Constraining Bridges Between Levels of Analysis: A Computational Justification for Locally Bayesian Learning

Adam N. Sanborn[a], Ricardo Silva[b]

[a]*Department of Psychology, University of Warwick*
[b]*Department of Statistical Science, University College London*

Abstract

Different levels of analysis provide different insights into behavior: computational-level analyses determine the problem an organism must solve and algorithmic-level analyses determine the mechanisms that drive behavior. However, many attempts to model behavior are pitched at a single level of analysis. Research into human and animal learning provides a prime example, with some researchers using computational-level models to understand the sensitivity organisms display to environmental statistics but other researchers using algorithmic-level models to understand organisms' trial order effects, including effects of primacy and recency. Recently, attempts have been made to bridge these two levels of analysis. Locally Bayesian Learning (LBL) creates a bridge by taking a view inspired by evolutionary psychology: Our minds are composed of modules that are each individually Bayesian but communicate with restricted messages. A different inspiration comes from computer science and statistics: Our brains are implementing the algorithms developed for approximating complex probability distributions. We show that these different inspirations for how to bridge levels of analysis are not necessarily in conflict by developing a computational justification for LBL. We demonstrate that a scheme that maximizes computational fidelity while using a restricted factorized representation produces the trial order effects that motivated the development of LBL. This scheme uses the same modular motivation as LBL, passing messages about the attended cues between modules, but does not use the rapid shifts of attention considered key for the LBL approximation. This work illustrates a new way of tying together psychological and computational constraints.

Keywords: rational approximations; locally Bayesian learning; trial order effects

Our goal when we model behavior depends on the level of analysis. If we analyze behavior at Marr (1982)'s computational level, then we aim to determine the problem that people are attempting to solve. Or, as more often found in psychology, we might be interested in the mechanism that drives behavior, placing us at Marr (1982)'s algorithmic level. In human and animal learning, both computational-level (Courville et al., 2005; Danks et al., 2003; Dayan et al., 2000) and algorithmic-level models (Rescorla & Wagner, 1972; Mackintosh, 1975; Pearce & Hall, 1980) have been developed. Models developed at different levels of analysis have different strengths and this can be seen in how these models of human and animal learning are applied: computational-level approaches are used to explain how organisms are sensitive to complex statistics of the environment (De Houwer & Beckers, 2002; Mitchell et al., 2005; Shanks & Darby, 1998) and algorithmic-level models are used to explain how organisms are sensitive to the presentation order of trials (Chapman, 1991; Hershberger, 1986; Medin & Edelson, 1988).

The computational and algorithmic levels provide different perspectives on model development, but an explanation is more complete if it works at both levels. Computational-level models that ignore the process can struggle with making fine-grained predictions (Sakamoto et al., 2008) and algorithmic-level models that ignore the computational level risk making incorrect or no predictions for task variants (Griffiths & Tenenbaum, 2009; Sanborn et al., 2013). A classic way to combine computational- and algorithmic-level insights is to begin with an algorithmic-level model developed to fit human behavior and then investigate its computational-level properties (Ashby & Alfonso-Reese, 1995; Gigerenzer & Todd, 1999). This is not the only possible direction, and recently researchers have begun at the computational level of analysis and then worked toward understanding the algorithm (Griffiths et al., 2012; Sanborn et al., 2010; Shi et al., 2010). Identifying the algorithm to associate with a computational-level model adds both psychological plausibility and explanatory power – computational-level models often are intractable, so the algorithm can provide a computationally tractable approximation while also explaining behavior that differs from predictions of the computational-level model as the result of the approximation. A major open question is how to select an approximation algorithm from the vast set of all algorithms, and again here human and animal learning provides examples of how this can be done.

Locally Bayesian Learning (LBL; Kruschke, 2006b) is one recent approach to bridging the computational and algorithmic levels in human and animal learning. LBL uses an approximation to a computational-level model to both improve computational tractability as well as better fit human trial order behavior. A driving motivation of LBL is a view inspired by evolutionary psychology: there are modular processes in the mind that have co-evolved. LBL assumes that each of these modules performs correct probabilistic updating, but each model must make due with only the information from messages it receives from other modules. By restricting the messages passed between modules, the predictions of LBL differ from that of the computational-level model it is based upon, Globally Bayesian Learning (GBL). LBL, unlike GBL, is able to successfully predict several effects of trial order on behavior, such as highlighting and the difference between forward and backward blocking. These effects are challenging because there are aspects for which earlier trials have greater influence, known as primacy effects, and aspects for which later trials have greater influence, known as recency effects, but computational-level models of behavior generally weight all trials equally.

Daw et al. (2008) motivate a bridge between the computational and algorithmic levels in a different way. Like with LBL, the approximation to the computational-level model is chosen because it reduces computational complexity while providing a better fit to human trial order behavior. However, computational instead of psychological considerations are used to select the approximation: A sequential updating algorithm is chosen from those that have been used in computer science and statistics to approximate complex probability distributions. The computational-level model is the Kalman filter (Kalman, 1960), which is a generalization of standard associative learning models (Dayan et al., 2000; Dayan & Kakade, 2001; Kruschke, 2008; Sutton, 1992), and it is approximated using Assumed Density Filtering (ADF; Boyen & Koller, 1998), an algorithm for sequential updating of a probability distribution. ADF approximates the full joint posterior distribution, which can contain dependencies between variables, with a factorized distribution that assumes the variables are independent. By using this and other approximations, the Kalman filter model is able to produce the same trial order effects that LBL does[1].

[1] Kalman filters were also used in a later version of the LBL by arranging two Kalman filters in a hierarchy and passing restricted messages between them (Kruschke, 2006a; Kruschke & Denton,

Both of the above motivations for bridging the computational and algorithmic levels have been criticized, each for not providing enough constraints. The restricted messages used by LBL have been criticized for having no specific computational justification (Daw et al., 2008), and thus leaving a great deal of freedom in selecting the content of messages and how they are passed between modules. In contrast, Kruschke (2010) argued that choosing an approximation from computer science and statistics is not very constraining, as there are a large number of plausible approximations from computer science and statistics that can be used.

Here we take the view that these motivations are not necessarily in conflict and that both psychological and computational motivations can be used to guide development of bridges between levels of analysis. We first describe LBL and review the trial order effects that are difficult for Bayesian models to produce. We then note that LBL constrains computation by assuming that a factorized posterior distribution is used to approximate the full posterior distribution on each trial. Using only this computational constraint and a standard measure of distance between probability distributions, we identify the message passing scheme that best approximates the full posterior distribution. This approximation is a form of ADF, the same approximation used to produce some trial order effects in the Kalman filter model. We show that the accumulation of approximation errors from a sequentially factorized representation alone produces these trial order effects, and that the rapid switching of the attended cues in the LBL messages is not necessary. We next give an example of where the predictions of LBL and the sequentially factorized representation differ. Finally, we discuss the implications for

2010).

attention, compare our approach to other approximations to rational models applied to human cognition, and discuss the prospects for integrating computational and psychological motivations.

1. Computational-Level Models of Learning

In human and animal learning studies, the problem that the organism faces is how to use a set of input cues x (e.g., lights or tones presented to an animal) to predict the outcome t (e.g., the food an animal receives). The statistical approach to this prediction problem is to view the relationship between the input cues and outcome as a probability distribution, $p(t,x)$. A full statistical treatment explains the joint probability of outcomes and input cues on a single trial $p(t,x)$, but we take as a starting point models of the conditional distribution $p(t|x)$, which is all that is needed for prediction of the outcome if the input cues are observed.

Computational-level analyses require both a set of possible hypotheses and a probability distribution over these hypotheses that describes the initial beliefs, called the prior distribution. Here the hypotheses are the possible mappings between the input cues and outcomes. There are many possible mappings, but a common choice is to start with outcomes that result from weighted sums of the input cues, as weighted sums are the basis of the classic Rescorla-Wagner (RW; Rescorla & Wagner, 1972) model. A prior distribution is then put over the possible weights, which completes the specification of the computational-level model.

One approach in this vein is the Kalman filter model (Dayan et al., 2000; Dayan & Kakade, 2001; Kalman, 1960; Kruschke, 2008; Sutton, 1992). The Kalman filter takes a weighted sum of input cues and maps it onto a Gaussian probability of a outcome. It also often assumes that the weights change over time, giving a built-in recency effect to the computational-level model because earlier data is less relevant than newer data. However, the Kalman filter model does not have a mechanism to produce primacy effects.

A second approach is GBL (Kruschke, 2006b), which begins with a model that has two levels of weighted sums. A schematic of the model is shown in Figure 1A. Unlike the Kalman filter or RW, GBL includes an early component that determines which input cues to attend to when computing the outcome prediction. The predicted outcome strength t is a sigmoid function of the weighted sum of the k attended cues y,

$$t = \mathrm{sig}\left(W_o y\right) \tag{1}$$

where $W_o y$ is the dot product (element-wise multiplication and then sum) of the $k \times 1$ vector representing the output weights, W_o, and the $k \times 1$ vector representing the attended cues, y. The weights were allowed to take discrete values for the sake of simplicity by (Kruschke, 2006b). Each output weight was allowed to take the values of -5, 0, or 5. The weights were combined with the activity of the attention cues and put through Equation 1, and raised to the power of 1.

