Kuhl 1991; Iverson and Kuhl 1996). In the present experiment, the ratings were constrained to a 7-point Likert scale. Once the ratings task was completed, participants were asked to provide three productions of /ka/ that were to be as representative as possible of the category. The productions were recorded by a condenser microphone positioned on the desk in front of the participants.

If we summarize the two training distributions based on raw frequency, their means and variances are identical (Figure 1a). On the other hand, if our parametric summaries are based on log frequency, the means of the two distributions are not equal but slightly shifted toward the tails (Figure 1b). This gives rise to specific predictions with respect to the typicality ratings. If frequency is represented veridically and learners average over the input, the ratings distributions of both training groups should completely overlap, centering over the common mean (the dotted, vertical line in Figure 1a). If repetition has diminishing returns and learners average over the input, the ratings of the two groups should not overlap but shift toward the tails and center over the respective means of the log-transformed distributions. Finally, if categories consist of histogram-like representations, the shapes of the training distributions should be recoverable from the ratings, with maximal goodness over the modes. With respect to the production results, the predictions are similar. If productions target category means, veridical frequency representation predicts an identical target for both groups, while the diminishing-returns account predicts somewhat higher pitch excursions in the Right group relative to the Left group. If productions instead target the category modes (e.g., Buz et al. 2016), then the Left group should shoot for higher excursions.

### 3 Results

All analyses were performed in R using linear mixed effects models provided in the lme4 package (Bates et al. 2015). The results of the rating and production tests are shown in Figure 2. Panel (a) shows the ratings for
each magnitude of pitch excursion, z-scored and averaged across the learners in each group. The plot suggests that the ratings distributions do not overlap completely, indicating a difference between the groups. Panel (b) illustrates this difference more clearly. Here, the ratings are fit with a loess smoother and shown without error bars to reduce visual clutter. With the superimposed training distributions serving as points of reference, it is clear that the goodness curves are of similar shape, but appear to be shifted relative to each other along the x-axis. Crucially, typicality ratings do not resemble a histogram of the learning experience or maxima over the modal values (cf. Buz et al. 2016). Instead, the ratings are slightly shifted toward the tails, as predicted by the error-driven account (compare the shift with the log-frequency means in Figure 1b).

Figure 2: Results of the experiment. Left-skewed distribution in red; right-skewed in blue. Panel (a): actual ratings; error bars are 95% CI. Panel (b): loess-smoothed ratings over training distributions. Panel (c): mean rise magnitudes produced by each training group.

To test for the significance of this apparent shift, the z-scores were predicted by the interaction of training group and centered pitch excursion (modeled as a parabola). The model included by-participant random slopes for the within-participant predictor. Of specific interest was the interaction of the linear term of the parabola with training group: a significant effect would support the horizontal shift. The model indicated that this interaction was indeed significant ($\beta = 2.02, t(64.5) = 4.19, p < 0.001$), lending support to the error-driven account.

Panel (c) on the right side of Figure 2 shows the production data. Recall that each participant provided three utterances of the LHL tone category. For this analysis, the magnitude of the F0 rise (LH) portion of each utterance was measured and converted to semitones relative to the initial L. This conversion normalized the measure on an utterance-by-utterance basis. As seen in the bar graph, the Right group appeared to have larger pitch excursions as their production targets than the Left group. This group difference was significant according to a mixed-effects linear model with random intercepts for participants ($\beta = 1.75, t(60.45) = 2.11, p < 0.05$). Thus, the pattern seen in production is consistent with the ratings results, providing converging evidence for the error-driven hypothesis.

4 Modeling error-driven distributional learning

To test whether distributional learning could be subserved by an error-driven learning mechanism, we modeled the ratings data seen in Figure 2a,b with the Rescorla-Wagner model (Rescorla and Wagner 1972), the most widely used error-driven learning model in the associative learning literature (see also Ramscar et al. 2010; Baayen et al. 2011; Milin et al. 2017, for recent applications to language acquisition). The model is a simple, two-layer network in which a layer of input units (called cues) is directly and fully connected to a layer of output units (outcomes). The model learns to predict the outcomes from the cues. Error-driven learning is used to update the connection weights. The crucial property of this kind of learning algorithm for the present purposes is that an observed event causes the learner to update beliefs about the cues to that event, and that beliefs are updated in proportion to how surprising the event is. The more surprising an event, the more one should update one’s beliefs about the circumstances in which it occurs, so that it is less of a surprise in the future.

A learning event consists of a subset of the cues occurring with a subset of the outcomes. Upon this pairing, the weights on the connections between the present cues and all outcomes connected to them undergo adjustment. The model derives a prediction for every outcome by summing the weights of the connections from the present cues to that outcome. To the extent that a present outcome is surprising (i.e. its activation is not strongly positive), the weights of connections from the present cues to that outcome are adjusted upward. Conversely, weights of connections from present cues to unexpectedly absent outcomes are adjusted downward. Formally, the weight from a cue C to a present outcome O at time t+1 is increased via the equation in (1), while the weight from a cue C to an absent outcome is decreased using the equation in (2).
where $\lambda$ and $\Lambda$ are free parameters, while the activation $a$ of the outcome $O$ at time $t$ is the sum of cue weights from all present cues to the outcome in question. Following common practice, the $\Lambda$ parameter, representing the maximum activation of an outcome, was set to 1. The $\lambda$ parameter, representing the learning rate, was varied (see below).

