Running head: Causality in time

Causality in time: Explaining away the future and the past

David E. Huber

University of California, San Diego

[Do not cite without permission]

David E. Huber

Department of Psychology

University of California, San Diego

9500 Gilman Drive

La Jolla, CA 92093-0109

phone: (858) 822-1630

fax: (858) 534-7190

e-mail: dhuber@psy.ucsd.edu

Abstract

This chapter is presented in three sections corresponding to three models that incorporate Bayesian explaining away between different sources. The first section considers primes and targets as potential sources without reference to time. The original ROUSE model is reformulated as a generative model, arriving at the original equations but with slightly different dependence assumptions. The second section considers a model in which past time steps explain away future times steps, thereby producing perceptual sensitivity to the onset of new objects (i.e., new events). The resultant dynamics are related to the dynamics of neural habituation in several important ways. The third section considers a model in which future time steps explain away past time steps, thereby producing sensitivity to the offset of old objects (i.e., old events). By cascading layers, a working memory system is developed that represents the temporal rank ordering of objects regardless of their specific durations (i.e., scale free sequential information).

In recent years, Bayesian models of cognition have effectively explained a wide variety of cognitive behaviors ranging from visual perception (Yuille & Kersten, 2006) and eye movements (Najemnik & Geisler, 2005) to episodic memory (Dennis & Humphreys, 2001; Glanzer & Adams, 1990; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) and implicit memory (Colagrosso, Mozer, & Huber, 2004; Mozer, Colagrosso, & Huber, 2002; Schooler, Shiffrin, & Raaijmakers, 2001) to semantic memory (Steyvers, Griffiths, & Dennis, 2006) and syntax (Dennis, 2005; Griffiths, Steyvers, Blei, & Tenenbaum, 2004). This approach brings together recent advances in computer science and engineering related to graph theory and probability theory (Pearl, 1988) in combination with the Rational Analysis of Anderson (1987; Schooler & Anderson, 1997). The Rational Analysis supposes that cognition is a problem of optimizing processes to reflect the environment that we live in. In other words, people are adapted to their environment, and this optimization problem is best understood with Bayesian statistics.

Despite the success of these Bayesian models, they say little about neural processing (although see Pouget, Dayan, & Zemel, 2003). In line with David Marr's three levels for theorizing about cognition (Marr, 1982), the Bayesian approach could be placed at the abstract *computational* level, rather than the lower levels of *algorithm* or *implementation*. However, as seen in this chapter, in some cases there is a more or less direct correspondence between the level of neural implementation and the level of abstract computation. The particular example under consideration examines dynamic neural processing and the ubiquitous observation of neural habituation. The claim is made that

neural habituation is the brain's trick for implementing Bayesian inference for sequences of events in time.

In this chapter: i) the notion of explaining away is introduced within a static Bayesian model; ii) the static model is augmented by explicitly representing time in a causal graph such that previous perceptions explain subsequent perceptions (a model of new events); and iii) the same model is considered under the assumption that temporal causality is reversed such that subsequent perceptions explain previous perceptions (a model of old events). In light of these results, inference over time may specify the dynamics of perception (explaining the near future) and short-term memory (explaining the recent past).

I. Responding Optimally with Unknown Sources of Evidence (ROUSE)

Before presenting the full causal model that explains the "why" of neural habituation, several background pieces are needed. First, immediate priming data are reviewed as explained by the ROUSE model, which includes the notion that primes serve as an explanation for subsequent targets (i.e., explaining away is the central contribution of the model). Second, 'generative' Bayesian models are introduced, in which the same causal structure that produces observations also guides inference. Third, the inferential process of 'explaining away' (otherwise known as discounting) is introduced, which arises from a particular kind of causal structure. Finally, the ROUSE model is reformulated as a truly generative model and the original equations are re-derived starting with a causal graph under the assumption that entire distributions of objects serve as causes. This first section

is solely concerned with static models (i.e., time is not explicitly represented), but the data and model serve as the basis for incorporating time in the second and third sections.

Explaining Away and Priming

There are many methods for examining the immediate effect of one word on another when we read. Such 'priming' effects usually produce facilitation when a prime word and a subsequent target word are related, with this facilitation resulting in faster responding for lexical decision or naming tasks or more accurate responding with threshold identification tasks. This is called priming because the first word 'primes the pump' for the second word. Such tasks have established a large number of prime-target relationships that produce facilitation, including but not limited to orthographic (Evett & Humphreys, 1981; Meyer, Schvaneveldt, & Ruddy, 1974; Peressotti & Grainger, 1999), phonemic (Goldinger, 1998; Lukatela, Eaton, Lee, Carello, & Turvey, 2002; Lukatela, Frost, & Turvey, 1998), associative-semantic (McKoon & Ratcliff, 1992; Mcnamara, 1992; McNamara, 1994; Meyer & Schvaneveldt, 1971; Perea & Gotor, 1997), and syntactic (Ferreira, 2003) similarity, to name just a few.

********** insert Figure 1 **********

In our research, we sought to understand the nature of priming itself rather than asking what primes what. Therefore, we modified the task of threshold identification so that we could examine the costs and benefits associated with priming (Huber, Shiffrin, Lyle, & Quach, 2002; Huber, Shiffrin, Lyle, & Ruys, 2001; Huber, Shiffrin, Quach, & Lyle, 2002; Weidemann, Huber, & Shiffrin, 2005). As seen in Figure 1, participants attempted to identify target words flashed at the perceptual threshold, with this threshold set separately for each participant. Rather than naming the flashed target, accuracy was

measured by means of a forced-choice between a target and a foil. Immediately before the flash of the target word, one or two primes were presented, which were on average non-diagnostic (i.e., primes were just as likely to indicate the wrong answer as the correct answer). Furthermore, participants were instructed that there was no effective strategy in relation to explicit use of the primes and trial-by-trial accuracy feedback reinforced this assertion. These procedures were implemented in order to assess the implicit inference process in the priming of perceptual evidence while reducing strategic responding.

Use of forced-choice testing allowed conditions that primed the correct answer (target-primed) but also other conditions that primed the incorrect answer (foil-primed). Not shown in the figure are still other conditions that primed neither choice word (neither-primed) or conditions that primed both choice words equally (both-primed). The example in Figure 1 and the data of Figure 2 tested repetition priming, which is found to sometimes produce facilitation but other times produce priming deficits (Bavelier, Prasada, & Segui, 1994; Evett & Humphreys, 1981; Hochhaus & Johnston, 1996; Huber et al., 2001). As seen in Figure 2, which was reported by Huber (submitted for publication), brief prime durations resulted in a ‘preference’ for whichever word was primed. This helped performance in the target-primed condition but harmed performance in the foil-primed condition (the neither-primed baseline condition lay between these conditions for all prime durations shown in Figure 2). However, for prime durations of 400 ms or longer, this preference reversed its direction, and the preference against choosing primed words lowered performance for the target-primed condition but raised performance when the foil was primed.

Complicating things further, the both-primed conditions, not shown in Figure 2, always revealed deficits as compared to the neither-primed baseline conditions (Huber, Shiffrin, Lyle et al., 2002; Huber et al., 2001; Huber, Shiffrin, Quach et al., 2002; Weidemann et al., 2005). In order to make sense of these results and other results with orthographic and semantic priming, Huber, Shiffrin, Lyle, and Ruys (2001) proposed a Bayesian model that included the offsetting forces of source confusion (unknown sources of evidence) and discounting (responding optimally). Source confusion assumes that people are confused about what was presented when and where, tending to mistake the prime for the target. Discounting is enacted at the level of individual features during the decision process in relation to the choice words and it assumes that features known to have been primed should be assigned a lower level of evidence.

********** insert Figure 2 **********

Beyond producing accurate fits to data such that the degree of source confusion and the degree of discounting can be quantified (e.g., similar to use of signal detection theory to quantify sensitivity and bias), ROUSE also proved remarkably effective in producing a priori predictions. For instance, the Bayesian discounting based on observed features predicts that even extreme discounting will fail to eliminate or reverse the direction of priming in some situations. Critically, even fully discounted features provide a measure of positive evidence, and, therefore, discounting efficacy relies upon a relative deficit. In other words, discounting only works if it serves to lower the evidence provided by features that would otherwise strongly indicate the target. Huber, Shiffrin, Lyle, & Quach (2002) analyzed the mathematics of ROUSE, revealing that discounting efficacy relies on 1) a sufficient number of features participating in the decision; 2) sufficient presentation

of the target; and 3) sufficient similarity between primes and primed alternatives. By separately increasing similarity between the choice words (e.g., choice between LIED and DIED, decreasing target duration (including a condition where no target was flashed), and decreasing similarity between primes (the number of letters in common), each of these predictions was empirically confirmed; these manipulations changed a condition that produced negative priming into one that produced positive priming.

Despite this success, the original ROUSE model was not truly generative because it did not use a causal structure to specify the joint probability distribution. Instead, it was assumed that in general choice words were independent and that features were independent, with these assumptions made in order to simplify the math rather than as the result of a particular causal structure. In what follows, I reformulate ROUSE as generative model by means of a plausible causal structure. In this reformulation, the original equations follow naturally from this structure, which stipulates instead that words are mutually exclusive and that features are only conditionally independent. These causal assumptions lend greater plausibility and wider applicability to ROUSE, releasing the original model from overly constraining assumptions of independence.

Generative Models of Cognition

Some Bayesian models can be termed ‘generative’, meaning that they define a causal structure that is used both for production and for inference. Examples of generative model are commonly found in the study of language where the same causal structure produces text and is used to understand the text produced by others (Dennis, 2005; Griffiths et al., 2004; Steyvers & Griffiths, 2005; Steyvers et al., 2006). Figure 3 provides

a more mundane example of a generative model that highlights the phenomenon of explaining away.

