Article

Habituation as optimal filtering

Samuel J. Gershman[1,2,*]

SUMMARY

Habituation, the reduction of responding to repetitive stimuli, is often conceptualized as a kind of attentional filter, amplifying salient signals at the expense of non-salient signals. No prior account has explicitly formalized filtering principles that can explain the major characteristics of habituation. In this paper, a simple probabilistic model is developed which permits analysis of the optimal filtering problem. This model exhibits the major characteristics of habituation, while also shedding light on other, relatively neglected, characteristics. These results demonstrate that habituation can be understood as a form of optimal filtering.

INTRODUCTION

Arguably the simplest and most ancient form of learning is *habituation*, the reduction of responding to repetitive stimuli.[1] Not limited to animals, it is exhibited by organisms as far flung as protozoa[2–5] and plants.[6,7] For example, the "sensitive plant" *Mimosa* closes its leaflets in response to mechanical stimulation, but with repeated stimulation the leaflets eventually reopen and cease to close. Similarly, the unicellular ciliate *Stentor* contracts in response to mechanical stimulation, but this contraction response attenuates with repeated stimulation. The ubiquity of habituation suggests that a universal principle may be at work. However, most theoretical treatments of habituation have focused on particular psychological[8–10] or neural[11–15] mechanisms, leaving the normative question—what is the logic of habituation?—unanswered.

One clue is the commonplace observation that defensive responses such as leaflet closing (in *Mimosa*) and cellular contraction (in *Stentor*) impede other activities such as photosynthesis and feeding. Thus, organisms should only respond defensively if the stimulus is really a threat. *Mimosa* reopens during rain in spite of the frequent mechanical stimulation. *Stentor*, adapted to living in turbulent ponds, returns to feeding in spite of mild disturbances to the water. Determining whether a stimulus indicates a threatening situation is fundamentally ambiguous, because the signals impinging on an organism's sensory apparatus may be similar for threats and non-threats. This imposes an inference problem: *what's out there*? In the language of signal processing, this corresponds to *filtering*—tracking a time-varying latent state variable based on noisy signals.[16] The interpretation of habituation as a form of attentional filtering has a venerable history (see the review by M. Ramaswami [15]). For now, we simply note that no previous model has formalized this idea in a way that explains all the major characteristics of habituation.

By making some assumptions about the structure of the environment and the sensors, we can derive a model of Bayes-optimal filtering. According to this model, the organism represents its state uncertainty in the form of a probability distribution, updated according to Bayes' rule. The distribution is used to compute the probability that the state is greater than a "danger" or "salience" threshold. The model assumes that the response probability or amplitude measured experimentally corresponds to the threshold exceedance probability. This simple model can explain the major characteristics of habituation (Box 1), as well as some subtle characteristics that have received less attention.

RESULTS

Model

We model an organism living in a time-varying environment (Figure 1). At time t, the organism collects sensory signal $x_t \in \mathbb{R}$, drawn from a probability distribution $p(x_t|\overline{x}_t)$ with expectation $\mathbb{E}[x_t] = \overline{x}_t$. We will refer to $\overline{x}_t$ as the *state* at time t. The state's time series is drawn from a probability distribution $p(\overline{x})$. The sensory signal is assumed to be measured on a logarithmic scale, consistent with psychophysical (i.e., the Weber-Fechner law[36]) and biophysical (i.e., fold-change detection[37,38]) principles. The model developed below does not intrinsically require logarithmic sensory transduction, but this is mathematically convenient since it does assume that the signals are real-valued.

The organism generates a response $y_t \in [0, 1]$ based on its inferences about the underlying state. In particular, we assume that the response corresponds to the probability that the state is greater than a threshold ψ:

$$y_t = p(\overline{x}_t > \psi|\mathbf{h}_t) = \int_{\overline{x}_t} \mathbb{I}[\overline{x}_t > \psi]p(\overline{x}_t|\mathbf{h}_t)d\overline{x}_t, \qquad \text{(Equation 1)}$$

[1]Department of Psychology and Center for Brain Science, Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA 02138, USA
[2]Lead contact
*Correspondence: gershman@fas.harvard.edu

Box 1. Major characteristics of habituation (adapted from [17])