Likewise, each attended cue's activation is a sigmoid function of a weighted sum of the input values,

$$y = \mathrm{sig}\left(W_h x\right) \tag{2}$$

where $W_h x$ is the matrix product between a $k \times k$ hidden weight matrix, W_h, and a $k \times 1$ vector of inputs, x. There were an equal number of input and attended cues, which were linked with one of two types of weights. Excitatory weights could take the value of 4 or 6 and inhibitory weights could take the value of 0 or -4.

Each input cue was linked to one attended cue with excitatory weights (in a one-to-one mapping) and linked to every other attended cue with inhibitory weights. As a result of this mapping, each attended cue could be identified with an input cue. To compute the activation of an attended cue, the weights of the present input were summed and put through a sigmoid (i.e., logistic) function as in Equation 2, and raised to the power of 6. This last operation was chosen by Kruschke (2006b) so that activation ranged from nearly zero to nearly one.

Given this specification, the learning done by the model is fixed. Bayes' rule is used to update the probability distributions over the hidden weights and hidden attentional cues based on the trials that have been experienced

$$p(W_h, W_o, y \mid x, t) \propto p(t \mid W_o, y) p(y \mid W_h, x) p(W_o, W_h). \tag{3}$$

The prior distribution on the output and hidden weights was independent, $p(W_o, W_h) = p(W_o)p(W_h)$. The prior $p(W_h)$ was a discrete uniform over all possible combinations of hidden weights. The prior $p(W_o)$ over sets of output weights was set to favor sets of weights that had more values of zero: a product of pseudo-Gaussian distributions[2] (ϕ) with mean zero and standard deviation five for each weight w_i in the set: $\prod_i \phi(w_i)$. Note that this is the prior for the first trial. Throughout this paper we consider the predictions of the model relative to a single trial, relegating information from previous trials to the prior to simplify the notation.The prior distribution over weights $p(W_h, W_o)$ is set to the posterior distribution from the previous trial.

[2] The discrete weights were assigned probability proportional to their density under a Gaussian distribution.

GBL learns in a probabilistically correct fashion from experience, but it is a poor fit to the experimental data: unlike the Kalman filter it is necessarily a stationary model and produces neither a primacy nor a recency effect. GBL can also quickly become intractable as the number of input cues grow, as it represents the probability of every combination of possible values of hidden weights and output weights. For k input cues and m outcomes, there are 2^{k^2} possibilities for W_h and 3^{km} possibilities for W_o, yielding $2^{k^2} * 3^{km}$ possibilities for the combinations of weights. As an illustration of how quickly the number of possibilities grows with the number of input cues, one input cue and one outcome produce six possible combinations of weights, but three input cues and one outcome produces over thirteen thousand weight combinations.

2. Locally Bayesian Learning

LBL is an approximation to GBL that both decreases the required computation and produces human-like trial order effects. LBL splits the network graph of GBL into two modules, as shown in Figure 1B. Each module is meant to represent a psychological process: the lower module takes input cues and maps it to attended cues, and the upper module maps the attended cues to the outcome. Each module is self-contained and only represents a probability distribution over its own set of weights. Splitting GBL into modules results in a much smaller representational complexity, instead of $2^{k^2} * 3^{km}$ possibilities, there are now $2^{k^2} + 3^{km}$ possibilities that need to be separately represented. In our illustration, three input cues and one outcome produce 539 combinations of weight values in LBL, less than 4% of the

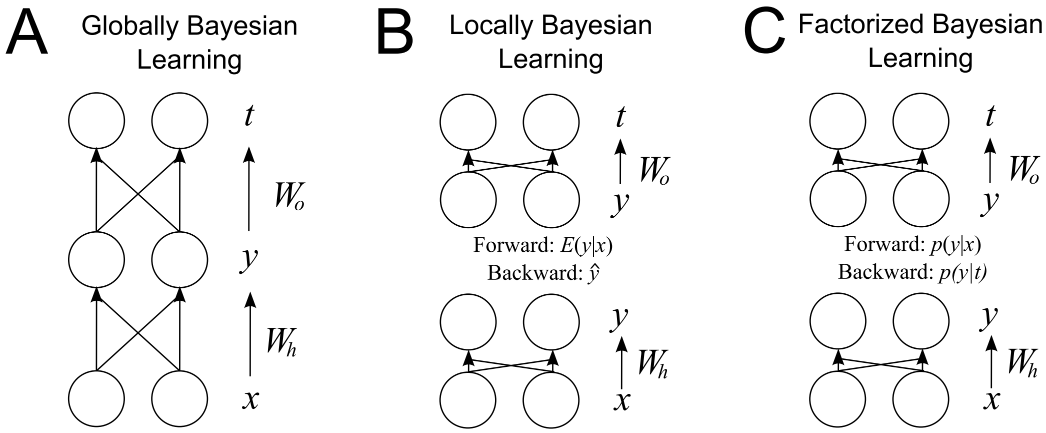

Figure 1: Diagrams of Globally Bayesian Learning (GBL), Locally Bayesian Learning (LBL), and Factorized Bayesian Learning (FBL). Input cues x are weighted by hidden weights W_h and transformed to produce attended cues y. Attended cues y are weighted by output weights W_o and transformed to produce outcomes t. Multiple outcome nodes are possible as shown here, though only one was required for the tasks we model. The hidden weights W_h, attended cues y, and output weights W_o are not observed and are instead inferred. In GBL, all of the hidden variables are inferred together. LBL splits GBL into two modules with copies of the attended cues y in each module. Messages are passed back and forth between the copies of the attended cues y, the expected value $E(y|x)$ is passed upward and the single $\hat{y}$ that maximizes the probability of the outcome is passed backward. FBL uses the same modules as LBL, but the messages passed between modules are distributions over the attended cues rather than a single set of attended cues.

combinations used in GBL for the same number of input cues and outcomes[3].

Each LBL module uses Bayes' rule to update its own representation, but each is prevented from observing the entire state of the environment or the probability distribution represented in the other module. Instead, a module receives some in-

[3] Continuous representations could also be used to reduce the complexity of the hypothesis space, such as in Kruschke & Denton (2010).

formation indirectly in the form of messages passed from the other module. Messages moving forward from the lower module to the upper contain the expected value of the attended cues $E(y|x)$. The upper module only observes the expected values of the attended cues and is blind to the input cues. Once the outcome t has been given, the output weights W_o are updated to be $p(W_o|E(y|x),t)$, instead of the $p(W_o|x,t)$ as they would be in GBL. The expected values of the attended cues given the input are used in the place of the probability distribution over the attended cues given the input.

Messages passed downward from the upper module to the lower module are of a different type. First the output weights are updated, then the value of y that maximizes the probability of the outcome t is passed downwards to the lower module,

$$\hat{y} = \arg\max_{y^*} \sum_{W_o} p(t|W_o, y^*) p(W_o|E(y|x), t) \qquad (4)$$

The lower module only observes $\hat{y}$ and the input cues, so the hidden weight prior $p(W_h)$ is updated to be $p(W_h|x,\hat{y})$. The restricted messages passed in LBL and the trial-by-trial updating of the representation result in its predictions depending on the order of the training trials, which we discuss in the next section.

3. Trial Order Effects

Two trial order effects in particular have so far proved difficult for computational-level modeling: the highlighting effect and the difference in strength between forward and backward blocking.

3.1. Highlighting

Because people are sensitive to the statistics of the environment, we expect that the frequencies of different outcomes would play a role in people's judgments: a higher frequency outcome should produce a stronger relationship than a lower frequency outcome. However, people can display unusual responses to the relative frequency of trials in an experiment, such as the inverse base-rate effect (Gluck & Bower, 1988; Medin & Edelson, 1988). The inverse-base rate effect occurs following two types of training trials. In the first, two input cues, *I* and *Pe*, are associated with an outcome E. We will write this as $I.Pe \to E$. The second set of trials pairs one of the old input cues, *I*, with a new input cue *Pl*, and a new outcome *L*. The labels associated with the input cues and outcomes indicate their roles (which participants must learn from experience): input cue *I* is an imperfect predictor, input cue *Pe* is a perfect predictor of early outcome E, and input cue *Pl* is a perfect predictor of late outcome *L*. When given a test trial with input cue *I* or with conflicting input cues *Pe* and *Pl*, it is reasonable to expect that the response chosen would depend on the relative frequencies of the two types of trials. If there were more $I.Pe \to E$ trials than $I.Pl \to L$ trials, then participants should respond E given input cue *I* or the conflicting input cues $Pe.Pl$. However, Medin & Edelson (1988) found that while participants chose the higher frequency outcome if given input cue *I*, they chose the lower frequency outcome if given the conflicting input cues $Pe.Pl$.

Later work showed that the inverse base-rate effect arises even if the relative frequency of training trial types is equated, but more $I.Pe \to E$ trials are presented early in training (Collins et al., 2011; Kruschke, 1996, 2009; Kruschke et al., 2005; Medin & Bettger, 1991; Sewell & Lewandowsky, in press). As a result, the inverse base rate effect has been renamed highlighting, the name following from an attentional explanation of the effect. Participants first learn that input cues *I* and *Pe* equally predict outcome *E*, because they are both equally predictive of the outcome in the $I.Pe \rightarrow E$ trials. However, the later $I.Pl \rightarrow L$ trials demonstrate the ambiguity of input cue *I* and highlight the relationship between input cue *Pl* and outcome *L*, so participants heavily weight this latter relationship. During test, there is both a primacy effect and a recency effect. The primacy effect is that *I* has a stronger relationship with outcome *E*. The recency effect is that if input cues *Pe* and *Pl* compete against each other, outcome *L* is chosen because *Pl* has a stronger relationship to *L* than *Pe* has to *E* (Kruschke, 1996).