********** insert Figure 3 **********

According to Figure 3, there are two causes of the observation that a light is out: Either the electricity is out or the bulb is broken. It is important that the connection between these causes and the observation is drawn with an arrow that points in a particular direction. These arrows indicate the direction of causation and the pattern of causal links across the entire graph specifies whether causes and effects are dependent or independent (Castillo, Gutierrez, & Hadi, 1997). The part of the figure to the left is the real world producing observations from this causal structure. Inside the observers head, cognition assumes this causal structure and the goal is to infer states of the world based upon limited observations (e.g., the light is out, but the cause is unknown).

In Figure 3, the two causes converge on a common effect, which defines a situation in which ‘explaining away’ can occur. In the absence of any observations regarding the light (i.e., if it is not known whether the light is on or off), the two causes are truly unrelated and independent; the status of the electricity has nothing to do with whether the filament in the light bulb is intact. However, once the common effect is observed, this introduces dependence between the causes. For instance, imagine that you are in a room with single light bulb that suddenly goes out, leaving you in the dark. The first thing you might do is attempt to turn on the stereo. But why would you do this and what does the stereo have to do with the light bulb? The answer is that you’re establishing the status of one of the two possible causes. If the stereo works, then you know that the electricity is not out and at that point you are essentially certain that the light bulb needs to be replaced. One can

write the equation, p(bulb broken | light out, electricity on), which has value 1.0. However, what is the corresponding value of p(bulb broken | light out, electricity off)? This second expression is close to 0, and equal to the prior probability that the bulb is broken in general. In other words, this second expression is the probability of a coincidence such that the bulb happened to be break at the same time that the power went out. This is explaining away; if the electricity out provides an explanation of the light out, then the probability that the bulb is broken is much less. This defines a dependency between otherwise independent causes--the probability of bulb broken depends on the status of the electricity, but this is true only when observing that the light is out.

Reformulating ROUSE as a Generative Model

As seen in Figure 4 and Table 1, ROUSE assumes three possible causal sources for the observed features in the word identification task (these features could be thought of as letters, but our results suggest higher level semantic features also play a role in forced-choice threshold identification). All features exist in one of two observed states: Either features are observed to be active, or they are observed to be inactive. The generative process defined by this causal structure stochastically produces patterns of active/inactive features according to the probabilities found in Table 1. The appropriate equation depends on which sources could produce a particular feature based upon the “true” prime and target presented on a particular trial (noise matches all features). After stochastically producing a pattern of active/inactive features, the model next performs inference based on the observed states of feature activation. In this reformulation of ROUSE, the causal structure in Figure 4 is used to derive this inference.

********** insert Figure 4 **********

The original ROUSE model assumed that each word was a binary distribution, similar to the features. Therefore, features were divided into those that matched the target choice alternative versus those that matched the foil choice alternative under the assumption that the choice words were independent. However, this reformulation employs probability distributions over all possible words rather than a binary word representation. Thus, the prime is a distribution over all words for the status of the word that was presented first (the prime) and the target is a distribution over all words for the status of the word that was presented second (the target). The key calculation is inference for the target distribution under the assumption that the prime is a known word (i.e., one word in the prime distribution is set to probability 1.0 and the rest set to 0.0). Although the distribution for the second word is referred to as the 'target distribution', it is important to keep in mind that this distribution contains probabilities both for the particular target word as well as the particular foil word that subsequently appear in the force-choice test display. Furthermore, because this is a distribution over all possible words, the entire pattern of feature activation over all possible features is considered, rather than just the features that pertain to the choice alternatives. Features that do not match either choice word may play an important role in other paradigms, such as naming, in which case all responses need to be considered. However, in light of the particular choices that appear at the end of a forced-choice trial, the large number of features that do not match either choice alternative provide a constant term that applies equally to both choices. Thus, for this particular situation, only features that match the target or match the foil matter in forced-choice testing.

The conditional probabilities that determine the generation of feature activation (α for primes, β for targets, and γ for noise) are not necessarily the same conditional probabilities used for inference. In other words, one set of parameters generates feature activation (source confusion), but a slightly inaccurate set of parameters might be used to determine accuracy from the observed pattern of activation (discounting). In this manner discounting is often incommensurate with source confusion, resulting in positive or negative priming. These potentially misestimated probabilities correspond to the prime symbols in the upper portion of Figure 4. Active features that are a common effect of both a particular prime and a particular target are features that are explained away (discounted). Such features do not provide strong evidence in favor of the target considering that these features may have arisen from the prime rather than the target. If the estimate of prime activation, α', is set too low, then there is too little explaining away, resulting in a preference for primed words, but if the estimate is set too high, then there is too much explaining away, resulting in a preference against primed words.

$$p(T,P,N,F_1,F_2,F_3,...) = p(T)p(P)p(N)\prod_i p(F_i|T,P,N) \qquad \text{Eq. 1}$$

To infer the target distribution (i.e., calculate the posterior distribution of potential target words based on the observed pattern of active/inactive features), the causal structure seen in Figure 4 is used to simplify the joint probability distribution through the implied independence relationships. For this causal structure, the target distribution, the target distribution, and the noise sources are independent of each other in the absence of knowing the status of the features. Therefore, in writing out the joint probability distribution (see Equation 1), these three probabilities, p(T), p(P), and p(N) are extracted first in order to simply the expression; because the features are not observed at this point

in the factorization, there is no need to write these causes as conditional probabilities. Furthermore, with these causes now specified, the features become independent of each other. Therefore, the factorization is completed in Equation 1 through the product over all possible features, i, conditioned on the prime, target, and noise distributions.

$$p(T=a|P,N,F_1,F_2,F_3,...) = \frac{p(T=a,P,N,F_1,F_2,F_3,...)}{p(P,N,F_1,F_2,F_3,...)} \quad \text{Eq. 2}$$

Equation 2 is the probability that the target is a particular word (T=a) conditioned on the prime, noise, and all observed features. By convention, use of italics indicates that a variable may take on any possible value whereas use of non-italics indicates that a variable is set to a specific value. Equation 2 is definitional and true regardless of the causal structure. Inclusion of the causal constraints from Equation 1 yields Equation 3.

$$p(T=a|P,N,F_1,F_2,F_3,...) = \frac{p(P)p(N)}{p(P,N,F_1,F_2,F_3,...)} p(T=a) \prod_i p(F_i|T=a,P,N) \quad \text{Eq. 3}$$

The original ROUSE model was formulated for forced-choice performance but Equation 3 (and more simply Equation 4) can be applied to many situations, such as lexical decision, speeded naming, or threshold identification with naming as the response. Performance in these tasks is a function of the probability that the target is the correct answer, p(T=a), as compared to all other possible answers contained in the target distribution. The ratio to the right of the equals sign in Equation 3 combines terms that are the same across the entire target distribution, and, thus, this ratio does not change across different conditions of interest in most situations (i.e., provided that all primes, features, and noise sources are equally likely in the various experimental conditions, this ratio is constant). For convenience, this term is dropped and replaced with a proportional relationship, yielding Equation 4.

$$p(T=a|P,N,F_1,F_2,F_3,...) \propto p(T=a) \prod_i p(F_i|T=a,P,N) \quad \text{Eq. 4}$$

Equation 5 can be converted into a relationship of equality by calculating this expression for all possible target words followed by normalization against the sum of the calculations. The term, p(T=a), is the prior probability that a particular word is the target. This term is typically the same across all conditions. However, comparisons between different classes of objects (for instance high versus low frequency words) may affect these priors. In addition, task specific manipulations may serve to make some classes of objects more likely (e.g., an experiment that only uses nouns as targets), which would likewise affect these priors. The final term in Equation 4 is the probability of observing features in their known active/inactive states given a particular target and the known values of the prime and noise; it is this last term that enacts explaining away.

For the particular task of forced choice testing, the situation is further simplified by taking Equation 3 as calculated for the correct target word (T=a) and dividing it by Equation 3 as calculated for the incorrect foil (T=b), yielding the likelihood of choosing the correct answer, seen in Equation 5.

$$\frac{p(T=a|P,N,F_1,F_2,F_3,...)}{p(T=b|P,N,F_1,F_2,F_3,...)} = \frac{p(T=a)}{p(T=b)} \prod_i \frac{p(F_i|T=a,P,N)}{p(F_i|T=b,P,N)} \quad \text{Eq. 5}$$

In order to implement ROUSE using Equation 5 (or more generally Equation 4), specific probabilities need to be entered for observed feature activation conditioned on the observed sources. The equations in Table 1 not only generate patterns of active/inactive features, but these same equations are used to the fill in appropriate values for the probability of particular features existing in the observed active/inactive state given the particular target word under consideration (a or b) and the known prime word. The only difference is that estimated probabilities are used in the equations of Table 1 instead of the “true” probabilities. Thus, simulation with ROUSE consists of a generative

pass, which stochastically determines the particular feature observations, followed by inference to determine accuracy for that trial (however, see Huber, 2006, for a method of implementing ROUSE without stochastic feature generation).

To see how these equations are used, consider the particular feature T___, the particular prime TRIP, and the particular target TOWN. This feature matches both prime and target (and all features match noise), and, thus, the probability that this feature will be active is found in the lower right-hand entry of Table 1. If indeed this feature is activated, this same conditional probability is also used to infer the probability for the particular possible target words TRIP in the target distribution of Equation 4, although in this case estimated values are used for the parameters.