1. Simple habituation: repeated application of a stimulus results in a progressive decrease in the response until an asymptotic level is reached.
2. Spontaneous recovery [18–21]: if the stimulus is withheld after response decrement, the response recovers at least partially over the observation time.
3. Potentiation [22,19,4]: after multiple series of stimulus repetitions and spontaneous recoveries, the response decrement becomes successively more rapid and/or more pronounced.
4. Frequency (rate) sensitivity [23,24,21]: other things being equal, more frequent stimulation results in faster and/or more pronounced response decrement, and more rapid spontaneous recovery.
5. Intensity sensitivity [19,25,23]: within a stimulus modality, the less intense the stimulus, the more rapid and/or more pronounced the behavioral response decrement. Very intense stimuli may yield no significant observable response decrement.
6. Stimulus specificity [26–28]: within the same stimulus modality, the response decrement shows some stimulus specificity. This characteristic distinguishes habituation from sensory adaptation/motor fatigue in neuroscience.
7. Dishabituation [29,19]: presentation of another (usually strong) stimulus results in the recovery of the habituated response.
8. Habituation of dishabituation [30–33,19]: upon repeated application of the dishabituating stimulus, the amount of dishabituation produced decreases.
9. "Below-zero" effects [19]: the effects of repeated stimulation may continue to accumulate even after the response has reached an asymptotic level. This effect of stimulation beyond asymptotic levels can alter subsequent behavior, for example, by delaying the onset of spontaneous recovery.
10. Long-term effects [34,35]: some stimulus repetition protocols may result in properties of the response decrement (e.g., more rapid rehabituation than baseline, smaller initial responses than baseline, smaller mean responses than baseline, less frequent responses than baseline) that last hours, days or weeks. We do not explicitly model these effects here, since the model does not commit to any particular timescale.

---

where $\mathbb{I}[\cdot] = 1$ if its argument is true (0 otherwise), and $p(\overline{x}_t|h_t)$ is the posterior over the state conditional on the signal history $h_t = \{x_{t'} : t' \leq t\}$, given by Bayes' rule:

$$p(\overline{x}_t|h_t) \propto p(h_t|\overline{x}_t)p(\overline{x}_t). \qquad \text{(Equation 2)}$$

Depending on the preparation under study, we can interpret the response y_t either as the probability of a binary action (e.g., the probability of contraction in response to an aversive mechanosensory signal) or as the amplitude of a continuous action (e.g., the amplitude of the flexor withdrawal in response to cutaneous electrical stimulation).

To obtain an analytically tractable model, we assume that the state is drawn from a Gaussian process[39] and then corrupted by additive Gaussian noise to generate the signal:

$$\overline{x} \sim \text{GP}(m, k) \qquad \text{(Equation 3)}$$

$$x_t \sim \mathcal{N}(\overline{x}_t, \alpha), \qquad \text{(Equation 4)}$$

where $m_t = \mathbb{E}[\overline{x}_t]$ is the mean function, $k_{t,t'} = \mathbb{E}[(\overline{x}_t - m_t)(\overline{x}_{t'} - m_{t'})]$ is the covariance function, and $\alpha > 0$ is the signal noise variance. We assume that the mean function is fixed to 0 for all t, which means that the organism tends to expect 0 signal amplitude in the absence of evidence to the contrary. The covariance function determines the timescale of habituation; we do not explicitly distinguish between "short-term" and "long-term" habituation (see Property 10 in Box 1) because the meaning of these terms (how long is long?) vary depend on the context and model organism.

Under these assumptions, we can derive a closed-form expression for Equation 1:

$$y_t = \Phi\left(\frac{\hat{x}_t - \psi}{\sigma_t}\right), \qquad \text{(Equation 5)}$$
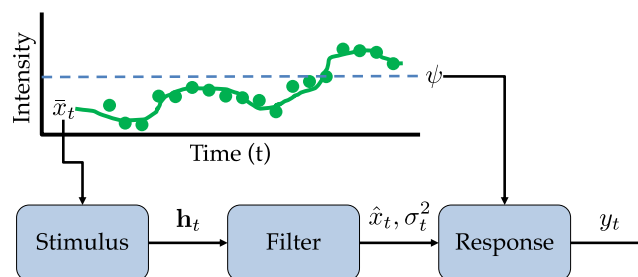

Figure 1. Structure of the model

A stimulus generates a sensory signal with intensity x_t (green circles), drawn from a Gaussian distribution with mean intensity $\overline{x}_t$ (green line). All the sensory signals up to time t are collected into the vector h_t, which is fed into the learning system (a Bayes-optimal filter). The output of the learning system is a probabilistic estimate of the mean intensity at time t, parametrized by a mean ($\hat{x}_t$) and variance (σ_t^2). Based on this estimate relative to a threshold ψ (dashed blue line), the organism generates a binary response with probability y_t.
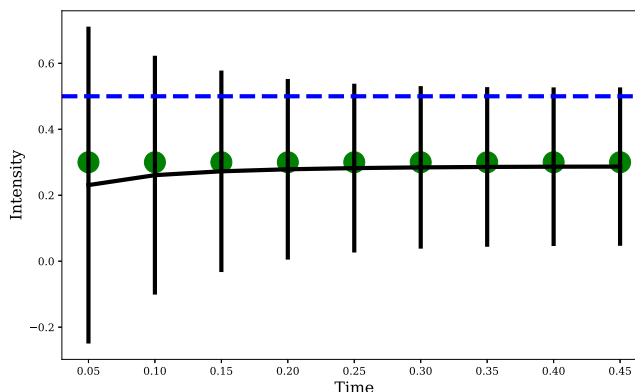
Figure 2. Illustration of the model