Kruschke (2006a,b) demonstrated that highlighting was an extremely challenging effect for Bayesian models of learning because of the equal number of training trials of each type. The predictions of GBL for two types of highlighting designs are shown in Figure 2. The first design was used in Kruschke (2006b) to demonstrate the models: seven trials of $I.Pe \rightarrow E$ followed by seven trials of $I.Pl \rightarrow L$. The second design follows more closely that used in Kruschke (2009, design given in the Appendix), in which the human data showed a strong highlighting effect.

Unlike GBL, the restricted nature of the messages in LBL causes it to predict a robust highlighting effect that matches human data, as shown in Figure 2. The prediction of highlighting was explained by attention to cues to that rapidly switched between trial types, like in the description of highlighting above. The message passed backward from the upper module to the lower module consisted of attended cues I' and Pe' in the early trials[4], so I' is activated on these trials and

[4] Each hidden attended cue can be identified with a specific input cue because there is a one-

is thus associated with E. However on the second block of $I.Pl \to L$ trials, as I already strongly activates E, the attended cue that is maximally consistent with the output weights is Pl' alone so I' is not activated. As a result, I more strongly activates E and Pl more strongly activates L than Pe activates E, producing the highlighting effect (Kruschke, 2006b).

3.2. Forward and Backward Blocking

The experimental effect of blocking demonstrates how input cues compete with each other during learning (Kamin, 1968). As a comparison, control trials consist of two input cues and a outcome, $A.B \to R$, and participants believe that B predicts R with some moderate strength. Forward blocking occurs if this set of training trials is preceded by a set of training trials in which $A \to R$. The initial learning of $A \to R$ blocks the establishment of a relationship of B with R in the $A.B \to R$ trials, as A by itself was sufficient to predict the outcome. After the two blocks of learning, participants believe that B predicts R only weakly.

Forward blocking is a straightforward prediction of RW, but a slight change to the design complicates associative explanations. In backward blocking, the order of the blocks is reversed so that the $A.B \to R$ trials occur before the $A \to R$ trials. Here the prediction of R from B is also reduced, though this effect is not as larger or as robust as forward blocking (Beckers et al., 2005; Chapman, 1991; Kruschke & Blair, 2000; Lovibond et al., 2003; Melchers et al., 2006; Shanks, 1985; Vandorpe et al., 2007). Essentially, participants retrospectively re-evaluate the strength of the relationship between B and R, reducing it because of the later

to-one mapping of positive weights between input cues to attended cues. All other weights were between input cues and attended cues are non-positive.

$A \to R$ trials. Backward blocking is not predicted by RW and so has been taken as evidence for statistical accounts of learning, though modified associative accounts are able to predict it (Van Hamme & Wasserman, 1994).

Backward blocking can be explained by Bayesian models of learning (Gopnik et al., 2004; Sobel et al., 2004; Tenenbaum & Griffiths, 2003), but a difference in strength between forward and backward blocking presents difficulties for many Bayesian models because the two designs differ only in the order of presentation of the training trials (but see Daw et al., 2008; Dayan & Kakade, 2001). Many experiments have shown a trace of this effect (Chapman, 1991; Kruschke & Blair, 2000), and it was shown to be statistically reliable in (Vandorpe et al., 2007). The difference in the size of the effects in this study was found to be between 10% and 20% of the range of the scale.

The predictions of GBL and LBL for forward and backward blocking are shown in Figure 3, using the same parameters as Kruschke (2006b). The control condition consisted of seven trials of $A.B \to R$, the forward blocking condition consisted of seven trials of $A \to R$ followed by seven control trials, and the backward blocking condition consisted of seven control trials followed by seven trials of $A \to R$. GBL shows the same decrement for both forward and backward blocking relative to control trials. LBL does predict a stronger influence of forward than backward blocking, which again was attributed to passing the maximally consistent value from the upper module to the lower module (Kruschke, 2006b). In forward blocking the initial block of $A \to R$ trials provides no information about what the outcome should be to input cue B, and thus there remains a good possibility that $B' \to \neg R$, where $\neg R$ is no outcome. Given the uncertainty about the outcome to B, when the later $A.B \to R$ trials appear it is best to attend to A and ignore *B*. This results in a weak relationship between *B* and *R*, as *B* is ignored during trials in which the relationship could be strengthened. In contrast, in backward blocking the maximally consistent message passed from the upper module to the lower module in the first block of backward blocking is to attend to both input cues, strengthening the relationship during this block of training trials. As a result, for LBL, *B* predicts *R* more strongly in backward than forward blocking.

4. Message Passing and Factorized Representations

The match between LBL and human trial order effects is due to a message passing scheme that was chosen on an ad-hoc basis. There are many possible schemes for passing messages between modules, varying in aspects such as which content is passed, in which sequence and at which loss of information. Despite the multiplicity of possible mechanisms, we argue that there are some general principles that can strongly constrain the possible algorithmic constructions for a computational model. In this section, we initially discuss how the message passing scheme of LBL approximates GBL. We then set the stage to introduce an alternative based on a more fundamental set of algorithmic principles, with the goal of largely retaining the predictive power of LBL without seemingly ad-hoc combinations of approximations.

To understand the design choices behind LBL, let us first summarize how GBL works. For GBL, the posterior distribution over the weights, $p(W_o, W_h|x,t)$, does not factorize into independent contributions from each of the weights, as in $p(W_o|x,t)p(W_h|x,t)$. Instead,

$$p(W_o, W_h|x,t) = \sum_y p(W_o|y,t)p(W_h|y,x)p(y|x,t). \tag{5}$$

GBL has a distribution over the possible attended cues, y, and this range of possibilities means that the posterior distribution does not factorize. A non-factorized distribution requires more memory to represent, and complicates updates when new data points are observed. However, we can see in Equation 5 what would happen if y were fixed at a single value: the summation would disappear and the posterior distribution would factorize. Exploiting this fact, the representation learned by LBL has the following structure:

1. Consider collapsing our uncertainty over $\mathbf{y}$ using an estimate $\tilde{y}$, which is assumed to be known with certainty (i.e., $p(\tilde{y}|x,t) = 1$)
2. To further simplify computation, do not construct a representation with the structure $p(W_o, W_h|x,t) \approx p(W_o|\tilde{y},t)p(W_h|\tilde{y},x)$: instead, use *two* different estimates where $p(W_o, W_h|x,t) \approx p(W_o|\tilde{y}_1,t)p(W_h|\tilde{y}_2,x)$. This means computation can be carried separately within two different modules, one for each factor
3. Under this formulation, use one estimate $\tilde{y}_i$ generated within one module to compute the other estimate $\tilde{y}_j$

Within the choices provides by this framework, LBL can be thought as having a single hypothesis, though different in the upwards and downwards messages, passed between modules. LBL's posterior distribution of $p(W_o, W_h|x,t)$ is a factorized distribution and can be written as the product of the individual weight distributions $p(W_o|E(y|x),t)p(W_h|\hat{y},x)$. LBL starts with a factorized prior $p(W_o, W_h) = p(W_o)p(W_h)$, and generates a factorized posterior $p(W_o, W_h \mid t,x) \equiv p^{new}(W_o)p^{new}(W_h) \equiv p(W_o|E(y|x),t)p(W_h|\hat{y},x)$. This probability distribution is then treated as a new prior for the next data point and the process is iterated.

Approximations that sequentially factorize the posterior distribution after each data point have been explored in computer science and statistics. This class of approximations is known as Assumed Density Filtering (ADF; Boyen & Koller, 1998). In ADF, the posterior distribution over parameters is approximated with a simpler distribution after each new data point is observed. This approximate posterior is used as the prior when processing the next point.

While LBL's message passing scheme falls within the class of ADF algorithms, there is still the question of whether LBL is a good approximation to the full posterior distribution. We can test LBL's message passing scheme by examining it within Minka (2005)'s unified framework for generating approximations, which encompasses and generalizes several techniques from machine learning, statistics, statistical physics, and information theory. One key aspect of this framework is that the choice of approximation is based on picking the approximation that is "closest" to $p(W_o, W_h \mid t, x)$ according to some definition of similarity between probability functions.

Although this similarity-maximization (or, analogously, divergence-minimization) principle might sound too broad, LBL does not seem to obey it. Namely, we have been unable to find any divergence measure $D(p,q)$ where, for $p = p(W_o, W_h \mid t, x)$ we have $q = p^{new}(W_o)p^{new}(W_h)$ as the factorized distribution that minimizes $D(p,q)$.

LBL's message passing scheme may not be justified by a divergence measure, but we can identify a message passing scheme that is justified. This removes several of the degrees of freedom of LBL's framework: we do not have a choice of different estimators of y, and how they are computed as a function of the other. The framework is predicated on the choice of a divergence measure. We adopt the standard procedure of ADF: considering all factorized distributions $q(W_o, W_h) \equiv$

$q_o(W_o)q_h(W_h)$, find the one that is closest to the true posterior $p(W_o,W_h \mid x,t) \equiv p_{x,t}(W_o,W_h)$. We term this approach Factorized Bayesian Learning (FBL) and a schematic of this model is shown in Figure 1C.