Conclusions from section I

Equation 5 is the same equation appearing in the original ROUSE model, and it specifies that forced-choice accuracy is the ratio of priors for the two choice words multiplied by the likelihood ratios that come from each of the features. Critically, derivation of Equation 5 in this reformulation did not assume that choice words were independent, such as was assumed in the original ROUSE model. Instead, Equation 5 was derived by assuming that the choice words are mutually exclusive because they are both part of the same discrete probability distribution for the second word presented in the sequence (i.e., the target distribution). Also, derivation of Equation 5 did not assume that the features are in general independent, such as was assumed in the original ROUSE model. Instead, Equation 5 was derived by assuming that the feature are in general dependent but that the features become independent when conditioned on the three sources. Because this reformulation is a generative model, these new assumptions were

drawn directly from the causal structure shown Figure 4. In developing this reformulation, Equation 4 was derived, which is generally applicable to any identification paradigm such as speeded naming, lexical decision, or threshold identification without forced choice testing. This provides an avenue for extension of ROUSE to these other identification paradigms.

II. Explaining Away the Future (what's new)

Despite the success of ROUSE in explaining a number of non-intuitive results related to similarity and to the efficacy of discounting (Huber, Shiffrin, Lyle et al., 2002), this static version of the model does not explain why brief prime durations result in too little explaining away while long duration primes result in too much explaining away. Instead, application of ROUSE requires a different set of parameters for each prime duration (i.e., fitting the data seen in Figure 3 would require 5 α values and 5 α' values). A better solution would be that these effects emerge naturally from a single set of parameters in a model that explicitly includes timing. Therefore, Huber and O'Reilly (2003) developed an alternative account based on the dynamics of neural habituation. Their model naturally captured the data of Figure 3 because brief primes result in lingering activation that blends with the target whereas long duration primes habituate and actually lessen the response to a repeated target. In this section, the neural habituation model is briefly summarized and it is demonstrated that a Bayesian model that explicitly incorporates time produces similar behaviors.

Neural Habituation

The brain includes many mechanisms that produce habituation in the presence of an ongoing stimulus. One way to quantify the joint action of these mechanisms is to drive a sending neuron (pre-synaptic membrane potential) while recording from a receiving neuron (post-synaptic membrane potential). This is precisely what Tsodyks and Markram did (1997), observing that the synapse appears to lose resources as a function of recent activity, thereby temporarily lowering the efficacy of the connection between the two cells. Abbott et al. (1997) developed a mathematical model of this habituation and Huber and O'Reilly (2003) derived a 'rate-coded' version of the model that does not require simulation of spiking neurons.

In the dynamic neural network of Huber and O'Reilly (2003), every simulated neuron within a 3 layer network activates (pre-synaptic membrane potential, *v*) according to Equation 6, in which *netin* refers to the summed excitatory input, L is a constant leak current, I is the strength of inhibition, and *o* refers to the post-synaptic effect of the sending cell.

$$\frac{\Delta v}{S} = (1-v)netin - v(L+Io) \qquad \text{Eq. 6}$$

Equation 6 specifies how the pre-synaptic membrane potential changes at every millisecond. Each layer of the model incorporates all-to-all inhibition (however, Equation 6 and the right-hand panel of Figure 5 only include self inhibition) and the simulated cells of each layer integrate information at a layer-specific rate, S.

$$\frac{\Delta a}{S} = [R(1-a)] - Do \qquad \text{Eq. 7}$$

Equation 6 is similar to a large number of other neural network models but the new element is a second time dependent variable, *a*, which captures the depletion, D, and

recovery, R, of synaptic resources in the face of ongoing synaptic output, *o*. Because D is greater than R, the synapse quickly loses its resources and requires some time to recover. Equation 7 presents the updating of this synaptic 'amplitude'. In order to produce output, *o*, the pre-synaptic cell needs to spike, which happens with probability *v*-θ (i.e., there is a threshold for activity), and the post-synaptic amplitude of each spike is *a*, resulting in Equation 8.

$$o = (v - \Theta)a \qquad \text{Eq. 8}$$

The three panels of Figure 5 show 1) observed behavior; 2) predicted behavior of the 3-layer dynamic neural network with habituation (the decision rule was to choose the word that reached its highest level of output first); and 3) an example of neural habituation for a simulated cell from the bottom visual feature layer of the model using Equations 6-8. Additional details are reported by Huber and O'Reilly (2003), but hopefully this is sufficient to give a qualitative sense of the model. As seen in Figure 5, there is a close correspondence in the initial build up and then elimination of the preference for repeated words as a function of prime duration and the build up and then habituation in the post-synaptic effect of visual features in the presence of ongoing input. Because the visual features project to orthographic features, this imparts lingering orthographic activation from the prime that reaches a peak but then habituates for longer primes. This lingering activation boosts a repeated target, but, with sufficient habituation, the lingering activation is reduced and, furthermore, the orthographic response becomes sluggish due to depletion of resources.

********** insert Figure 5 **********

A Cascaded Hidden Markov Model for New Events

In order to extend the ROUSE model by explicitly including inferences over time, a temporal causal structure is needed. Hidden Markov Models (HMMs) provide a mechanism to represent time in a causal model because time is implemented as a continually unfolding chain of steps, with each step causally related to the prior time step. HMMs successfully capture sequential processing and have been applied to long-term priming (Colagrosso et al., 2004; Mozer et al., 2002) and other phenomena. Perhaps their most well known use is in the realm of phonology and their application in speech recognition software (Jaffe, Cassotta, & Feldstein, 1964; Mari, Haton, & Kriouile, 1997; Ostendorf, Digalakis, & Kimball, 1996; Seward, 2004). The bottom half of Figure 6 (not including the event layer) portrays the classic form of an HMM, with links pointing forward in time and other links pointing from objects towards observations. For the remainder of this chapter, the more neutral term of 'objects' is used rather than 'words' because the models are broadly applicable to a wide variety of stimuli and paradigms. A standard HMM is a generative model because real objects in the world cause observations and inference is performed on these observations to determine which objects were likely to have produced them (i.e., perception as an inferential problem).

********** insert Figure 6 **********

It is often useful to breakdown the problem of inference over time into different layers of abstraction, such as with a generative grammar model of sentence processing (Crocker & Brants, 2000), or a perceptual layer followed by a response layer (Mozer et al., 2002). These models are referred to as 'cascaded HMMs' and are typically implemented by performing inference in lowest layer based on observations and then

performing inference in the next layer based on the results of the lower layer, etc. In other words, these models are implemented by severing the dependence relationships between levels of the inferential cascade. This is done because including dependence relationships between layers is computationally difficult, but, as presented next, inclusion of these dependencies naturally includes explaining away over time, thus producing behaviors that are similar to the neural dynamics of habituation.

In the notation applicable to Figure 6, specific points in time are indicated by the superscript on variables and a particular layer in the HMM is indicated by the subscript. In light of the model presented in section III, nodes are referred to with variable B, indicating Buffer. Dependencies in time (i.e., horizontal connections) correspond to conditional probabilities with parameter α and dependencies in detection (i.e., vertical connections) correspond to conditional probabilities with parameter β. In this manner, the static version of ROUSE is related to this dynamic version in terms of the explaining away from prior time steps (corresponding to the prime with parameter α) versus explaining away from new events, such as the second word in the sequence (corresponding to the target with parameter β). As with the static version of ROUSE, every node also includes noise as a potential source of activation with probability γ (not shown in Figure 6). Unlike the static version, this version does not include features and, instead, all nodes are entire distributions over all possible objects. This was largely for reasons of simplification, and the object node at each time step could be replaced with all possible feature nodes at each time step, thereby maintaining the ability of the model to represent similarity through proportions of shared features.

The model seen in Figure 6 includes an additional layer, ‘new events’, which is not typically included in an HMM. However, this is still in keeping with a generative model framework if it is assumed that events are the underlying cause of new objects. Thus, while things tend to persist over time with a previous object causing its ongoing presence at the next point in time (the α links), new events are the reason that objects change (the β links). Although the new event layer is the same distribution over objects as appears at the object layer, its interpretation is that of the event of a new object--thus, the new event layer is inherently transient. However, inclusion of α links in the event layer allows that events are not entirely discrete in time such that events linger for awhile (e.g., an ongoing event). This latter assumption is critical in explaining why neural habituation is not instantaneous. Traditional HMMs do not include explaining away and have no converging arrows. However, by including a second layer with dependencies between layers, this model is considered a factorial HMM (Ghahramani & Jordan, 1997) because the distribution at B_0 depends both on the previous time step and on the second layer at that point in time. In this manner, previous time steps explain away ongoing objects and the inference at the new event layer becomes mainly sensitive to the onset of new objects.

Unlike the static version of ROUSE, exact inference in this model is intractable considering that every time step depends on every other time step and every time step contains a factorial connection (Ghahramani & Jordan, 1997). Instead, approximate inference is achieved by means of Gibbs sampling (Albert & Chib, 1993). In the Gibbs sampling algorithm, the entire chain is initially randomly sampled according to the priors (e.g., every node at every time step is randomly set to a particular object as drawn from the uniform discrete distribution). Then, the entire chain is repeatedly stepped through in

permuted order. With each step, one node is chosen, the posterior probability distribution at that node is calculated based on the current values of the other nodes (i.e., local inference), and, finally, a new particular value for the chosen node is sampled from the newly calculated local posterior distribution. After an initial 'burn-in' period (set to 1,000 iterations in the reported results), the model sufficiently departs from the initial priors to allow collection of 'counts' that are used to determine the posterior distribution over the entire chain of time steps. Counts are the number of times that each object is sampled for every node and every time step. For the reported simulations, 100,000 counts were collected. Between collection of each count, the chain was stepped through 4 times in permuted order to provide relative independence between counts. The value of 4 was determined by examining the asymptotically low value for the correlation between one count and a subsequent count.