Green dots show the stimulus series and blue dashed line shows the response threshold (as in Figure 1). The black curve shows the posterior mean (with standard deviation error bars) just prior to the signal at each time point. In this illustration, the mean doesn't change very much over time, while the standard deviation shrinks gradually. This has the effect of decreasing the probability y_t that the signal mean $\bar{x}_t$ is above the threshold ψ, thus producing habituation.

with posterior mean and variance given by:

$$\hat{x}_t = \mathbf{k}_t^\top (\mathbf{K}+\alpha\mathbf{I})^{-1} \mathbf{h}_t \qquad \text{(Equation 6)}$$

$$\sigma_t^2 = k_{t,t} - \mathbf{k}_t^\top (\mathbf{K}+\alpha\mathbf{I})^{-1} \mathbf{k}_t, \qquad \text{(Equation 7)}$$

where $\mathbf{k}_t$ is the vector of covariances between t and all other time points, $\mathbf{K}$ is the matrix of covariances evaluated at all time points, and the signal history $\mathbf{h}_t$ is organized into a column vector. Note that the expressions for the posterior mean and variance hold for any choice of covariance function (including non-stationary and non-smooth ones). Figure 2 illustrates how the model works on an example habituation protocol.

For simplicity, the model above was presented for the single stimulus case. It can be straightforwardly generalized to settings with multiple stimuli by assuming that the covariance function is defined over time-stimulus pairs $\mathbf{z} = [t,\mathbf{s}]$, where $\mathbf{s} \in \mathbb{R}^D$ denotes a D-dimensional stimulus.

Further modeling details can be found in the STAR methods section.

Simulations

Let us begin with the basic phenomenon of habituation (Property 1 in Box 1): why does it happen? The filtering model asserts that habituation arises from the process of learning that the current state is non-threatening. Intuitively, a repeated signal, provided that its intensity is below the threshold ψ, will drive the mean $\hat{x}_t$ of the posterior distribution toward a sub-threshold value. In addition, the variance of the posterior distribution σ_t^2 will be driven toward 0 as evidence accumulates. Together, these dynamics imply that the response y_t should diminish with repetition—i.e., simple habituation.

Parametric studies of stimulus frequency (or rate) and intensity have revealed a more complex pattern. A common finding (e.g., [21,24]) is that habituation is stronger for high stimulus frequency. The filtering model accounts for this (Figure 3) by virtue of the fact that high frequency drives the posterior uncertainty down more aggressively, yielding a sharper reduction in responding. Critically, however, this depends on the intensity being below threshold. If it is above threshold, *sensitization* (an *increase* in responding with repetition) should be observed. This is consistent with the findings of Groves and Thompson,[23] who further demonstrated that high frequency stimuli produce stronger sensitization. In other words, frequency controls the slope of the learning curve, which goes in opposite directions depending on stimulus intensity. This is expected based on the different roles for the posterior mean and variance in determining response dynamics. Specifically, intensities below the response threshold will produce habituation, whereas intensities above the response threshold will produce sensitization. Because high frequency reduces the variance more quickly, this translates to a faster learning curve.

Most studies that manipulate parametric properties of the stimulus series confound learning (encoding and retention of information) and performance effects (expression of learned knowledge); see the review by R.M.Colwill et al.[40] for further discussion of this issue. A few studies have used a "common test" procedure, where the learning conditions are varied across groups of subjects but all subjects receive the same test trials. Here, we focus on a study reported by Davis,[24] who showed that responses to test stimuli were weaker for low frequency stimuli when measured using a common test. The filtering model captures this pattern (Figure 4).