Our choice of a message passing scheme depends on our measure of divergence. We propose that the choice of q should be the one that minimizes the *Kullback-Leibler* (KL) divergence. KL divergence is a popular criterion for choosing approximations, since $KL(p \mid\mid q) = 0$ if and only if $p = q$, and is positive otherwise (Minka, 2001). It has a long history in information theory (Cover & Thomas, 1991), and has an interpretation based on coding: if a string/sample is generated from distribution p, but encoded using a scheme based on q, the KL divergence is how many extra bits (or nats in our case, because we use base e) are needed to encode the message relative to the optimal code based on p. KL divergence can be written as

$$KL(p \mid\mid q) = \int\int p(W_o,W_h \mid x,t) \ln \frac{p(W_o,W_h \mid x,t)}{q_o(W_o)q_h(W_h)} dW_o dW_h \tag{6}$$

where the target approximation $q(W_o,W_h)$ takes the shape $q_o(W_o)q_h(W_h)$. The distribution over hidden attended cues y is implicit, since the problem of choosing $q_o(W_o)q_h(W_h)$ to minimize Equation 6 is equivalent to choosing the one that minimizes

$$-\int\int\int p(W_o,W_h,y \mid x,t) \ln \left[q_o(W_o)q_h(W_h)\right] dW_o dW_h dy \tag{7}$$

Minimizing Equation 6 with respect to $q_o(\cdot)$ and $q_h(\cdot)$ results in $q_o(W_o) = p(W_o \mid x,t)$ and $q_h(W_h) = p(W_h \mid x,t)$ regardless of the functional form of $p(\cdot \mid x,t)$ (Minka, 2001). This can be shown as follows. For simplicity, assume all random variables are discrete, as this will be the case in our case study. Then Equation 7 can be rewritten as

$$-\sum_{W_o} p(W_o \mid x,t) \ln q_o(W_o) - \sum_{W_h} p(W_h \mid x,t) \ln q_h(W_h) \tag{8}$$

We have to optimize this function with respect to the entries of $q_o(W_o)$ and $q_h(W_h)$ such that such entries are non-negative and $\sum_{W_o} q_o(W_o) = 1$, $\sum_{W_h} q_h(W_h) = 1$. Using Lagrange multipliers for this constrained optimization problem and ignoring for now the non-negativity constraints, this gives the following objective function:

$$-\sum_{W_o} p(W_o \mid x,t) \ln q_o(W_o) - \sum_{W_h} p(W_h \mid x,t) \ln q_h(W_h) + \lambda_o(\sum_{W_o} q_o(W_o) - 1) + \lambda_h(\sum_{W_h} q_h(W_h) - 1) \tag{9}$$

Taking the derivative of Equation 9 with respect to an arbitrary entry $q_o(W_o)$, we obtain

$$-p(W_o \mid x,t)/q_o(W_o) + \lambda_o = 0 \tag{10}$$

which implies $q_o(W_o) \propto p(W_o \mid x,t)$ for all values W_o. Because $\sum_{W_o} q_o(W_o) = 1$, it follows that $q_o(W_o) = p(W_o \mid x,t)$. The reasoning is analogous when deriving $q_h(W_h) = p(W_h \mid x,t)$.

The role of message-passing and prior factorization have algorithmic implications due to the calculation of the marginals. For the output weights,

$$\begin{aligned} p(W_o|x,t) &\propto \textstyle\sum_{W_h} \sum_y p(t \mid W_o, y) p(y \mid W_h, x) p(W_o) p(W_h) \\ &= p(W_o) \textstyle\sum_y p(t \mid W_o, y) \sum_{W_h} p(y \mid W_h, x) p(W_h) \\ &\equiv p(W_o) \textstyle\sum_y p(t \mid W_o, y) m_x(y) \end{aligned}$$

where $m_x(y) \equiv \sum_{W_h} p(y \mid W_h, x) p(W_h)$ is the message passed from the lower module to the upper module that encapsulates all the information content provided by

x for a given value of y. It is possible to decouple this module from the output module only because the prior over W_o and W_h factorizes as $p(W_o)p(W_h)$.

Analogously,

$$\begin{aligned} p(W_h|x,t) &\propto \textstyle\sum_{W_o}\sum_y p(t \mid W_o,y)p(y \mid W_h,x)p(W_o)p(W_h) \\ &= p(W_h)\textstyle\sum_y p(y \mid W_h,x)\sum_{W_o} p(t \mid W_o,y)p(W_o) \\ &\equiv p(W_h)\textstyle\sum_y p(y \mid W_h,x)m_t(y) \end{aligned}$$

where $m_t(y) \equiv \sum_{W_o} p(t \mid W_o,y)p(W_o)$ is the message passed from the upper module to the lower module that encapsulates all the information content provided by t for a given value of y.

The approximated posterior given by the optimal $q_o(W_o)q_h(W_h)$ is the new prior when processing the next data point. Hence the prior used at the beginning of each trial is always factorized. The approximation used in LBL produces a non-standard projection of the true posterior into the space of factorized distributions, so FBL will always be as good or better than LBL in the KL divergence sense, all without the need of choosing a way of collapsing y. An example of a full posterior distribution and the factorized approximation made by FBL is shown in Figure 4.

Given the generality of the KL divergence metric, the main degree of freedom in the algorithmic procedure of FBL is the choice of approximation by a factorized distribution. This is shared with LBL and is motivated by requiring, in general, exponentially fewer bits of information (as a function of the number of parameters) to be represented than a full joint distribution, and by allowing a message passing formulation when calculating marginals. And, as we will show next, this factorization property will imply artifacts of reasoning that match human behavior.

5. Factorized Representations for Trial Order Effects

Surprisingly, Figure 2 shows that FBL produces the highlighting effect for the design in Kruschke (2006b) and that the size of the effect matches human data if the model is trained with the same number of stimuli that participants were. Figure 3 shows that FBL also produces the trial order effect for blocking. The FBL highlighting effect and the FBL blocking effect can be better matched to the size in the human data by adjusting the parameters of the model[5], but we used the original parameters to demonstrate that the FBL produces the same qualitative effects as LBL with the same parameters. Instead of an explanation that is due to passing the maximally consistent message backwards, this effect is due to the more basic separation of GBL into two modules and the sequential approximation of trials that then results.

Effects of approximation have a long history in comparisons of the most general artificial algorithmic system, the digital computer, against abstract computational models such as the Turing machine. The field of numerical analysis, in particular, tackles the issue on how problems of mathematical analysis can be solved in practice, considering the accumulation of errors due to the sequential processing of numerical operations using a digital representation. One can, for instance, analyze the computational complexity of a procedure for matrix inversion (Cormen et al., 2009), but its numerical stability depends upon the control

[5] Changing the parameters to fit highlighting data must be done carefully. For some parameter settings, the GBL does predict a highlighting effect because the critical test items I and $Pe.Pl$ consist of different numbers of cues, and inhibition only occurs if more than one cue is presented. For example, if the prior on inhibition is strong, the single cue I has no other cue to inhibit it and so favors E, but strong inhibition between cues makes $Pe.Pl$ favor L.

of rounding errors that accumulates as the sequence of steps in the algorithm is followed. ADF, and in particular FBL, uses an approximation as input to the next approximation. As the literature of numerical analysis shows us, a combination of biases and sequential processing might lead to results that do not match what the computational model entails.

ADF gives the best approximation *with respect to a given prior*. However, since in a sequential update scheme the "prior" represents compiled evidence of previous observations, errors will propagate. Minka (2001) suggests, for instance, that ADF is particularly prone to bad approximations of marginals if the input sequence of data points differs considerably from what it should be obtained by a randomized sequence. Hence, FBL dispenses with the necessity of a Kalman filter formulation, since an ordering effect is automatically accounted for by approximation errors.

An example that will prove useful for explaining both the highlighting and blocking predictions is shown in Figure 5. GBL and FBL both begin with the same prior distribution and are updated with either data that produce "complex" likelihoods or data that produce "simple" likelihoods. The weights are dependent in the complex likelihood, but independent in the simple likelihood. If FBL and GBL are updated with a simple likelihood, they produce the same posterior distributions, as can be seen in the first and second posterior distributions if the simple likelihood is presented first. However, when GBL and FBL are updated with a complex likelihood, their joint posterior distributions diverge.

Especially interesting is the case in which GBL and FBL are updated multiple times with a complex likelihood. Here not only do the joint distributions diverge, the sequential updating procedure also results in the marginal posterior distributions diverging. The final marginal distributions depend on whether FBL is updated with the two simple likelihoods first or updated with the two complex likelihoods first. The difference between the final marginal distributions (shown at the bottom of Figure 5) is small for such a small number of training trials, but the most likely value is smaller and the other values larger if the complex likelihoods are presented first compared to if the simple likelihoods are presented first. As we explain below, this is what drives both the highlighting and blocking effects. Unlike in the LBL, the trial order effects do not arise from the rapid nature of changes to which cues are attended to, but instead follow directly from computational considerations.