Gibbs sampling turns an intractable global inference problem into a relatively easy local inference problem. With this method, the causal graph need only provide the joint probability distribution for a local area as dependent upon just one node of the causal chain over time. Because the other values of the other nodes are known through stochastic sampling, this 'severs' the dependence relationships from propagating down the entire chain. The observations, *I*, are specified for a particular simulation (e.g., this is perceptual input) and so only two equations are needed to perform local inference. As before, these are derived from the causal structure (Figure 6 in this case). The first of these equations (Equation 9) is the probability that a node in the object layer, B_0^t, takes on the value j, conditional on the other nodes.

$$p(B_0^t = j) \propto p(B_0^t = j \mid B_0^{t-1}, B_1^t)\ p(B_0^{t+1} \mid B_0^t = j, B_1^{t+1})\ p(I^t \mid B_0^t = j) \qquad \text{Eq. 9}$$

As in the static version of ROUSE, there is no need to express this as an equality relationship and the posterior distribution is turned into a true distribution by dividing Equation 9 by the sum of Equation 9 as calculated for all values of the object j. The first term two terms in Equation 9 to the right of the proportional symbol include explaining away based on the prior time step and event perception (i.e., both an α link and a β link, with noise γ always as a source). The third term only includes a β link between the object node and an observation. In calculating the local posterior distribution for a particular object node, the equations of Table 1 are used for the terms of Equations 9 as dictated by the existence of a possible α source or a β source and whether the nodes match or mismatch in object value.

$$p(B_1^t = j) \propto p(B_1^t = j \mid B_1^{t-1})\ p(B_1^{t+1} \mid B_1^t = j)\ p(B_0^t \mid B_0^{t-1}, B_1^t = j) \qquad \text{Eq. 10}$$

Equation 10 is the local posterior probability distribution of an event node. This equation consists of two α only links and one explaining away term with α and β. In theory, the causal chain continues endlessly into the future and endlessly into the past. However, simulations are over a prescribed number of time steps. Equations 9 and 10 are modified slightly for the first and last time steps because there is no known value for the time step prior to the first or the time step after the last time step. Instead, it is assumed that these time steps beyond the boundaries are set to uniform priors.

The neural habituation model of Huber and O'Reilly (2003) not only produces realistic habituation functions with ongoing stimulation, but it also produces behaviors at the offset of a stimulus that correspond to the finding that some cells linger in their response while others exhibit a rebound effect and fall below baseline firing rates with the offset of the preferred stimulus (Duysens, Orban, Cremieux, & Maes, 1985).

Lingering responses are seen for a simulated cell that has a 0 baseline firing rate (the low baseline panel of Figure 7) and are due to gradual leaking of membrane potential. In contrast, a rebound effect is seen for a simulated cell that is driven above zero even in the absence of the preferred stimulus (the high baseline panel of Figure 7). For both the high and low baseline situations there is a lingering depletion of synaptic resources past removal of the preferred input. However, synaptic depression is not the same as post-synaptic inhibition, and can only produce a relative deficit. Therefore, the slower process of recover only produces the apparent rebound in the non-zero baseline situation, serving to keep post-synaptic depolarization below the ultimate baseline level until synaptic resources are fully recovered.

As seen in Figure in Figure 7, not only is the cascaded new event HMM capable of producing something analogous to neural habituation, but it also produces offset behaviors similar to simulated cells with low or high baseline firing rates. In these simulations, the cascaded HMM was shown 3 different objects in succession with each object observed for 10 time steps. Unlike the static version of ROUSE, this was not done by generating stochastic patterns of activation. Instead, the goal here was to examine the temporal properties of the inference process and so an idealized sequence of observations was presented to the model. Thus, while this model is capable of generating observations, this aspect of the model is not currently utilized. The figure shows the probabilities at the object and event layers appropriate to the second object (the first and third objects are presented to place the model in an appropriate “baseline” state before and after the second object). At the time of the onset of the second object, the object layer reaches its peak probability for that object, and, thus, the event probability reaches its peak value as

well. However, subsequent time steps are progressively explained away by prior time steps within the object layer and so the event probability is progressively lowered (but it retains some probability due to the inclusion of α links within the event layer).

********** insert Figure 7 **********

With offset of an observation, the probabilities return to baseline levels. However, the path that the probabilities take in reaching baseline is a function of how many other potential objects exist (high versus low priors) as well as the strength of the α links. Based on the forward directed α links, the time step just prior to offset stipulate that the object probabilities should continue unabated. However, because a new observation overrides object perception, this forward prediction is not born out. Therefore, the event layer has even a greater reason to expect anything but the previous object (i.e., a rebound due to excessive explaining away). Conversely, the α links within the event layer work against this effect and tend to produce maintenance of the previous event. When there are few alternative possibilities (5 total possible objects for high priors), the rebound effective is large enough to overcome a maintenance, but when there are many alternative possibilities (100 total possible objects for low priors), the rebound effect is negligible and the model produces some degree of persistence.

Conclusions from Section II

Having developed a cascaded HMM with event detection, the goal is not necessarily to promote its widespread use in modeling experimental results, particular in light of the difficult nature of inference. Instead, this serves as an existence proof, demonstrating the similar dynamic properties in both the Bayesian model and the dynamics of neural habituation. If the cascaded HMM is treated as an ideal observer model under the

assumed causal structure, then neural habituation can be viewed as a useful mechanism for calculating approximate inference over time. Thus, neural habituation is not merely an artifact or some sort of capacity limitation, but, instead, neural habituation is a trick that has evolved to solve the difficult problem of deciding whether a particular observation is something new within the ongoing stream of observations (i.e., event detection), or, whether the observation should instead be considered a lingering response to something that was previously identified. In relating the causal graph to neural processing, it is suggested that neural habituation is the natural result of a perceptual system that is constantly trying to predict the future and determine what's new.

III. Explaining Away the Past (what's old)

This section considers the same cascaded HMM from section II with the assumption that the causal time links are reversed in their direction, pointing towards the past rather than the future. This is done, not in the service of any particular behavioral result, but, rather, simply as an exploration of the computational properties under this assumption. The claim is not that causation is literally flowing backward (although note that the equations of Physics are symmetric in time). Instead, the question is whether inference based on backwards causation may be useful. The model in section II assumed forwards causation and produced behaviors that were predictive of the near future, which is likely to be useful in many situations in which a response is required when the environment changes. In contrast, a model with backwards causation "postdicts" the past, which could be useful when events from the recent past are no longer at hand, but perhaps important

for sequential processing. As seen below, this inferential process can be used to build a working memory system that keeps track of recent events in the order in which they occurred (see Miyake & Shah, 1999 for different views on working memory).

Working memory buffers are created by cascading one layer into the next within this backwards causation model. Analogous to event detection in section II, this produces the most recent past event in the layer above object perception, the event before that in the next layer, and so on. Thus, the model keeps track of the last N objects in the order in which they arrived. This representation may be desirable for sequential learning because it is 'scale free' over time, keeping track of the rank order relationship of the last N objects regardless of how quickly or slowly they occurred.

Higher-order Dependencies in a Hidden Markov Model

Figure 8, not including the dashed links, is a classic HMM with forward directed links from the last time step. For reasons of computational complexity, applications of HMM's typically only include these first-order dependencies (i.e., what follows what). However, some applications of HMM's find that higher-order dependencies produce more accurate identification based on the expanded temporal context (Mari et al., 1997). Inclusion of the dashed links in Figure 8 is an example of a third-order HMM. Beyond reasons of computational complexity, one advantage of first-order HMM's is that first-order relationships are scale free and are preserved regardless of the rate at which a sequence is presented. However, higher-order sequential information is important in many situations and the model presented next provides an alternative method for including higher-order links that depends on the sequence of objects, rather than the timing of objects.

********** insert Figure 8 **********

A Cascaded Hidden Markov Model for Old Events

Figure 9 gives an example of the backwards causation model with one layer for object perception and three additional layers for past events (working memory buffers). The model can be built with any number of layers although each additional layer adds complexity and becomes progressively imprecise in representing past objects. The reported simulations used 4 total layers but the dashed links in the figure were not implemented. The dashed links are included in the figure to portray the analogue of higher-order links from past objects in the sequence. Even though these dashed links appear within the same time step, they are effectively links from the past, similar to the dashed links in Figure 8. However, unlike Figure 8, these links constitute a scale-free higher-order dependence because the representations in the buffers are largely independent of the rate at which objects are presented.

********** insert Figure 9 **********

As with the forwards causation model, approximate inference in this backwards causation model is implemented with Gibbs sampling and the necessary equations are the joint probability distributions for a node at each layer, with all other nodes existing with known values. In general, the local posterior probability distribution for some layer n, where n is neither the object layer 0 nor the top buffer N, is calculated according to Equation 11, which includes three different explaining away terms with both α and β links (as before, noise is always a source with probability γ). Equations 12 and 13 modify this expression for the object layer B_0, and the top buffer B_N, respectively.

$$p(B_n^t = j) \propto p(B_n^t = j \mid B_n^{t+1}, B_{n+1}^t)\, p(B_n^{t-1} \mid B_n^t = j, B_{n+1}^{t-1})\, p(B_{n-1}^t \mid B_{n-1}^{t+1}, B_n^t = j) \quad \text{Eq. 11}$$

$$p(B_0^t = j) \propto p(B_0^t = j \mid B_0^{t+1}, B_1^t)\, p(B_0^{t-1} \mid B_0^t = j, B_1^{t-1})\, p(I^t \mid B_0^t = j) \quad \text{Eq. 12}$$

$$p(B_N^t = j) \propto p(B_N^t = j \mid B_N^{t+1})\ p(B_N^{t-1} \mid B_N^t = j)\ p(B_{N-1}^t \mid B_{N-1}^{t+1}, B_N^t = j) \quad \text{Eq. 13}$$

The two simulations reported in Figure 10 were designed to demonstrate the buffering capacities and scale-free nature of this backwards causation model. The parameters for the model in section II were set to produce abrupt event detection through explaining away ($\alpha = 1$ and $\beta = 1$), such that the last time step fully explains a persistent object, but an event also fully explains a new object. In contrast, the backwards causation model is used to build a working memory system and must attempt to satisfy the contradictory goals of maintaining objects while also allowing that new objects enter the buffering system. This was achieved by setting maintenance to a decent but not excessive level ($\alpha = .7$) and by setting detection to a sizable but nonetheless smaller value ($\beta = .3$).