This observation poses an apparent paradox, since the studies reviewed previously (which did not use the common test procedure) showed stronger habituation (i.e., weaker responses) for high frequency stimuli. The paradox can be resolved by recognizing that responding recovers more quickly for high frequency compared to low frequency stimuli. In terms of the model dynamics, the same reduction in posterior uncertainty that drives amplification of sensitivity to changes in the posterior mean also drives faster updating (similar to the effects of uncertainty on learning in other domains; e.g., [41–44]). These two roles can work against each other, thus explaining some of the non-monotonicities
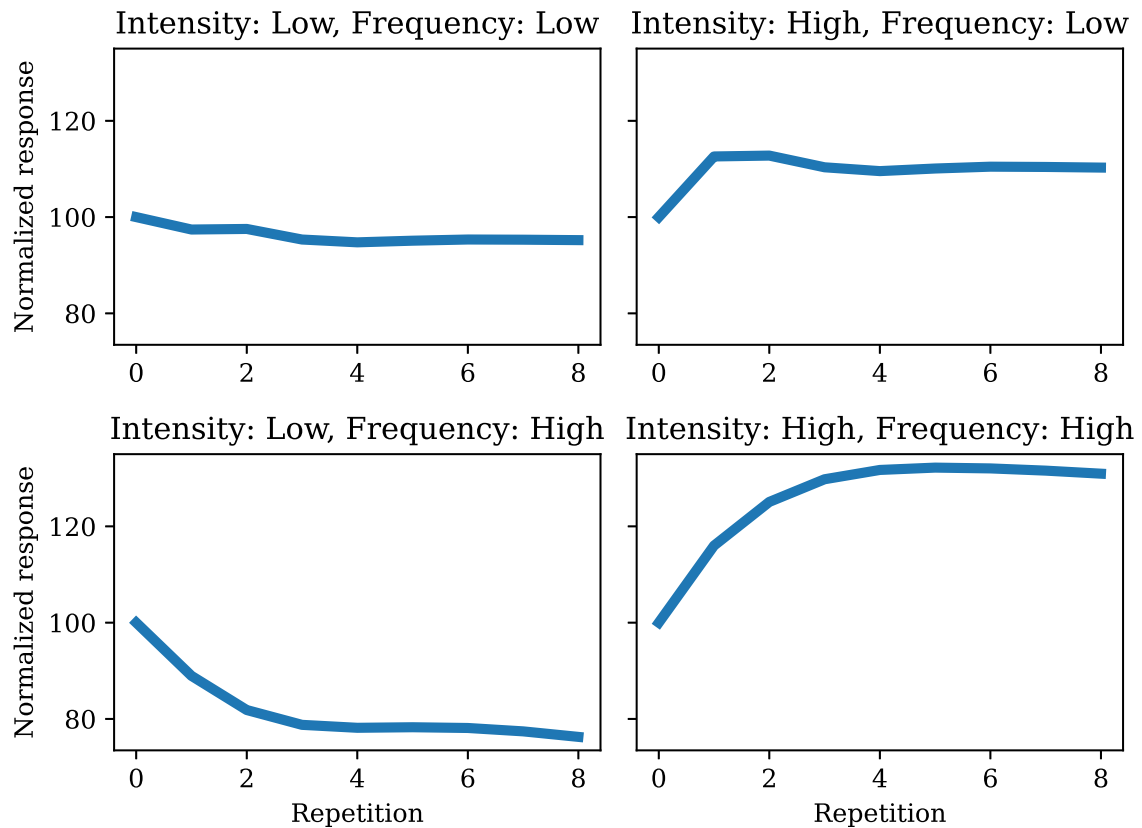
Figure 3. Stimulus frequency and intensity effects
When intensity is low, higher frequency stimulus presentations lead to stronger habituation. When intensity is high, habituation gives way to sensitization (increased responding as a function of stimulus repetition), with stronger sensitization for high frequency stimulus presentations.

in the learning curves shown in Figure 3. Which role dominates will depend on both the parameters of the model and the structure of the stimulus series.

Our discussion of the common test procedure highlights the fact that habituated responses recover spontaneously after a rest period.[18–21] Recovery is slower following extended habituation, even after responding has gone to zero or some asymptotic level (so-called "below-zero"
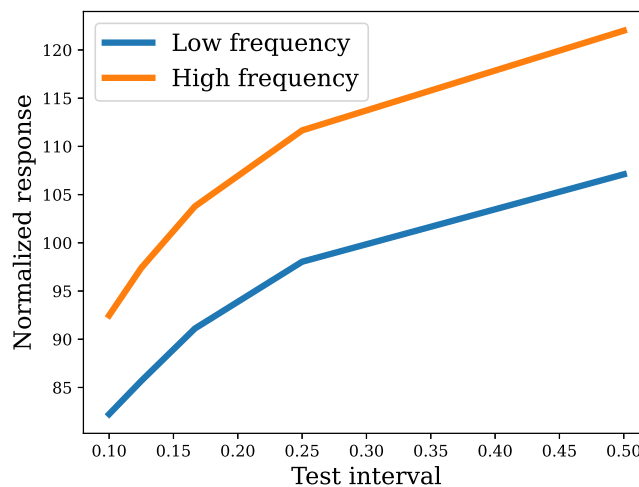

Figure 4. Common test procedure
In an initial habituation series (not shown here), the model received either low or high frequency stimulation. It was subsequently tested on multiple intermediate intervals in random order.
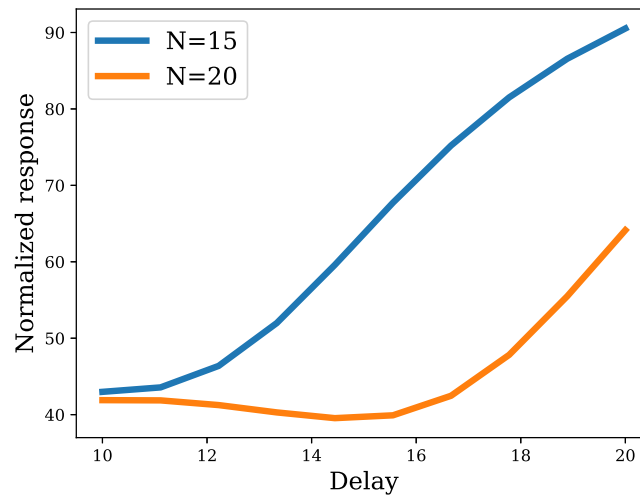
Figure 5. Spontaneous recovery

After a rest period in which no stimuli are presented, the response recovers from habituation. Recovery is slower when the initial stimulus series was longer. Note that this plot shows single tests at different delays, not repeated tests.

habituation; though see the studies by LE Gardner and D. Stephenson et al.[45,46] for divergent results). The filtering model captures this phenomenon (Figure 5): extended habituation, even after asymptotic responding, further decreases the posterior variance, such that it subsequently requires more time to return to baseline.