5.1. Predicting Highlighting

We present an example of how the posterior distribution changes in FBL in Figure 6. The final posterior distribution is a result of training with an early block of seven $I.Pe \rightarrow E$ trials followed by a block of seven $I.Pl \rightarrow L$ trials (model predictions are shown in Figure 2). The hypothesis space of FBL is summarized in Figure 6 to make the relevant patterns easier to see. The vertical dimension of each plot shows possible hypotheses about the hidden weights, grouping those hypotheses that do and do not result in the attended cues exceeding an arbitrary threshold of 0.5 in activation. The horizontal axis groups the probabilities of the output weights by showing the probability of the largest weights for one hypothesis or the other, excluding the probabilities of the indifference weight. The two rows separately summarize the hypotheses relevant to $I.Pe$ and $I.Pl$. This is not a partition of the hypotheses, some of the hypotheses about I are reused between the two rows of plots and other hypotheses about indifference in output weights do not contribute at all, but it provides a useful summary.

The likelihoods for the first and second set of training trials have the same structure as those we used as examples in Figure 5. $I.Pe \rightarrow E$ trials are presented first, so the first seven likelihoods for I and Pe in the first row of the plot are complex, where if only I is attended (activating its corresponding attended cue I') then both hypotheses in which $I' \rightarrow E$ are more likely than the hypothesis in which $I' \rightarrow L$. Likewise, if only Pe is attended, then both hypotheses in which $Pe' \rightarrow E$ are more likely than the hypothesis in which $Pe' \rightarrow L$. However, the $I.Pe \rightarrow E$ trials tell us nothing about Pl. So for the first likelihood of the $I.Pl$ hypotheses, we learn that $I' \rightarrow E$ is more likely than $I' \rightarrow L$, but it has no interactions with the attended cues, a likelihood which is simple. The second seven likelihoods are exactly the reverse. The $I.Pl \rightarrow L$ training trials gives us a likelihood that is complex between hidden and output weights for $I.Pl$ hypotheses, but is simple for $I.Pe$ hypotheses. This gives us the same orderings of simple and complex likelihoods as in Figure 5.

As a result, we find the same effect on the marginal distributions that we found in Figure 5. When the complex likelihood is first, then the first column is reduced and the third column is boosted compared to when the simple likelihood is first. We can see the highlighting prediction arise from this difference in the marginals[6]. The combined marginal in which $I' \rightarrow E$ are greater than the combined marginals in which $I' \rightarrow L$, giving the prediction for the irrelevant input cue. For $Pe.Pl$, the combined marginals in which $Pe' \rightarrow E$ are less than the combined marginals in which $Pl' \rightarrow L$, giving the prediction of $Pe.Pl \rightarrow L$.

[6] We ignore the hidden weights here because the test input cue I alone means that there is no cross-cue inhibition, so I is likely activated. Also, the test input cues $Pe.Pl$ are a novel combination, so inhibition should not have changed much from baseline.

5.2. Predicting Blocking

FBL predicts a difference in strength between forward and backward blocking for the same reason: the ordering of complex and simple likelihoods. Figure 7 summarizes the hypothesis space for FBL for blocking. Again, the vertical dimension of each plot shows possible hypotheses about the hidden weights, grouping those hypotheses that do and do not result in the attended cues exceeding an arbitrary threshold of 0.5 in activation. The horizontal axis groups the output weights by showing the probability of the largest weight and smallest weight for reward. The two rows separately summarize backward and forward blocking. Hypotheses about indifference in output weights are not included in this figure.

Like highlighting, the likelihoods for blocking are either complex or simple, though here the complex or simple likelihoods apply to the entire hypothesis space. In backward blocking the first set of trials are $A.B \rightarrow R$, so the combinations of hypotheses that lead to a greater prediction of R are given greater likelihood. This leads to a dependence between hidden and output weights because if only A' is activated, then $A' \rightarrow R$ is more likely than $B' \rightarrow R$, and the relative ordering of the probabilities reverses if B' is activated. In forward blocking the first likelihood is simple. We are only learning about A with $A \rightarrow R$ trials, so we do not learn anything about which cues should be attended.

The end result again is that FBL produces lower marginals for the first column and higher marginals for the third column for the complex likelihood first compared to the simple likelihood first. This produces the blocking trial order effect as well. The probability of $B' \rightarrow R$ is higher for backward blocking than forward blocking as a result, reproducing the experimental effect.

6. Differences Between Locally Bayesian Learning and Factorized Bayesian Learning

We have shown FBL is a more principled approximation than LBL and here we demonstrate how the more principled approximation can lead to different predictions. The example we use is a classic in both the artificial intelligence and the human and animal learning literatures: the exclusive-OR (XOR) problem for which the learner is trained to respond to cues singly but not in combination. XOR is a simple version of a nonlinearly separable problem that cannot be learned by a single layer linear network (Minsky & Papert, 1969), but has been shown to be learnable by both animals and humans (also known as negative patterning; Pavlov, 1927; Harris & Livesey, 2008; Harris et al., 2008; Rescorla, 1972, 1973).

A simple XOR design was used in which we trained the models on three types of trials: the single cues $T \rightarrow R$ and $U \rightarrow R$ the compound cue $T.U \rightarrow 0$. We presented these cues in that order for seven repetitions and show the predictions for GBL, LBL, and FBL in Figure 8. GBL and FBL were both able to learn that the single cues were better associated with R than the compound cue. However, LBL does not learn this distinction, predicting essentially the same outcome for both the single and compound cues. This inability arises from choosing to train the lower module to produce the attentional cues that maximize the probability of the outcome. Even for the single cues T and U, the outcome is maximized if both attentional cues T' and U' are activated. However, in order to predict a higher response to the single cues than the compound cue the two cues must be trained to inhibit one another when both are present. GBL and FBL can both learn this inhibition scheme because they do not use the max message passed downward in LBL.

Of course a common way to attempt to account for XOR problems is to introduce configural units (Minsky & Papert, 1969; Spence, 1952), and indeed Kruschke (2006b) proposed this solution for LBL. There is some evidence that configural units make the wrong sort of predictions for human and animal behavior (Harris & Livesey, 2008; Harris et al., 2008), but if we allow them then this demonstration serves to illustrate a difference between FBL and LBL that could be potentially tested in experiments with more complex XOR designs.

7. Discussion

We have shown how a computationally justified version of LBL can be used to produce human-like trial order effects, and additionally how the FBL potentially better matches human behavior in XOR tasks. Here we investigate the implications for rapid shifts of attention, relate the approximation used in FBL to other approximations hypothesized to be in use in the mind, discuss the hypothesis that modularity corresponds to factorization, and conclude.

7.1. Implications for Rapid Shifts of Attention

The success of LBL in producing the effect of highlighting and the difference between forward and backward blocking was attributed to rapid shifts of attention (Kruschke, 2006b), like those used in the error-driven connnectionist model EXIT (Kruschke, 2001a,b). These rapid shifts were identified with the maximization messages passed backward from the upper module to the lower module. Later work with the Kalman filter model demonstrated that rapid shifts of attention were not necessary to produce highlighting, because it could be produced with a single layer network instead, though the approximation used in the single layer network needed to be complex (as discussed below) to produce the blocking results as well (Daw et al., 2008).

The current results go beyond these to demonstrate that even in the two layer network of the LBL, in which the output is based on attended cues rather than the observed cues, rapid shifts of attention are not necessary to predict highlighting or the difference between forward and backward blocking. Instead, FBL predicts these results using only the factorization of the probability distributions over the layers of the weights of the network. The message passed backwards from the upper module to the lower module is not a maximization message, but is instead the actual marginal distribution of the attended cues given the outcome of the trial. This indicates that the separation between the modules imposed by the factorization is sufficient to produce these trial order effects, and that the particular kinds of messages associated with rapid shifts of attention are not necessary.

7.2. Kinds of Approximations

The computationally-justified message passing scheme we developed, FBL, uses the same class of approximations as Daw et al. (2008) used in their approximation to the Kalman filter model, but the explanations differ in their details. Both approaches are sequential updating algorithms that factorize the posterior distribution after each trial, but the explanations of why the trial order effects occur differ because of the different computational-level model structures. The Kalman filter model uses a single layer to map input cues to outcomes and conceptualizes attention as uncertainty about the weights. In contrast, FBL uses an explicit module for activating attended cues before a second module maps the attended cues to the outcome. The effect that factorizing the posterior distribution has in each model differs as well. For FBL, the sequentially factorized posterior produces both the highlighting effect and the difference in strength between forward and backward blocking. For the Kalman filter, factorizing the posterior distribution produces highlighting, but causes backward blocking to disappear. In order to produce both effects, interpolations were made between the sequentially factorized posterior distribution that produces highlighting and the full posterior distribution that produces backward blocking.

The effectiveness of LBL, FBL, and the Kalman filter approximations in trial order effects has wider implications for how we attempt to build bridges between computational- and algorithmic-level analyses. Other research has used sampling algorithms from computer science and statistics to bridge computational- and algorithmic-level analyses. This has been done in wide variety of areas, such as categorization (Sanborn et al., 2010; Shi et al., 2010), sentence parsing (Levy et al., 2009), prediction (Brown & Steyvers, 2009), perceptual bistability (Gershman et al., 2012), and even human and animal learning (Lu et al., 2008; Rojas, 2010) to explain trial order effects. Sampling algorithms tend to come with asymptotic guarantees: with enough samples any computation done with these algorithms will be indistinguishable from computation done with the full probability distribution. To allow for computational tractability and to produce deviations from the computational-level model, far fewer samples are used. While in some situations we can choose among sampling algorithms to best approximate the posterior distribution (e.g., Fearnhead, 1998) and the number of samples that best balances reward with opportunity cost can at times be computed (Vul et al., 2009), the quality of approximation given by a sampling algorithm is generally not made explicit.