********** insert Figure 10 **********

For both simulations there were 100 possible objects and the sequence provided to the observation nodes consisted of the same 3 objects cycling in the same order. These three objects correspond to the solid, dashed, or dotted probability lines, respectively. The simulation reported in a) changed objects every 10 times steps whereas the simulation in b) changed objects every 5 time steps. This is analogous to use an HMM for speech recognition with higher-order dependencies as applied to talkers who differ in rate of speech by as much as a factor of two. The vertical bars in each simulation demonstrate that both simulations are able to faithfully represent the order of the last 3 objects (assuming that the object with the highest probability is selected); in both simulations the object layer B_0 represents the current object, the first working memory buffer B_1 represents the most recent object, and the second working memory buffer B_2 represented the object before that. The third working memory buffer is not shown in Figure 10 for

reasons of simplicity, but, also, because the last buffer often fails to faithfully represent previous objects because it does not have a layer above to explain away its representation from objects even further in the past. Figure 10 was accomplished with N=3 (4 layers). Additional simulations with N set to higher values revealed that the model can continue to buffer past objects, but that with each additional layer the model becomes progressively inaccurate and the probabilities become closer and closer to uniform priors. Presumably this is because even very small amounts of noise result in temporal uncertainty ad the noise is propagated across multiple inferential steps.

Conclusions from Section III

The cascaded HMM developed in section II, which proved useful for detecting new events, is likewise useful as a model of past events and maintenance of previous objects when the direction of the causality in time is reversed. Furthermore, cascading one event layer into the next produced a buffering system such that the cascade keeps track of previous objects in the order they were presented. This could be viewed is a scale-free rank order working memory system that encodes sequences in a similar manner regardless of the rate that they occur. Implementation of causal links from higher layers directly onto current object perception (i.e., the dashed links in Figure 9) could be used to establish sequential learning that applies to similar sequences even if the sequences do not occur at the same rate.

The reported simulations were implemented with identity transition matrices. In other words, α and β links consisted of α or β for the conditional probability values along the diagonal for each object as a function of every other object, and the off diagonal elements were all set to zero. This was done because the simulations were an exploration of

dynamic properties rather than an analysis of perceptual competence or sequential learning. Future applications of the model could include these off diagonal elements so that previous objects predict different subsequent objects, perhaps as learned from large corpora of follows and precedes data (Dennis, 2005; Griffiths et al., 2004), or perhaps as specified by associative norms (Nelson, McEvoy, & Schreiber, 1994). The same off diagonal elements may play a role in working memory too, providing an explanation as to why known sub-sequences enable apparent increases in capacity through chunking (Baddeley, 1994).

The new event perception model in section II used parameters to promote abrupt onsets. In contrast, the parameters for the past events model of working memory were set to allow some degree of maintenance as well as sensitivity to new offsets of objects. The setting of maintenance (α) and change detection (β) is tricky in several respects. If change detection is too high, then presentation of a new observation immediately propagates through the entire cascade of buffers. Conversely, if maintenance is set too high, then it is difficult to swap objects between buffer states. What appears to work is that both parameters are set to sizable levels with maintenance greater than change detection. In addition, noise needs to be set as low as possible.

In theory the backwards causation model is also a generative model, although, like the model in section II, it was not used for its ability to produce sequences of observations. Nevertheless, conditional probabilities (the parameters α and β) that produce desirable inferential characteristics are perhaps interpretable in light of environmental statistics; presumably the ratio of α to β is something like the frequency for maintenance of objects as compared to the frequency for events (object change).

Recently we performed a series of experiments that manipulated the statistics of repetition frequency within different blocks of trials in threshold word identification (Weidemann, Huber, & Shiffrin, submitted). In using the static version of ROUSE to explain the results, we found that the most sensible account assumed the system adopts different estimates of source confusion (i.e., different α‘ levels) to handle repetition frequency. Furthermore, the empirical evidence suggested that these adaptations to the local statistics were rapid as revealed by the lack of differences between the first block of trials with a new frequency and a subsequent block that contained the same frequency of repetitions. Despite this rapid adaptation, discounting behavior was also sensitive to the global statistics of the entire experiment and with each change in the local statistics, behavioral changes in discounting were attenuated. Therefore, we modeled the current estimate of source confusion as a mixture of the current statistical regularities combined with the previous statistical regularities. This suggests that learning appropriate parameters for detection and maintenance is multifaceted, involving both short-term and long-term statistical regularities.

Conclusions

Section I of this chapter set the stage by presenting an overview of generative models based on causal graphs, with the reformulation of the ROUSE model providing a specific example. Sections II and III were an exploration of this same explaining away model as applied to time steps rather than temporally unspecified primes and targets. In section II, it was demonstrated that explaining away from past time steps produces a model that is

sensitive to the onset of new objects (i.e., new events) and that these inferential dynamics are similar to the dynamics of neural habituation both in terms of the lessening of a response in light of an ongoing observation, but, also in terms of the offset behavior with either a lingering response or a rebound response. In section III it was demonstrated that explaining away from future time steps produces a model that is sensitive to the offset of old objects (i.e., past events) and that this can be used to produce a scale-free working memory system that buffers objects in rank order, regardless of the timing of the objects. Combing the models form sections II and III, these two types of temporal inference may serve as fundamental building blocks for perception (predicting the future) and working memory (postdicting the past).

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science, 275*(5297), 220-224.

Albert, J. H., & Chib, S. (1993). Bayes Inference Via Gibbs Sampling of Autoregressive Time-Series Subject to Markov Mean and Variance Shifts. *Journal of Business & Economic Statistics, 11*(1), 1-15.

Anderson, J. R. (1987). A Rational Analysis of Human-Memory. *Bulletin of the Psychonomic Society, 25*(5), 342-342.

Baddeley, A. (1994). The Magical Number 7 - Still Magic after All These Years. *Psychological Review, 101*(2), 353-356.

Bavelier, D., Prasada, S., & Segui, J. (1994). Repetition blindness between words: nature of the orthographic and phonological representations involved. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 1437-1455.

Castillo, E., Gutierrez, J. M., & Hadi, A. S. (1997). *Expert systems and probabilistic network models*. New York: Springer-Verlag.

Colagrosso, M. D., Mozer, M. C., & Huber, D. E. (2004). Mechanisms of skill refinement: A model of long-term repetition priming. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. (pp. 687-692). Hillsdale, NJ: Erlbaum Associates.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research, 29*(6), 647-669.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science, 29*(2), 145-193.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452-478.

Duysens, J., Orban, G. A., Cremieux, J., & Maes, H. (1985). Visual cortical correlates of visible persistence. *Vision Res, 25*(2), 171-178.

Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33A*(4), 325-350.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language, 48*(2), 379-398.

Ghahramani, Z., & Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning, 29*(2-3), 245-273.

Glanzer, M., & Adams, J. K. (1990). The Mirror Effect in Recognition Memory - Data and Theory. *Journal of Experimental Psychology-Learning Memory and Cognition, 16*(1), 5-16.

Goldinger, S. D. (1998). Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem. *Perception & Psychophysics, 60*(6), 952-965.

Griffiths, T., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating Topics and Syntax. In L. K. Saul (Ed.), *Advances in Neural Information Processing Systems* (Vol. 17, pp. 537-544): MIT Press.

Hochhaus, L., & Johnston, J. C. (1996). Perceptual repetition blindness effects. *Journal of Experimental Psychology-Human Perception and Performance, 22*(2), 355-366.

Huber, D. E. (2006). Computer simulations of the ROUSE model: An analytic method and a generally applicable techniques for producing parameter confidence intervals. *Behavior Research Methods, 38,* 557-568.

Huber, D. E. (submitted for publication). Immediate Priming and Cognitive Aftereffects.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science, 27*(3), 403-430.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Quach, R. (2002). Mechanisms of source confusion and discounting in short-term priming 2: effects of prime similarity and target duration. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*(6), 1120-1136.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review, 108*(1), 149-182.

Huber, D. E., Shiffrin, R. M., Quach, R., & Lyle, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming: 1. Effects of prime duration and prime recognition. *Memory & Cognition, 30*(5), 745-757.

Jaffe, J., Cassotta, L., & Feldstein, S. (1964). Markovian Model of Time Patterns of Speech. *Science, 144*(362), 884-&.

Lukatela, G., Eaton, T., Lee, C. H., Carello, C., & Turvey, M. T. (2002). Equal homophonic priming with words and pseudohomophones. *Journal of Experimental Psychology-Human Perception and Performance, 28*(1), 3-21.

Lukatela, G., Frost, S. J., & Turvey, M. T. (1998). Phonological priming by masked nonword primes in the lexical decision task. *Journal of Memory and Language, 39*(4), 666-683.

Mari, J. F., Haton, J. P., & Kriouile, A. (1997). Automatic word recognition based on second-order hidden Markov models. *Ieee Transactions on Speech and Audio Processing, 5*(1), 22-25.

Marr, D. (1982). *Vision : a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724-760.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: mediated priming revisited. *J Exp Psychol Learn Mem Cogn, 18*(6), 1155-1172.

Mcnamara, T. P. (1992). Theories of Priming .1. Associative Distance and Lag. *Journal of Experimental Psychology-Learning Memory and Cognition, 18*(6), 1173-1190.