Rehabituation with a 2nd series of stimulus repetitions produces a learning curve that decreases more quickly than habituation to the 1st series, a phenomenon known as *potentiation*.[4,19,22] Importantly, the learning curve starts at roughly the same response level (provided there is adequate recovery time). This implies that potentiation is not simply a reflection of residual habituation from the 1st series—an "inactive memory"[47] must be present, encoding information about the stimulus history that is not immediately expressed in behavior. Figure 6 shows that the filtering model captures potentiation. The response returns to baseline after the recovery period, but the posterior variance is lower (due to having more data compared to the beginning of the 1st series), which sharpens the sensitivity of the response function to changes in the posterior mean.

So far we have focused on studies with a single stimulus. We now turn to studies with multiple stimuli. The most elementary observation is that habituation exhibits stimulus specificity: responding increases when tested on another stimulus.[26–28] This has sometimes been interpreted to reflect generalization,[19,48] in the sense that responding is intermediate between baseline and the habituated response to the familiar stimulus. Figure 7 shows that the filtering model captures graded stimulus specificity as a function of distance between the familiar and novel stimulus. This result is essentially baked into the structure of the covariance function (see STAR methods), which assumes that covariance drops off exponentially as a function of Euclidean distance in stimulus space. In other words,
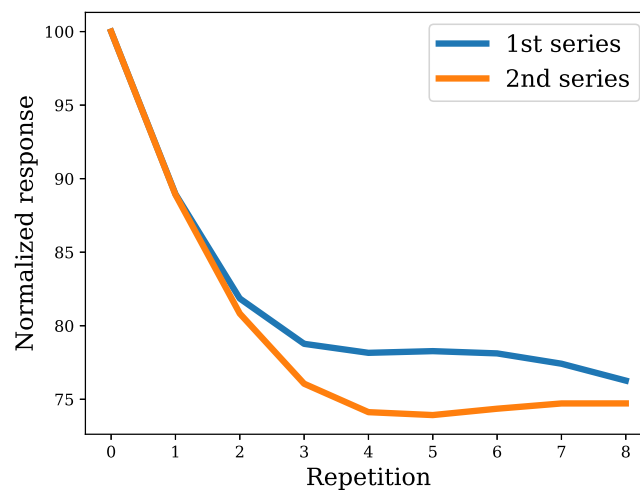

Figure 6. Potentiation

Rehabituation to a 2nd series is faster than initial habituation to the 1st series.
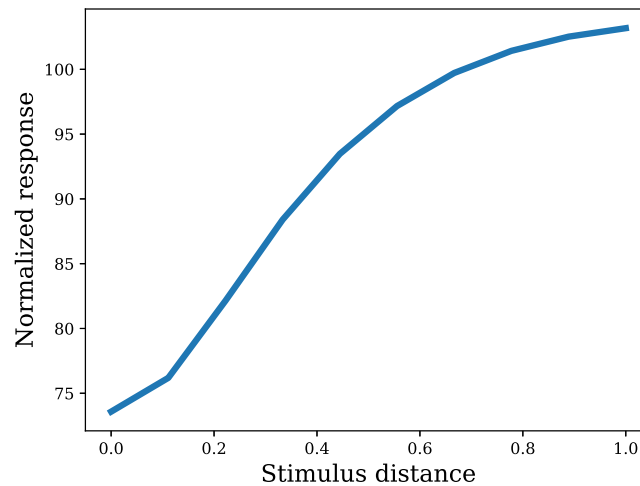
Figure 7. Stimulus specificity
Presentation of a novel stimulus produces a generalization decrement of habituation.

presenting a new stimulus moves the recent stimulus history into a new part of the stimulus space, producing a generalization decrement.

Another important multiple-stimulus phenomenon is *dishabituation*, the increase in responding to the familiar stimulus after presentation of a novel stimulus.[19,29] A typical (though not universal) finding is that dishabituation is stronger with high intensity novel stimuli. This arises in the model (Figure 8) because a strong stimulus is farther away in stimulus space from the familiar stimulus; the posterior probabilities of the two stimuli are coupled together, thus pulling the posterior mean for the familiar stimulus away from where it was at the end of habituation. Dishabituation itself habituates over the repeated presentations of the novel stimulus.[19,30–33] This is also predicted by the model due to the reduction of uncertainty in the new part of stimulus space, which mitigates the generalization decrement due to moving the recent stimulus history away from the habituated part of stimulus space.