In contrast, the approximations used in LBL, FBL, and the Kalman filter are all examples of sequential variational approximations. Instead of representing the distribution with a series of points chosen stochastically from the true distribution, variational approximations are deterministic and approximate a target distribution by choosing a more tractable distribution as a stand in. In terms of trying to fit to human data, variational approximations have the advantage of introducing biases that can be explicitly justified by a divergence measure from the true distribution given particular computational constraints. This opens up a new set of algorithms that can be used for developing rational process models.

7.3. Factorization and Modularity

In addition to computational constraints, FBL incorporates the psychological intuitions about modularity in the mind that motivated LBL. Intuitions about modularity have taken many forms. Fodor (1983) gave criteria for evaluating the strength of modularity in the mind. The form used here is very weakly modular, because top-down information can have an effect which breaks Fodor's property of information encapsulation. Modularity has been supported for peripheral processes, as envisioned by Marr (1982), though it has been found that in some cases modularity is more of a useful heuristic than a complete description of separation in visual processing (Schenk & McIntosh, 2010). Other researchers have proposed modules for central processes, claiming with the "massive modularity" hypothesis that there are task-specific modules, such as for cheater-detection (Carruthers, 2006; Cosmides & Tooby, 1992). Still another proposal is that central modules perform particular information-processing tasks, especially those that have been identified by psychologists as underlying performance across a range of tasks (Bechtel, 2003). LBL is appealing from this final viewpoint, dividing processes along traditional psychological definitions of attention and learning.

LBL and FBL both cast modules as factorized probability distributions that are coordinated by statistically-motivated message passing – resulting in central modules with extraordinary flexibility. Other computational approaches have either worked out how to co-ordinate the output of peripheral modules (Bülthoff & Yuille, 1996) or cast modules as complete central procedures that are context-dependent (Jacobs et al., 1991). Here we have introduced modularity that results from sequential co-ordination of modules, and the use of message passing opens up ideas for much more active and principled co-ordination between modules.

One interesting case of modularity is the case where factorization does no harm: when the information is actually independent given the interpretation. For example, participants could be given visual and auditory information in order to estimate an object's location. Here factorization does not result in the loss of information because these sources are assumed to be independent. Interestingly, participants in this task take into account information about the variability of the cues, and give more weight to cues that are more reliable (Alais & Burr, 2004; Ernst & Banks, 2002). This sort of result is more congruent with FBL than LBL, because FBL passes along an entire distribution over outputs while LBL only passes along the mean of a distribution without information about its variability.

7.4. Conclusions

Kruschke (2006b) introduced the idea that trial order effects that involve both primacy and recency, such as highlighting, could be produced by using message passing between locally Bayesian modules to approximate full Bayesian models. Our work builds on this approach by developing a closely related alternative that is computationally justified, can also predict human-like trial-order effects with appropriate and not overly rapid shifts of attention, and may make better predictions for other experimental designs. Connections between existing models and machine learning algorithms give cognitive scientists access to a rich resource for developing alternative models that produce a range of behavior. Aside from psychological and computational constraints, an exciting prospect is that other constraints can be introduced by neural considerations. The approximations used in the brain are still a new area of investigation, though some work has been done on explaining neural activity using both variational (Friston, 2010; Gershman & Wilson, 2010) and sampling explanations (Fiser et al., 2010). By constraining our search it is hoped that the approximations used in the mind can be identified.

The authors thank John Kruschke, Stephen Denton, Peter Dayan, and David Shanks for helpful discussions.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Bechtel, W. (2003). Modules, brain parts, and evolutionary psychology. In S. J. Scher, & F. Rauscher (Eds.), *Evolutionary Psychology: Alternative Approaches* (pp. 211–227). Norwell, MA: Kluwer Academic Publishers.

Beckers, T., De Houwer, J., Pineo, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 238–249.

Boyen, X., & Koller, D. (1998). Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (pp. 33–42).

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Bülthoff, H. H., & Yuille, A. L. (1996). A Bayesian framework for the integration of visual modules. In T. Inui, & J. L. McClelland (Eds.), *Attention and Performance 16: Information Integration in Perception and Communication* (pp. 49–70). Cambridge, MA: MIT Press.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 837–854.

Collins, E. C., Percy, E. J., Smith, E. R., & Kruschke, J. K. (2011). Integrating advice and experience: learning and decision making with social and nonsocial cues. *Journal of Personality and Social Psychology*, *100*, 967–982.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms*. MIT Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind* (p. 163228). Oxford: Oxford University Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*. MIT Press.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: the case of trial order. In N. Chater, & M. Oaksford (Eds.), *The Probabilistic Mind* (pp. 431–452). Oxford, UK: Oxford University Press.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In *Advances in Neural Information Processing Systems*.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

De Houwer, J. D., & Beckers, T. (2002). Second-order backward blocking and unovershadowing in human causal learning. *Experimental Psychology*, *49*, 27–33.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fearnhead, P. (1998). *Sequential Monte Carlo methods in filter theory*. Ph.D. thesis University of Oxford.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119–130.

Fodor, J. A. (1983). *The Modularity of the Mind*. Cambridge, MA: MIT Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

Gershman, S. J., & Wilson, R. (2010). The neural costs of optimal control. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 712–720).

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Harris, J. A., & Livesey, E. J. (2008). Comparing patterning and biconditional discriminations in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 144–154.

Harris, J. A., Livesey, E. J., Gharaei, S., & Westbrook, R. F. (2008). Negative patterning is easier than a biconditional discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 494–500.

Hershberger, W. A. (1986). An approach through the looking-glass. *Learning & Behavior*, *14*, 443451.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, *15*, 219–250.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, *82*, 35–45.

Kamin, L. J. (1968). 'Attention-like' processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–33). Coral Gables, FL: University of Miami Press.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 3–26.

Kruschke, J. K. (2001a). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1385–1400.

Kruschke, J. K. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Kruschke, J. K. (2006a). Locally Bayesian learning. In R. Sun (Ed.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 453–458). Hillsdale, NJ: Erlbaum.

Kruschke, J. K. (2006b). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*, 677–699.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*, 210–226.

Kruschke, J. K. (2009). Highlighting: A canonical experiment. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (pp. 153–185). volume 51.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*, 636–645.

Kruschke, J. K., & Denton, S. E. (2010). Backward blocking of relevance-indicating cues: Evidence for locally Bayesian learning. In C. J. Mitchell,

& M. E. L. Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour* (pp. 273–304). Oxford, UK: Oxford University Press.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*, 830–845.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 937–944).

Lovibond, P. F., Been, S.-L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, *31*, 133–142.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 185–190). Austin, TX: Cognitive Science Society.

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

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Medin, D. L., & Bettger, J. G. (1991). Sensitivity to changes in base-rate information. *The American Journal of Psychology*, *104*, 311–332.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Melchers, K. G., Lachnit, H., & Shanks, D. R. (2006). The comparator theory fails to account for the selective role of within-compound associations in cue-selection effects. *Experimental Psychology*, *53*, 316–320.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis MIT Boston.

Minka, T. (2005). *Divergence measures and message passing*. Technical Report MSR-TR-2005-173 Microsoft Research Ltd. Cambridge, UK.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Massachusetts: The MIT Press.

Mitchell, C. J., Lovibond, P. F., & Condoleon, M. (2005). Evidence for deductive reasoning in blocking of causal judgments. *Learning and Motivation*, *36*, 77–87.

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. New York: Dover Publications.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552.

Rescorla, R. A. (1972). “Configural” conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology*, *79*, 307–317.

Rescorla, R. A. (1973). Evidence of unique stimulus account of configural conditioning. *Journal of Comparative and Physiological Psychology*, *85*, 331–338.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rojas, R. R. (2010). *Explaining Human Causal Learning Using a Dynamic Probabilistic Model*. Ph.D. thesis University of California, Los Angeles.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: mechanistic versus rational approaches. *Memory & Cognition*, *36*.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to the rational model of categorization. *Psychological Review*, *117*, 1144–1167.

Sanborn, A. N., Mansinghka, V., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*, 411–437.

Schenk, T., & McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, *1*, 52–62.

Sewell, D. K., & Lewandowsky, S. (in press). Attention and working memory capacity: Insights from blocking, highlighting, and knowledge restructuring. *Journal of Experimental Psychology: General.*, .

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology*, *97B*, 1–21.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 405–415.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychological Bulletin and Review*, *17*, 443–464.

Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2004). Childrens causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303–333.

Spence, K. W. (1952). The nature of the response in discrimination learning. *Psychological Review*, *59*, 89–93.

Sutton, R. S. (1992). Gain adaptation beats least squares? In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 161–166).

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In *Advances in Neural Information Processing Systems 15* (pp. 35–42). Cambridge, MA: MIT Press.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*, 127151.