When an outcome occurs, weights from the present cues to that outcome are increased to make it more expected in the future, and weights from the present cue to all other outcomes are decreased. The existence of the $(\lambda - a_O^t)$ and $(0 - a_O^t)$ terms is what makes the Rescorla-Wagner model error-driven: if the outcome that occurs on a given trial was fully expected to occur given the preceding cues, its activation is equal to $\lambda$ while activations of all other possible outcomes are equal to zero and no learning happens. The closer an outcome’s activation is to zero, the more unexpected it is. As it moves in the direction of 1, the maximum (mean equilibrium) value, it is more and more expected. Hence, the adjustments to the weights become increasingly smaller.

In our implementation, both the cues and the outcomes are LHL excursion magnitudes. The model was trained on exactly the same data as the participants, including trial order (which was randomized and differed for each participant). Its task was to predict a stimulus from the three trials that preceded it. Because the trial order was random, the preceding trials are actually not predictive of the upcoming stimulus, and modifying their number has no effect on the model. Furthermore, the model is insensitive to perceptual similarity between the stimuli because excursion magnitude is treated as a nominal variable. For these reasons, the only factor that drives the predictability of a particular pitch excursion value is the long-term relative frequency of that pitch excursion in the input. The present simulation is therefore intended to demonstrate that error driven learning predicts the current results solely by imposing a decelerating relationship between frequency and expectation.

We evaluated this prediction by determining how strong the weights of connections to each outcome are across the cues, i.e. the sum of the absolute values of cue weights for each particular outcome (the column 1-norm of the weight matrix; Arnold et al. 2017; Milin et al. 2017). The column 1-norm is a measure of prior availability of an outcome, reflecting the strength of the network’s beliefs about that outcome. In recent studies of word recognition, it has been shown to be an excellent predictor of lexical decision latencies, one that outperforms simple word frequency (Milin et al. 2017).

The left panel in Figure 3 presents the modeling results when the learning rate parameter is set at 0.25, with the column 1-norm plotted against pitch excursion magnitude. As seen in the panel, the model succeeds at predicting the mean shift observed in the ratings data: the mean of the right-skewed distribution in blue is to the right of the mean of the left-skewed distribution in red. Note also that the shape of the distributions approximates Figure 1b, with relatively fat tails: this is the surprisal effect.

A mixed-effects model evaluating the effect of skew on predicted scores for each subject modeled using a second-degree polynomial indicates that skew is predicted to shift the category along the horizontal axis in the expected direction. In other words, the difference between the two vertical lines in Figure 3 is statistically reliable ($\beta = 0.24; z = 4.55, p < 0.001$). The right panel of Figure 3 plots the correlation between the model predictions and the observed ratings, averaged across participants. Linear regressions revealed a significant model effect for both groups (Left: $\beta = 0.13$, $t(11) = 4.95$, $p < 0.001$, adjusted $r^2 = 0.66$; Right: $\beta = 0.16$, $t(11) = 2.52$, $p < 0.05$, adjusted $r^2 = 0.31$).
5 Discussion

Having argued that distributional learning of phonetic categories is an error-driven process, we now explore the effect of varying the learning rate (\(\Lambda\)) parameter of the network model. Figure 4 shows the model predictions if learning is so slow as to prevent activations from approaching \(\lambda\) by the end of training (the training data are identical to above). Clearly, the model is sensitive to rate: slow learning results in a faithful reproduction of the input.

![Figure 4: Model predictions for a slow learning rate.](image)

The lack of a mean shift in Figure 4 is explained by the fact that some amount of learning is required in order for the frequent outcomes to become expected. Recall that the equation in (1) decreases weights of connections to an outcome only if the activation of the outcome is strongly positive. Because all weights in the simulations were initialized at zero, the activations were also close to zero. For this reason, the weights grow positive before some of them turn negative to fine-tune expectations by context. With the slow learning rate in Figure 4, very few negative weights have time to develop, and the column 1-norm is virtually identical to the sum of input activations. With the faster learning rate in Figure 3, many weights are strongly negative. The negative weights themselves do not appear to be particularly important for the prediction of peak shift observed in our data: deriving model expectations by summing over only the positive weights leads to the same predictions. However, it does appear important for the learner to have learned enough to form strong expectations about both when a stimulus will occur and when it will not occur. For the error-driven nature of the Rescorla-Wagner model to make a large difference in predictions, activations of outcomes on training trials must be approaching the lambda attractor. 8