An interesting prediction of this theory is that stimulus specificity and dishabituation should be positively correlated: dishabituation should be stronger to the extent that the original habituation is stimulus-specific. This prediction follows from the fact that both stimulus specificity and dishabituation arise in the model due to the same underlying property, namely generalization decrement due to a change in the posterior mean.

Alternative parametric assumptions

The flexibility of the Gaussian process model allows us to explore other modeling assumptions. We assumed that the covariance function is smooth over a characteristic timescale (see STAR methods). Alternatively, we could assume that the process is very non-smooth over this timescale (close to independent samples across time points), or even smoother (close to a constant mean across time points). What are the consequences of these alternative modeling assumptions?

We will use the frequency and intensity simulations as a case study (compare with Figure 3). When the covariance function is very non-smooth (length-scale parameter $\lambda = 0.001$), habituation and sensitization go away completely (Figure 9). This happens because there is no generalization across time points. When the covariance function is close to constant (length-scale parameter $\lambda = 100$), habituation for low intensity and sensitization for high intensity is still observed, but frequency dependence goes away (Figure 10). This happens because changes in frequency around the characteristic timescale of the experiment are effectively invisible to the model, which can only detect changes at much slower timescales.

In summary, capturing the key phenomena of habituation requires a choice of covariance function that is smooth at the characteristic timescale of the experimental protocol.

DISCUSSION

This paper has shown that the major characteristics of habituation can be accounted for by a model of habituation as optimal filtering. The key computations are: (i) tracking the posterior probability of a hidden state over time using Bayesian inference; (ii) mapping the posterior distribution to a response by computing the probability that the state exceeds a "threat" or "salience" threshold. These computations allow the organism to filter out unimportant signals and amplify important ones.

Related ideas have a long history in the study of habituation. Sokolov[49] was perhaps the first to suggest that habituation reflects a form of internal model, which he conceptualized as stimulus expectation (though he never formalized this idea). Response strength depends on a comparison between the stimulus and the expectation. In other words, response strength is proportional to prediction error. Variations of
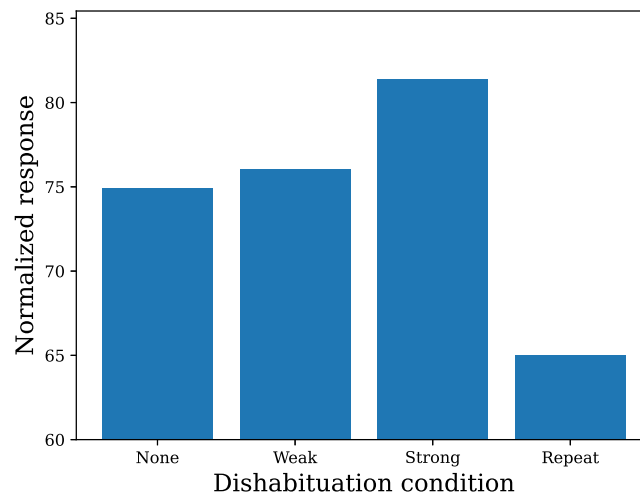
Figure 8. Dishabituation

Presentation of a novel stimulus increases responding to the familiar stimulus. This is particularly pronounced for high intensity (strong) compared to low intensity (weak) novel stimuli. Repetition of the novel stimulus causes habituation of dishabituation.