Table 1: Canonical Highlighting Design

Phase	Number of Trials	Items
First	$2 * N_1$	$I.Pe \rightarrow E$
Second	$3 * N_2$	$I.Pe \rightarrow E$
	$1 * N_2$	$I.Pl \rightarrow L$
Third	$1 * (N_2 + N_1)$	$I.Pe \rightarrow E$
	$3 * (N_2 + N_1)$	$I.Pl \rightarrow L$

Vandorpe, S., De Houwer, J., & Beckers, T. (2007). Outcome maximality and additivity training also influence cue competition in causal learning when learning involves many cues and events. *Quarterly Journal of Experimental Psychology*, *60*, 356–368.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

8. Appendix

In this appendix we provide a description of the design of the canonical highlighting experiment. The "canonical" design, shown in Table 1, equalizes the number of $I.Pe \rightarrow E$ and $I.Pl \rightarrow L$ trials over the entire experiment. Within each phase the trials were randomly ordered. We used a canonical design in which $N_1 = 10$ and $N_2 = 5$, repeating the experiment 100 times for each model to average over order effects.

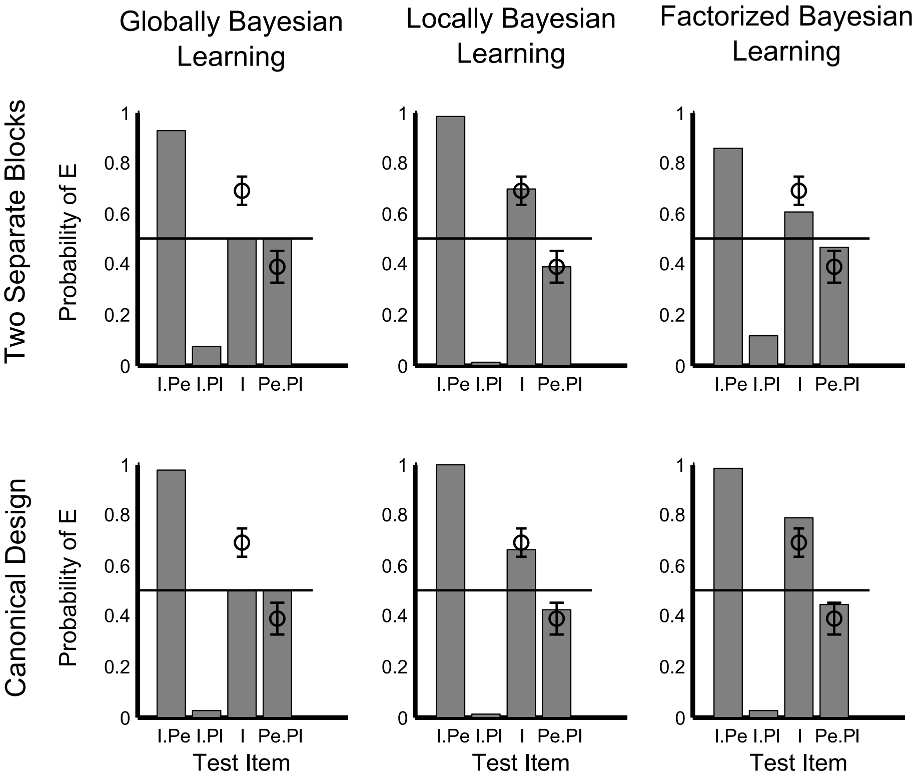

Figure 2: Highlighting predictions for Globally Bayesian Learning (GBL), Locally Bayesian Learning (LBL), and Factorized Bayesian Learning (FBL) for two experimental designs (explained in the main text). Experimental results from Kruschke (2009) are plotted on each graph with circles. Error bars around the circles show 95% confidence intervals for the human data. The bar plots show the model predictions of outcome E, where the line marks equal preference between predictions of E and L. A standard set of input cues is tested in each model: the original training sets of input cues $I.Pe$ and $I.Pl$, as well as the critical tests of input cue I and input cues $Pe.Pl$. Each set of model predictions was made using the same parameters as used in Kruschke (2006b) for highlighting.

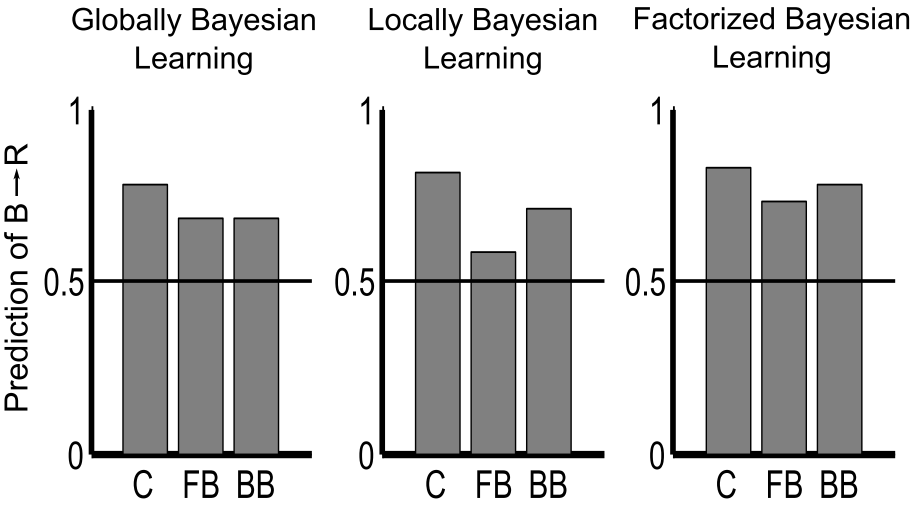

Figure 3: Blocking predictions for Globally Bayesian Learning (GBL), Locally Bayesian Learning (LBL), and Factorized Bayesian Learning (FBL). Each bar plot shows the strength of the $B \to R$ prediction after the control trials (C), after all of the forward blocking (FB) training trials, and after all the backward blocking (BB) training trials. Each set of model predictions was made using the same parameters as used in Kruschke (2006b) for highlighting.

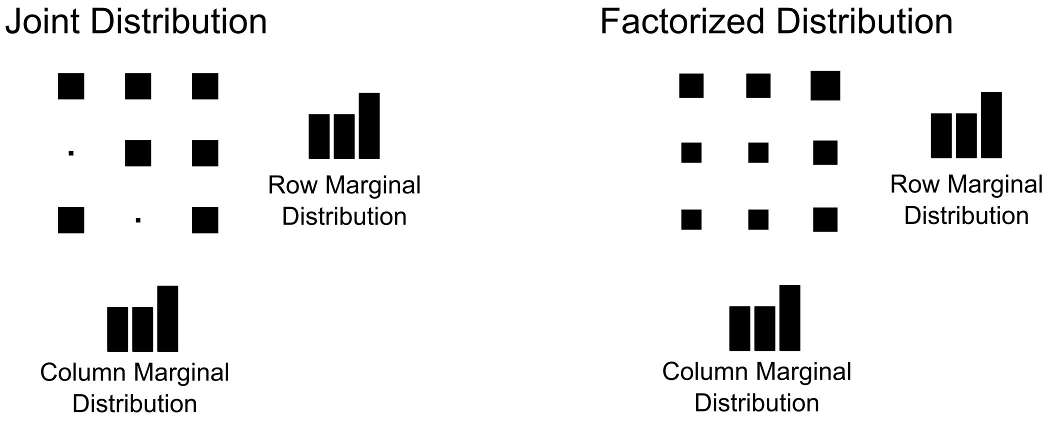

Figure 4: Example of a factorized distribution constructed from the marginals of a joint distribution. Each cluster of nine boxes shows a joint probability distribution, where the probability is equal to the area of a box (akin to a Hinton plot). The row of a box indexes the value of one variable, while the column of the box indexes the value of a second variable. The bar plots are marginal distributions of either the row or the column variable.