In addition to learning relatively quickly, it is crucial that the model predicts upcoming sounds, i.e. category members, rather than predicting the category identity, or making inferences on its basis. With predictions proceeding in this direction, individual category members are outcomes. As discussed above, expectedness of the outcome determines the degree of belief updating in the model, strengthening connections to outcomes that are unexpected and making tokens of unexpected category members particularly influential. We consider this kind of prediction to be likely in our task: the learner can’t help but predict upcoming auditory events, even when not tasked to do so – and especially when not tasked to do anything else. As argued by Baayen et al. (2011), this is likely why learners keep track of transitional probabilities in auditory input (Aslin et al. 1998). When the learner does not predict category members, but rather uses them to predict something else, we expect the learner to instead learn the most when experiencing a category member that has been highly discriminative/predictive in his/her prior experience (e.g. Ramscar et al. 2010; Dye et al. 2017). In many situations, the most discriminative category members are precisely the most expected ones: ones that are known to reliably cue the category. Therefore, when learners are asked to make predictions based on category members, they should learn less – rather than more – from exposure to peripheral members of the category. See Kleinschmidt (2016: Ch. 4) for evidence consistent with this prediction. 9

While the model is constantly trying to predict upcoming auditory stimuli based on preceding ones, its expectations are continually being disconfirmed in our experiment because the order of stimuli in our training data is random. For this reason, one might think that the model is committing a fallacy, not unlike the gambler’s fallacy or the hot hand fallacy: independent events are perceived as dependent (de Laplace 1951; Gilovich et al. 2002).
However, as many researchers have pointed out, committing this fallacy is in fact adaptive: randomness is rare in the real world (e.g. Pinker 1997: 346; Ayton and Fischer 2004) and the reward for noticing a real sequential dependency is usually greater than the punishment for postulating a spurious one. Viewed in this light, the imposition of sequential structure on the world by grasping at the straws of spurious dependencies is a sensible learning strategy.

6 Conclusion

Previous work in categorization and distributional learning has assumed that the center of a category corresponds to some measure of the central tendency of the experienced frequency distribution. By investigating the learning of skewed distributions, we have demonstrated that the center of an experienced phonetic distribution does not correspond to any such measure because the distribution is not represented veridically because experienced exemplars do not contribute equally to the category representation. Instead, an encounter with a rare member of the category influences the category more than an encounter with a frequent member of the same category does. Here, we have shown that this result is predicted by predictive error-driven learning. Under this view, the learner is not passively experiencing the category while building up a mental histogram. Rather, s/he is actively anticipating upcoming perceptual stimuli. The resulting category representation is not a faithful reproduction of environmental frequencies but rather is tailored to rapidly learning to predict the future.

Funding

Alexander von Humboldt-Stiftung (Grant/Award Number: ‘1141527’).

Notes

1 While there is some debate on whether the relationship between practice and performance is precisely logarithmic, there is consensus that it is a decelerating function, with the sharpness of deceleration the remaining issue (Keuleers et al. 2015).
2 Pitch excursion magnitude was chosen as the dimension of interest because it does not play a linguistic role in the participants’ native language (English). Therefore, participants should not have strong prior beliefs about locations of category boundaries along this dimension and should be especially susceptible to distributional learning (see also Escudero and Boersma 2004).
3 The data and analysis code are available at https://osf.io/bf62k/.
4 The shift is also confirmed by a non-parametric GAMM analysis, as well as by a group difference between the participant means in panel (c); (\(\hat{\beta} = 1.02, (64) = 4.17, p < 0.001, \text{adjusted } r^2 = 0.20\)).
5 Model code is available at https://osf.io/bf62k/.
6 This sense of ‘error-driven’ includes many learning models proposed in probabilistic phonology, both connectionist and probabilistic/Bayesian (e.g. Boersma and Hayes 2001; Kleinschmidt and Jaeger 2015; see also the discussion of surprisal in Hume and Mailhot 2013).
7 Only the points for trained stimuli shared between the two conditions (5ST to 17ST) are plotted in the right panel because the model does not extrapolate beyond the training data due to the nominal coding of pitch excursion. This limitation can perhaps be overcome by introducing cues that represent broader intervals on the pitch excursion continuum alongside the smaller intervals they contain, implementing confusability in the mapping between the input and the model’s intake and representing the continuum using cochlear-like acoustic cue representations recently developed in Arnold et al. (2017). Such manipulations are beyond the scope of the present paper.
8 It is also possible to set \( \Lambda \) so high (above \( \sim 0.33 \)) that the weights grow to plus and minus infinity instead of being constrained in the interval between \(-1\) and \(1\). For example, rate = 0.36 causes the weights to skyrocket into the \(10^{4\text{–}10}\) range. The model still predicts that the mean will shift towards the long tail of the input distribution but entirely mangles the shape of that distribution. We consider such learning rates to be outside of the reasonable parameter space for any application of the Rescorla-Wagner model, as overly rapid learning impairs the basic functioning of its learning algorithm, as well as other connectionist learning algorithms (see White 1989; McClelland et al. 1995: 436–438).
9 These data are also captured by Kleinschmidt and Jaeger’s (2015) probabilistic Bayesian model, which is also error-driven. It remains to be seen whether the relationship between surprisal and learning rate on a trial is always normative, as expected by Bayesian models, or dependent on overall learning rate, as expected by ours.

References