this "comparator" model were later explored by Konorski[50] and Wagner,[51] framed in terms of comparison with memory representations ("gnostic units") associatively activated by the stimulus history. Comparator theory resonates with modern computational ideas about predictive coding.[15] However, with the exception of Wagner's SOP model and Staddon's models (discussed further below), few versions of comparator theory have been described in sufficient formal detail (and implemented in computational models) to be evaluated systematically. The adequacy of Wagner's model as a descriptive account has been questioned by Mackintosh.[52] For example, Wagner's model posits
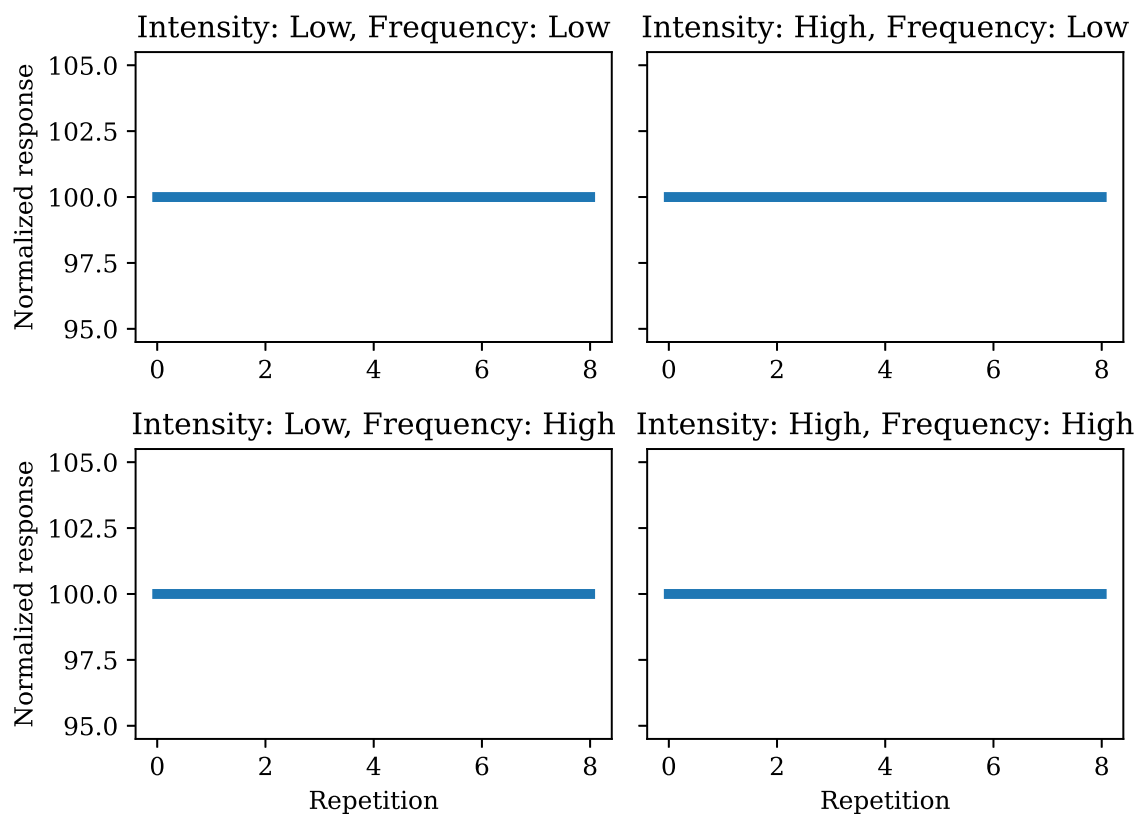

Figure 9. Stimulus frequency and intensity effects with a short length-scale
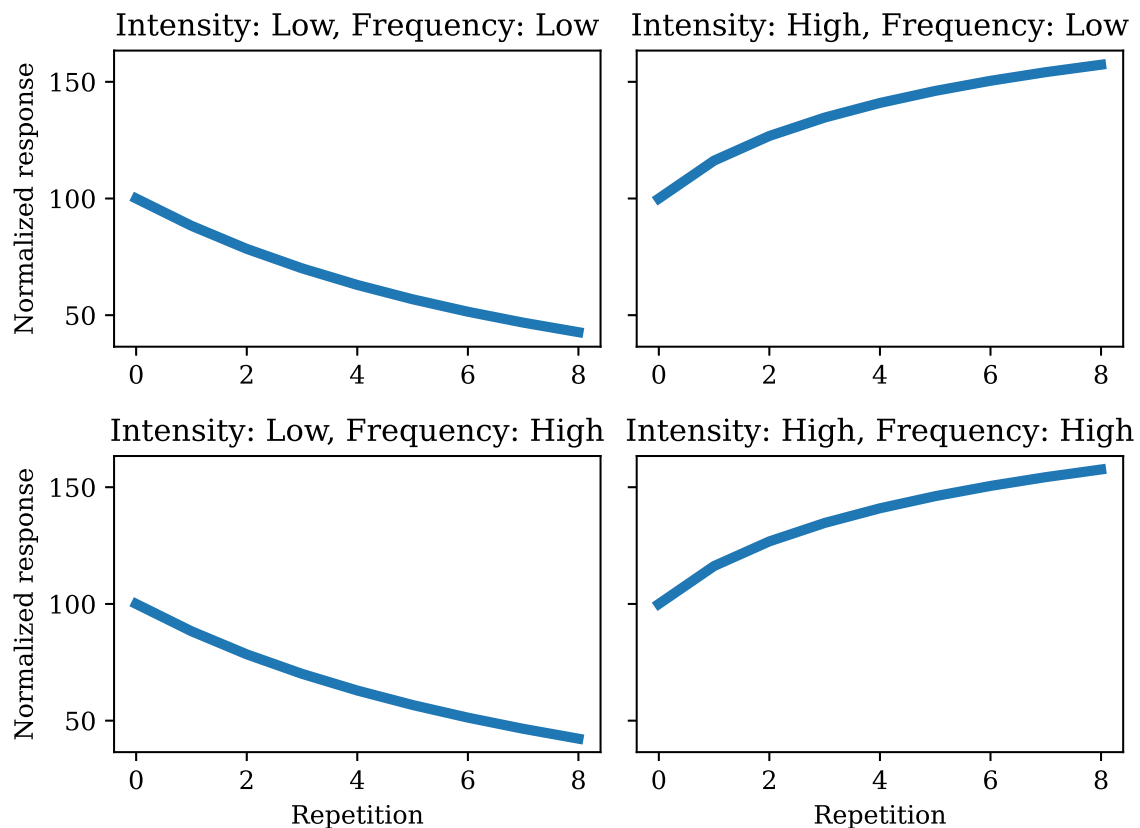
Compare with Figure 3.