McNamara, T. P. (1994). Theories of priming: II. Types of primes. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*(3), 507-520.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227-234.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1974). Functions of Graphemic and Phonemic Codes in Visual Word-Recognition. *Memory & Cognition, 2*(2), 309-321.

Miyake, A., & Shah, P. (1999). *Models of working memory : mechanisms of active maintenance and executive control*. Cambridge ; New York: Cambridge University Press.

Mozer, M. C., Colagrosso, M. D., & Huber, D. E. (2002). A rational analysis of cognitive control in a speeded discrimination task. In T. G. Dietterich, S. Becker & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 51-57). Cambridge, MA.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*(7031), 387-391.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1994). *The University of South Florida word association, rhyme and word fragment norms.*Unpublished manuscript.

Ostendorf, M., Digalakis, V. V., & Kimball, O. A. (1996). From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *Ieee Transactions on Speech and Audio Processing, 4*(5), 360-378.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. San Mateo, Calif.: Morgan Kaufmann Publishers.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition, 62*(2), 223-240.

Peressotti, F., & Grainger, J. (1999). The role of letter identity and letter position in orthographic priming. *Perception & Psychophysics, 61*(4), 691-706.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381-410.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32*(3), 219-250.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108*(1), 257-272.

Seward, A. (2004). A fast HMM match algorithm for very large vocabulary speech recognition. *Speech Communication, 42*(2), 191-206.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Steyvers, M., & Griffiths, T. (2005). The topics model for semantic representation. *Journal of Mathematical Psychology, 49*(1), 92-93.

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*(7), 327-334.

Tsodyks, M. V., & Markram, H. (1997). The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc Natl Acad Sci U S A, 94*(2), 719-723.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2005). Spatiotemporal confusion and compensation in visual word perception. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 40-61.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (submitted). Prime diagnosticity in short-term repetition priming: Is primed evidence discounted even when it reliably indicates the correct answer? *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301-308.

Author Note

David E. Huber, Department of Psychology, University of California, San Diego. This research was supported by NIMH Grant MH063993-04. Correspondence concerning this article should be addressed to David E. Huber, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail may be sent to dhuber@psy.ucsd.edu.

Table 1. Conditional probabilities for combinations of sources

Possible Sources	Inactive/Mismatch	Active/Match
Noise	$(1-\gamma)$	$1-(1-\gamma)$
Prime or Noise	$(1-\alpha)(1-\gamma)$	$1-(1-\alpha)(1-\gamma)$
Target or Noise	$(1-\beta)(1-\gamma)$	$1-(1-\beta)(1-\gamma)$
Target or Prime or Noise	$(1-\beta)(1-\alpha)(1-\gamma)$	$1-(1-\beta)(1-\alpha)(1-\gamma)$

Figure Captions

Figure 1. Sequence of events for the priming data shown in Figure 2. Primes were presented for 17, 50, 150, 400, or 2,000 ms. Target flash durations were determined separately for each participant to place performance at 75% on average (perceptual threshold). Mask durations were set such that the total time between target flash onset and test display onset was 500 ms. Trial-by-trial feedback was provided to minimize strategic responding.

Figure 2. Accuracy for the experiment shown in Figure 1 (Huber, submitted for publication). As prime duration increased, there was initially a preference to choose primed words, resulting in better performance when the target was primed and worse performance when the foil was primed. For longer prime durations this preference changed to a preference against choosing primed words. This experiment also included the neither-primed condition, which appeared roughly halfway between the target-primed and foil-primed conditions at all prime durations.

Figure 3. Example of a generative model that includes real causes (electricity out and broken bulb) and an observation (loss of light). Electricity out and broken bulb are usually independent but become dependent once the loss of light is observed. Given the loss of light, if one cause is known to exist, then the probability of the other cause is less (explained away).

Figure 4. The causal structure that guides reformulation of ROUSE as a generative model. The Prime and Target nodes are distributions over all possible words. The features are binary, taking on value 0 or 1, depending on whether they have been activated by any of the three sources. The probability of feature activation conditional on only the Prime, only the Target, or only Noise, is equal to α, β, or γ respectively. For the Prime and Target these conditional probabilities only apply to features contained within a particular word whereas Noise always applies. The task of inference is to determine the Target distribution given the observed feature states and the observed Prime. This inference is achieved with estimated conditional probabilities, α', β', and γ'. When α is underestimated there is too little explaining away from the prime and when α is overestimated there is too much explaining away from the prime. Dark gray indicates active features and white indicates inactive features. Light gray is used to indicate that the particular prime word within the discrete Prime distribution is known.

Figure 5. Results with the 3-layer dynamic neural network of Huber and O'Reilly (2003), which included neural habituation. The right-hand panel illustrates neural habituation using parameters for the bottom visual feature layer as dictated by Equations 6-8. The simulated cell was driven with *netin*=1.0 and the other parameters were: S=.054, D=.324, R=.022, I=.3, θ=.15, and L=.15.

Figure 6. A cascaded hidden markov model that includes a layer for inferring objects based on observations and a second layer for inferring new events based on objects. Each vertical column represents a different time step. Because of the converging causal arrows

in the object layer, an object from the last time step can explain away a new event. Therefore, event perception is mainly sensitive to the change of objects (i.e., object onset). All nodes represent discrete distributions over all possible objects and all nodes include noise as a possible source, with probability γ (not shown).

Figure 7. Comparison between the forward directed cascaded HMM seen in Figure 6 and the neural habituation model of Huber and O'Reilly (2003). Neural habituation was simulated with the same parameters as Figure 5 for a 500 ms presentation with *netin* set at 1. Low baseline corresponded to *netin* at 0 all other times and high baseline corresponded to *netin* at .3 all other times. For the cascaded HMM, α and β were both set to 1.0 in order to produce strong explaining away from prior time steps as well as strong sensitivity to new observations. 5 possible values existed for every node in the high prior simulation and γ was set at .2. 100 possible values existed for every node in the low prior simulation and γ was set at .02.

Figure 8. A traditional HMM with first-order dependencies is shown for the solid links. Inclusion of the dashed links creates a third-order HMM in which the next time step is predicted based upon the factorial combination of the previous three time steps.

Figure 9. A cascaded hidden markov model that includes backwards causation. This produces a model that is sensitive to order of past events. Therefore, these layers implement working memory buffers that represent the sequence of objects regardless of the rate of the sequence. The dashed links represent higher-order terms from past events

and are analogous to the dashed lines in Figure 8, except that these apply to past events rather than past time steps.

Figure 10. Results from simulating the backwards causation model seen in Figure 9 with a repeating sequence of three objects. Simulation in the left graph presented a new object every 10 time steps while simulation in the right graph presented a new object every 5 time steps. The solid, dashed, and dotted lines are the probabilities of the three presented objects. The vertical bars in each simulation highlight a point in time to demonstrate that both simulations represent with highest probability the current object in B_0, the most recent object in B_1, and the object before than in B_2, despite the fact that one sequence was twice as fast. All nodes consisted of distributions over 100 possible objects. The other parameters were $\gamma = 1 \times 10^{-10}$, $\alpha = .7$, and $\beta = .3$.

Figure 1.

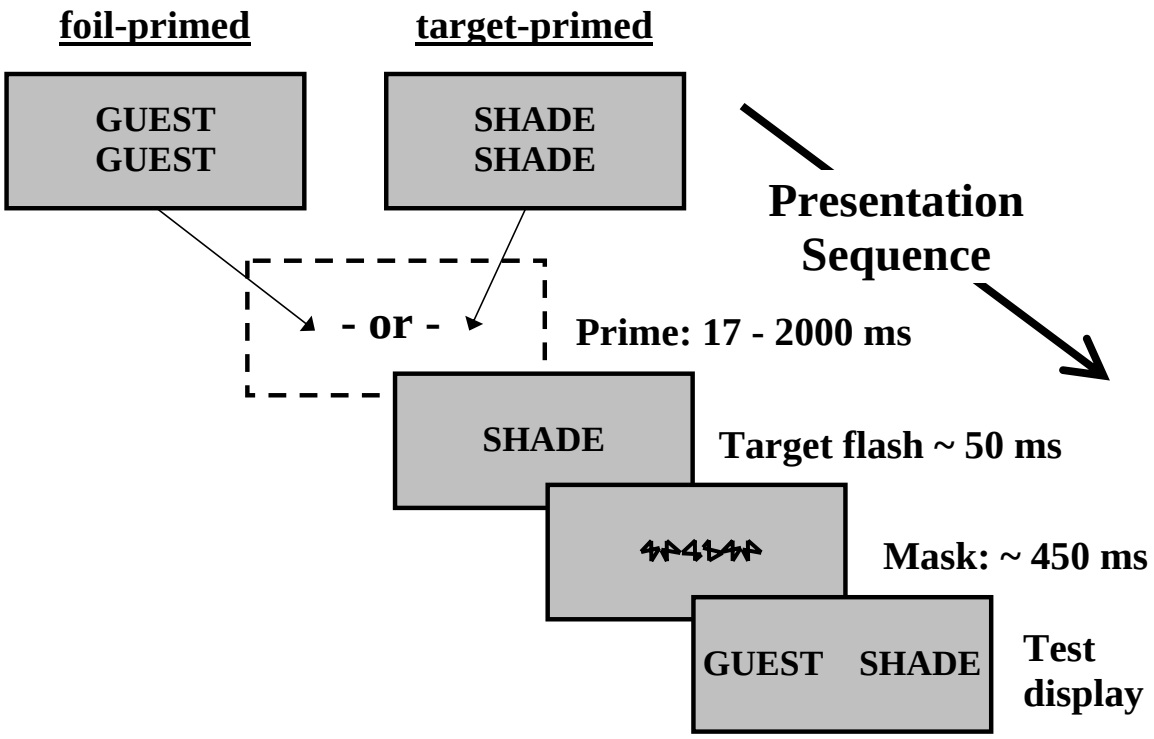
foil-primed
target-primed
GUEST
GUEST
SHADE
SHADE
Presentation
Sequence
- or -
Prime: 17 - 2000 ms
SHADE
Target flash ~ 50 ms
Mask: ~ 450 ms
GUEST SHADE
Test
display

Figure 2.