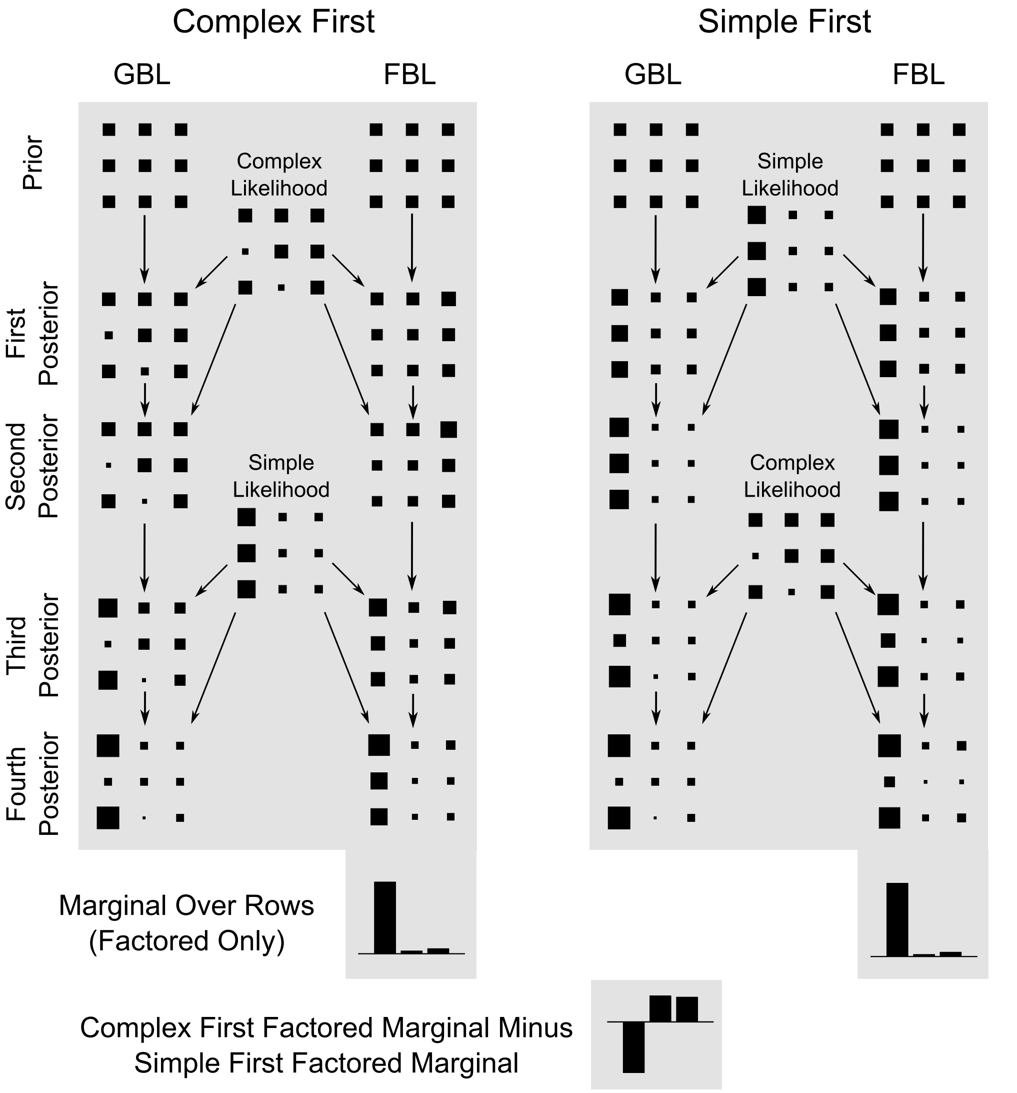

Figure 5: Example comparison of the updating process of Globally Bayesian Learning (GBL) and Factorized Bayesian Learning (FBL). Each cluster of nine boxes represents a joint probability distribution, and probabilities are equal to the areas of the boxes (akin to a Hinton plot). The column of a box indexes the setting of the first variable, while the row of the box indexes the setting of the second variable. Within each gray area, GBL and FBL begin with the same prior distributions and are updated with the same likelihoods. The left gray box shows the results of updating GBL and FBL with a complex likelihood before a simple likelihood, while the right gray box shows the results of updating GBL and FBL with the same likelihoods in the reverse order. At the bottom of the figure, the final marginal distributions for each column are shown for FBL. The difference plot at the bottom illustrates how sequential updating has produced order-dependent marginals, with the vertical axis rescaled to emphasize the differences.

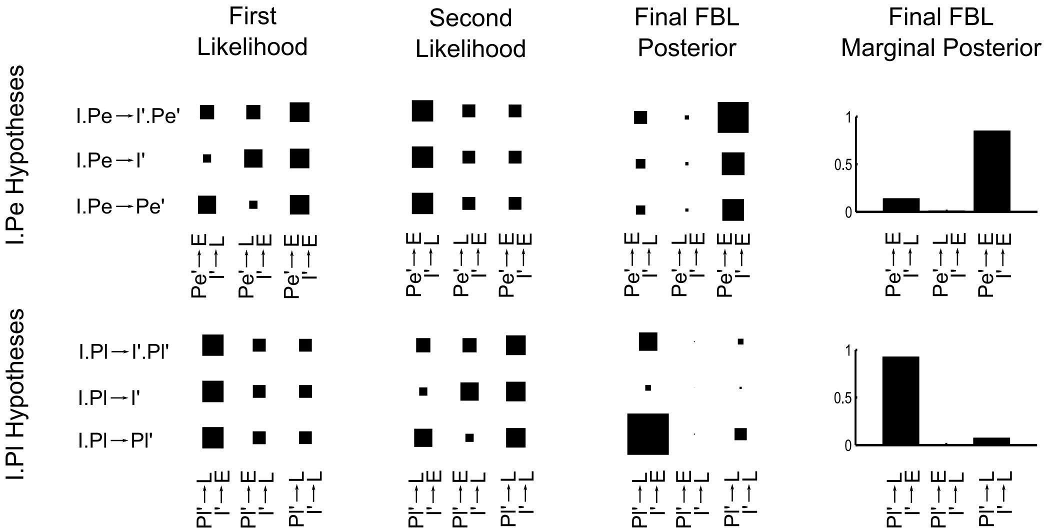

Figure 6: Illustration of how the effects of highlighting arise from Factorized Bayesian Learning (FBL). Each cluster of nine boxes shows a joint probability distribution, where the probability is equal to the area of a box (akin to a Hinton plot). Each row within a cluster corresponds to a different set of hypotheses about how the input cues activate the attended cues. Each column within a cluster corresponds to how the attended cues activate the outcome. Each row of plots corresponds to a different set of hypotheses. The first and second columns display the likelihoods used in the first and second block of trials respectively. The third column displays the posterior distributions of FBL following all training trials and the fourth column shows the final posterior distributions again, but marginalized over the hypotheses about how input cues activate attended cues.

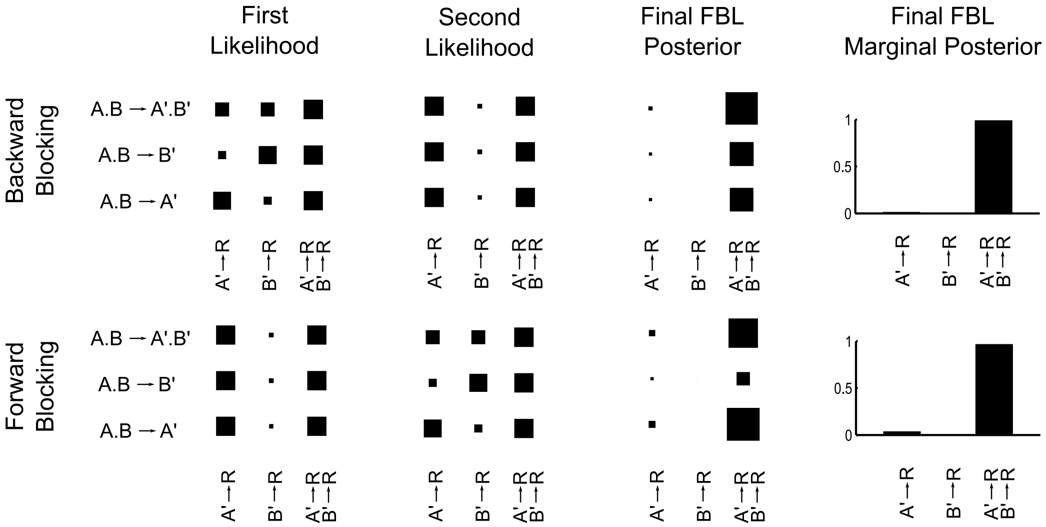

Figure 7: Illustration of how the effects of blocking arise from Factorized Bayesian Learning (FBL). Each cluster of nine boxes shows a joint probability distribution, where the probability is equal to the area of a box (akin to a Hinton plot). Each row within a cluster corresponds to a different set of hypotheses about how the input cues activate the attended cues. Each column within a cluster corresponds to how the attended cues activate the outcome. Each row of plots corresponds to a different training order condition. The first and second columns display the likelihoods used in the first and second block of trials respectively. The third column displays the posterior distributions of FBL following all training trials and the fourth column shows the final posterior distributions again, but marginalized over the hypotheses about how input cues activate attended cues.

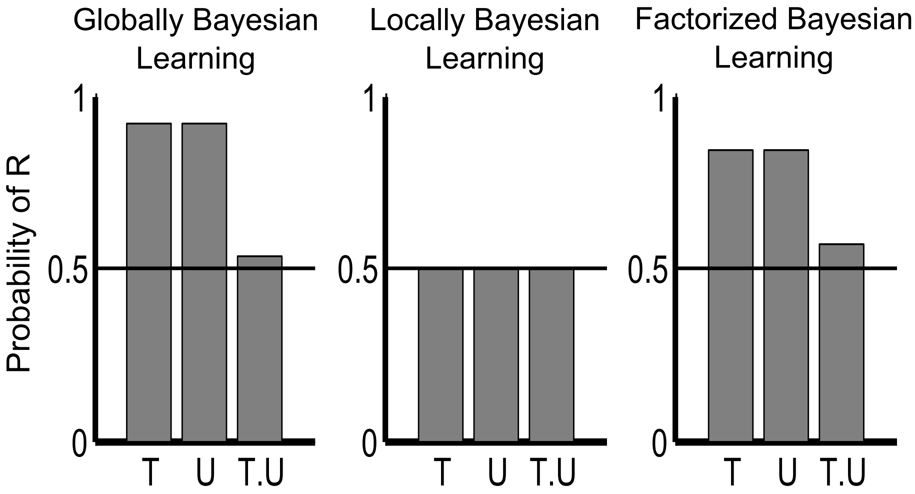

Figure 8: Exclusive-OR (XOR) predictions for Globally Bayesian Learning (GBL), Locally Bayesian Learning (LBL), and Factorized Bayesian Learning (FBL). Each bar plot shows the probability of *R* prediction after testing with the single cues alone (T and U) or with the compound cue ($T.U$). Each set of model predictions was made using the same parameters as used in Kruschke (2006b) for highlighting.