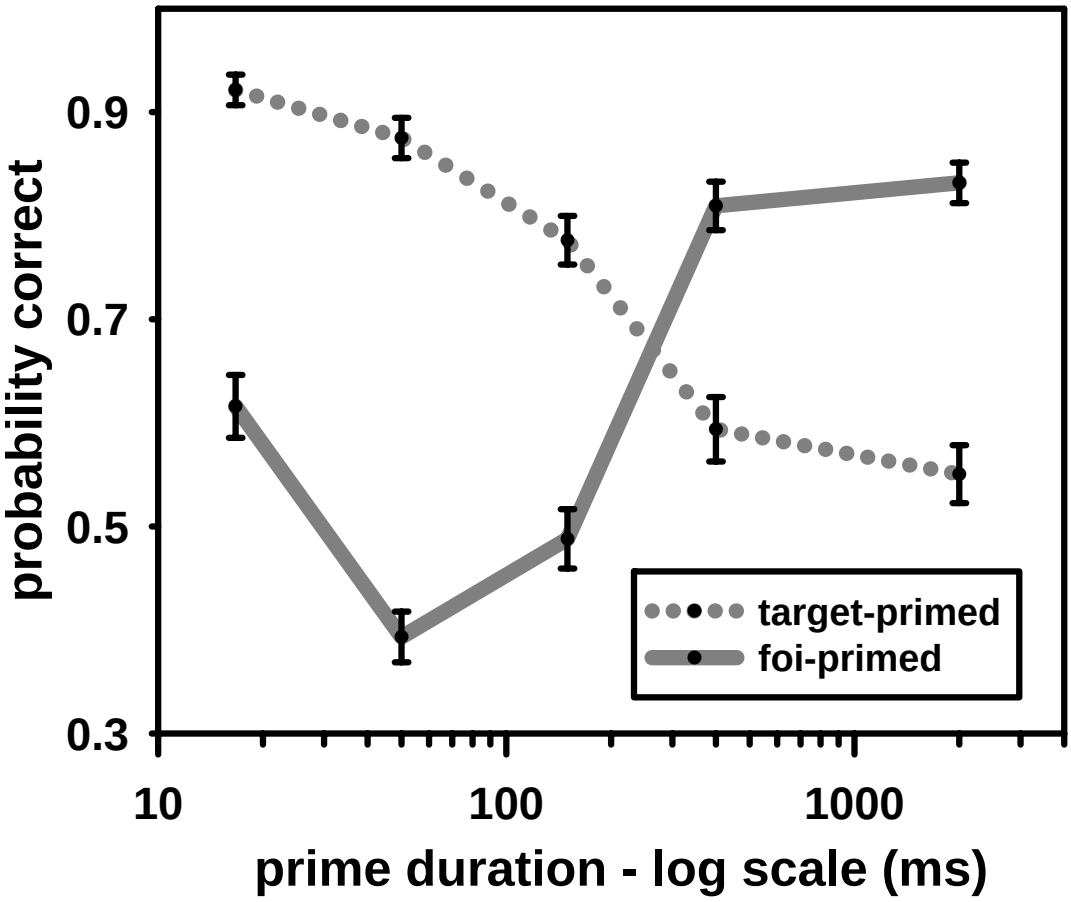
probability correct
0.9
0.7
0.5
0.3
10
100
1000
prime duration - log scale (ms)
target-primed
foi-primed

Figure 3.

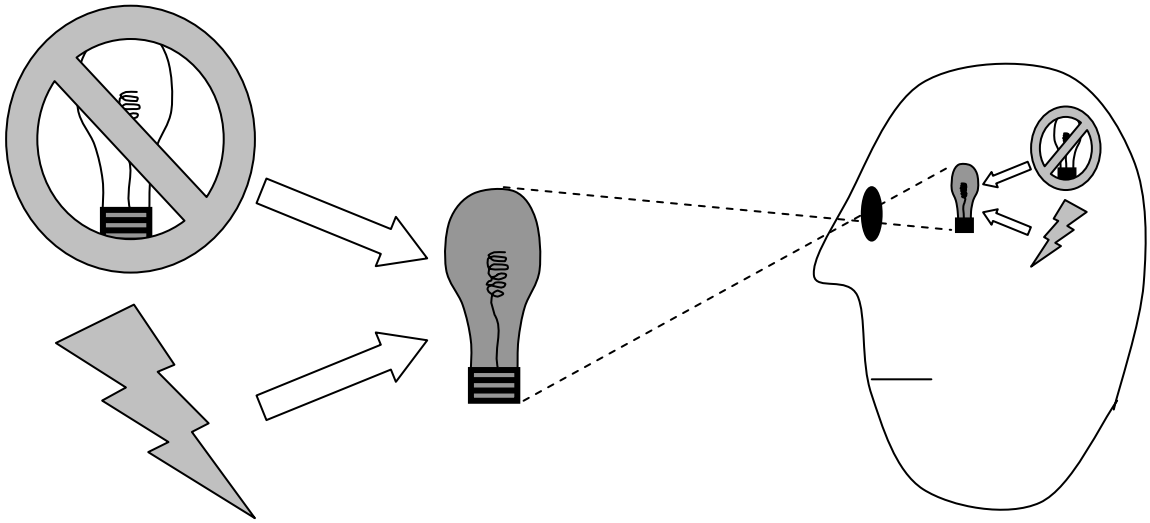

Figure 4.

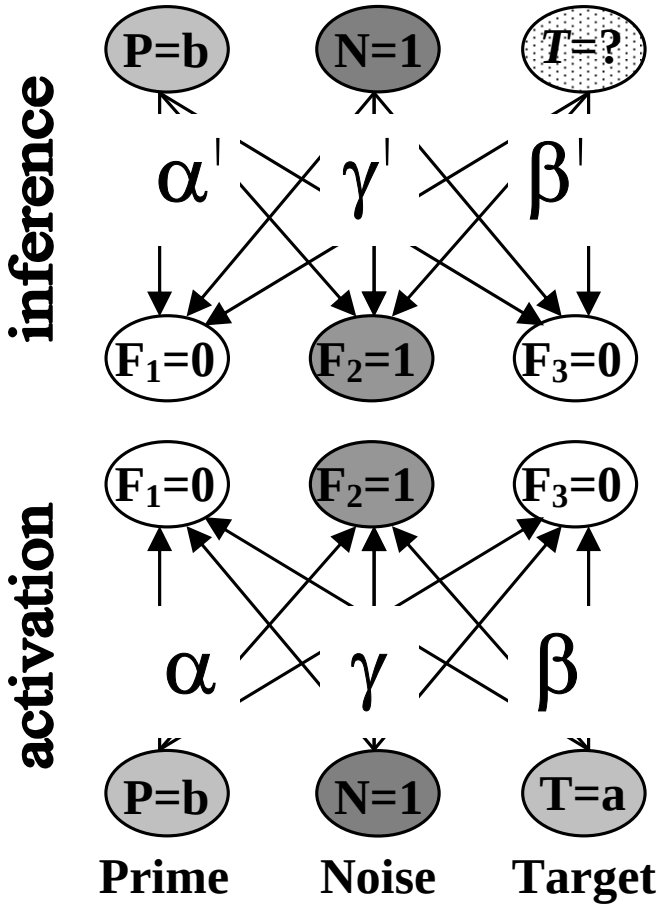
P=b
N=1
T=?
α'
γ'
β'
inference
F1=0
F2=1
F3=0
F1=0
F2=1
F3=0
α
γ
β
activation
P=b
N=1
T=a
Prime
Noise
Target

Figure 5.

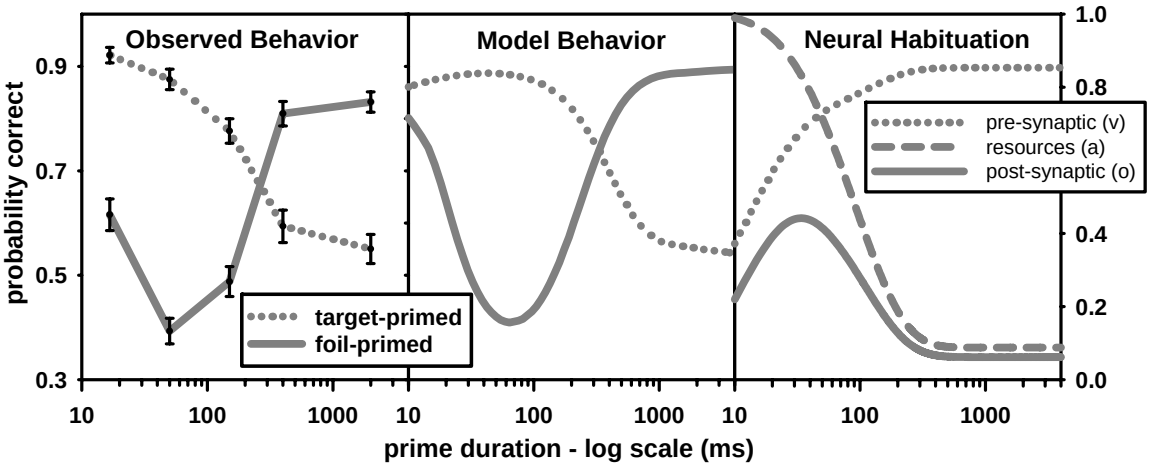
Observed Behavior
Model Behavior
Neural Habituation
probability correct
0.9
0.7
0.5
0.3
1.0
0.8
0.4
0.2
0.0
10
100
1000
prime duration - log scale (ms)
target-primed
foil-primed
pre-synaptic (v)
resources (a)
post-synaptic (o)

Figure 6.

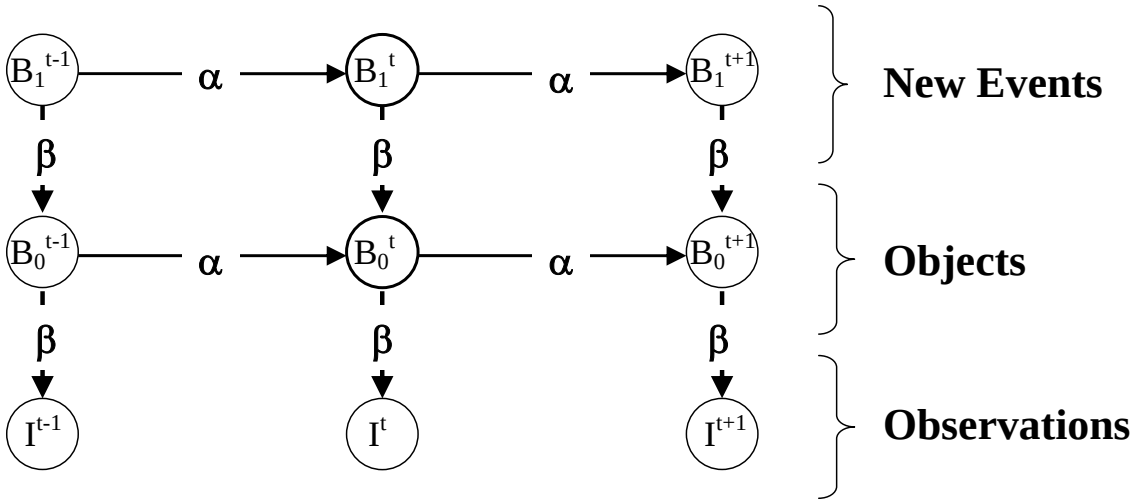
B_1^{t-1}
α
B_1^{t}
α
B_1^{t+1}
β
β
β
New Events
B_0^{t-1}
α
B_0^{t}
α
B_0^{t+1}
β
β
β
Objects
I^{t-1}
I^{t}
I^{t+1}
Observations

Figure 7.

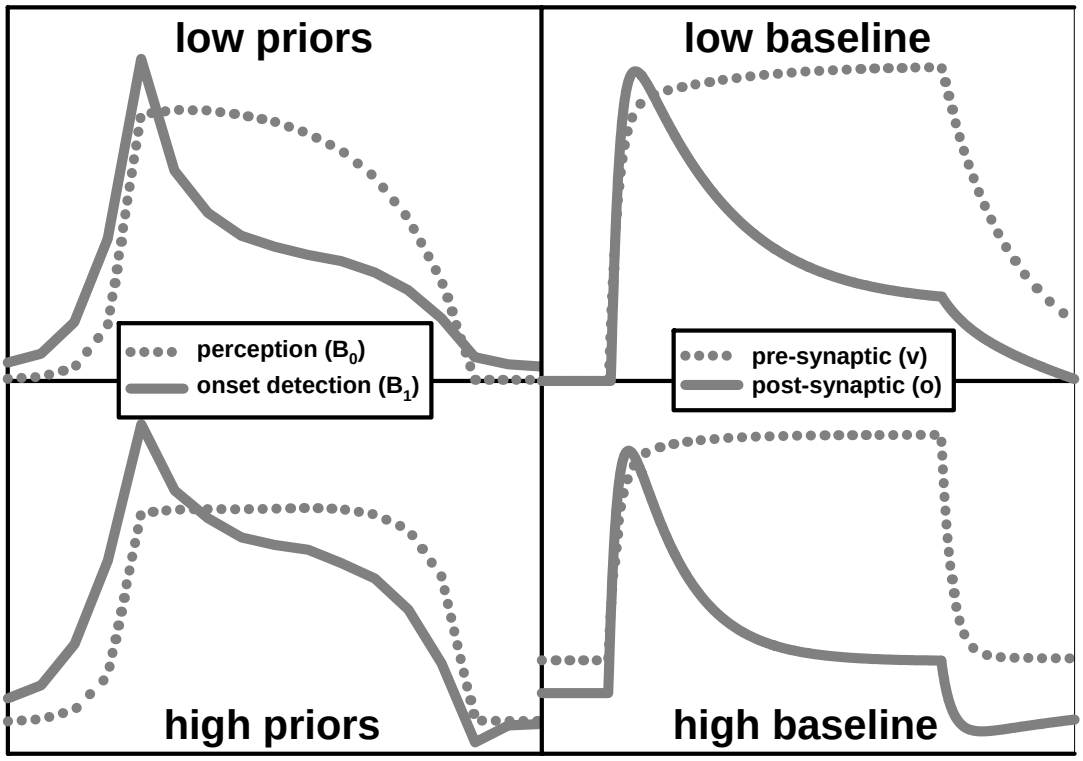
low priors
low baseline
perception (B_0)
onset detection (B_1)
pre-synaptic (v)
post-synaptic (o)
high priors
high baseline

Figure 8.

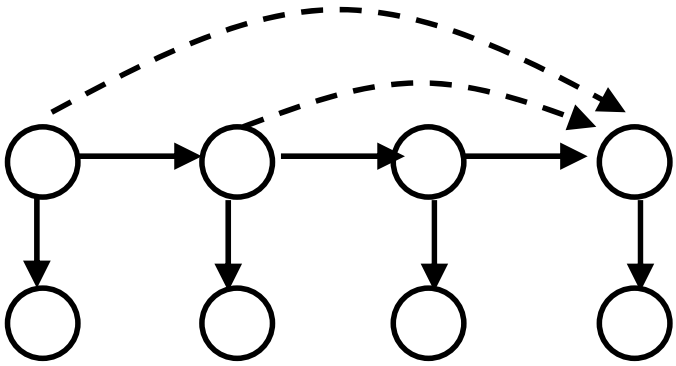

Figure 9.

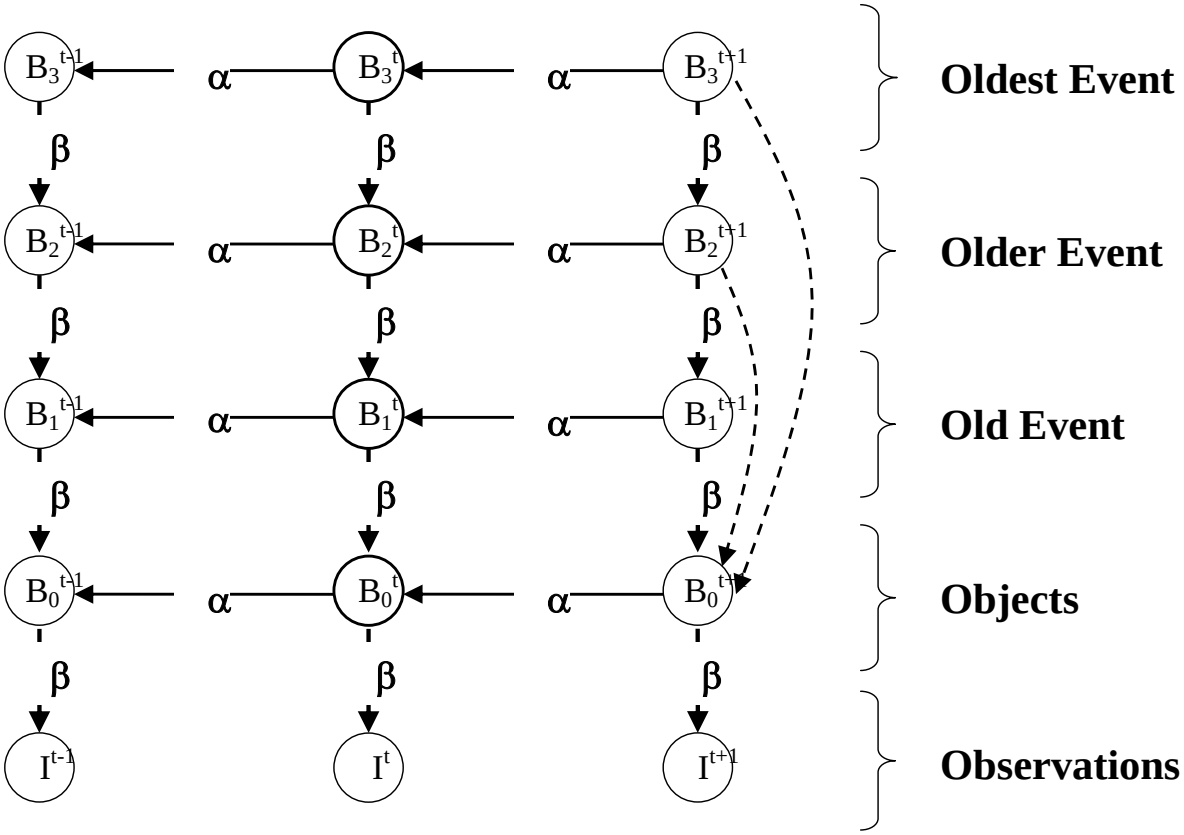
B_3^{t-1}
B_3^{t}
B_3^{t+1}
B_2^{t-1}
B_2^{t}
B_2^{t+1}
B_1^{t-1}
B_1^{t}
B_1^{t+1}
B_0^{t-1}
B_0^{t}
B_0^{t+1}
I^{t-1}
I^{t}
I^{t+1}
α
β
Oldest Event
Older Event
Old Event
Objects
Observations

Figure 10.