Figure 10. Stimulus frequency and intensity effects with a long length-scale
Compare with Figure 3.

separate short-term and long-term habituation processes to explain phenomena like the frequency sensitivity, but Mackintosh points out that it could be explained by a single habituation process; indeed, the filtering model only appeals to a single process.

The filtering model shares with comparator models the idea that habituation reflects interpretation of signals through an internal model. It also hypothesizes a comparison operation, though not between the stimulus and the model expectation. Instead, the comparison is between an inferred state and a threshold. In this respect, it bears similarity to "optimal approach" models developed in behavioral ecology,[53,54] where the critical computation for organisms is the decision about whether an unfamiliar object is safe to approach.

Within the human developmental literature, normative models of habituation have been used to explain a quite different set of empirical phenomena (mainly infant looking time data). These models also invoke statistical inference, linking habituation to optimal information acquisition.[55–57]

None of these prior models have been comprehensively applied to all the major characteristics of habituation, as defined in Box 1. The most systematic attempt is the recent paper by Uribe-Bahamonde et al.,[10] which could explain many aspects of habituation but nonetheless did not offer an account of several phenomena addressed here (sensitization, frequency dependence of rehabituation). Other prominent models of habituation, such as the cascaded integrator models explored by Staddon et al.[8,9] and Bonzanni et al.,[58] also only explain a subset of the relevant data. For example, Staddon's models do an excellent job explaining the frequency dependence of habituation, but they do not account for sensitization. They also only apply to single stimuli, and hence cannot explain stimulus specificity or dishabituation.

Our goal is not to pit the filtering model against these other models, because they are conceptualized at different levels of analysis. Unlike these other models, the filtering model is deliberately agnostic about psychological and neural mechanisms; it adopts a level of abstraction that allows us to recognize the *function* of habituation. In principle, a number of the mechanistic ideas formalized by other models could be used to implement the computations required by optimal filtering. For example, it is possible to cast many Gaussian processes, either exactly or approximately, in a "state space" form governed by linear-Gaussian dynamics.[59] The optimal filtering solution is then given by a Kalman filter, which is essentially a kind of leaky integrator model similar to those studied by Staddon and colleagues. Much work remains to be done in bridging these levels of analysis.

Another way in which there is convergence across levels of abstraction concerns the use of memory, which plays a prominent role in Wagner's SOP model and Staddon's cascaded integrator models. These models assume that a memory trace of stimulus history is maintained over time and compared with incoming stimuli. The filtering model also maintains a memory trace of the stimulus history ($\mathbf{h}_t$), but this is

an infinite-capacity idealization which we don't assume is physically encoded. Because the covariance function determines the characteristic timescale over which memory needs to be maintained, very old memories can be safely discarded. One way to view the filtering model is that it stipulates what memory needs to be stored, given some assumptions about the environment. This provides a normative framework for constraining mechanistic models of habituation.

Adopting a high level of abstraction is important for understanding habituation, because evidence suggests that there may not be a single mechanism underlying all its manifestations. Studies of organisms with very simple nervous systems like *Aplysia* have suggested that homo-synaptic plasticity is the mechanism of habituation,[60] but this hypothesis does not generalize to organisms with more complex nervous systems, where multi-cellular circuit mechanisms come into play.[15,61] Even for *Aplysia*, the mechanisms for habituation appear to be much more complex than originally envisioned.[62] At the other extreme, unicellular organisms such as ciliates and ameobae exhibit habituation despite having no nervous system at all. Plants and even isolated cell lines[63] are in the same boat. Models like the one developed here may help us understand what computational principles these radically different systems have in common.

Limitations of the study

The downside of abstraction is that the model does not have much to say about the specific biological mechanisms underlying habituation in particular systems. For example, why does habituation appear to be a multicellular phenomenon in some systems but not others? Future work will need to ground the abstract model in particular biological implementations.

Another limitation is that the model assumes that the underlying latent variable generating sensory signals varies smoothly in time. However, other temporal structures might be more natural in certain environments. For example, some environments contain periodic or non-stationary temporal structures. Fortunately, these kinds of structures can be easily accommodated by the model through different choices of covariance function. We leave an exploration of this question to future work.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Covariance function
  - Model parameters
  - Response normalization

AUTHOR CONTRIBUTIONS

S.G. conceived and wrote the paper.

DECLARATION OF INTERESTS

The author has no competing interests to report.

REFERENCES

1. Thompson, R.F. (2009). Habituation: a history. Neurobiol. Learn. Mem. *92*, 127–134. https://doi.org/10.1016/j.nlm.2008.07.011.

2. Jennings, H.S. (1906). Behavior of the Lower Organisms (Columbia University Press). https://doi.org/10.5962/bhl.title.57732.

3. Boisseau, R.P., Vogel, D., and Dussutour, A. (2016). Habituation in non-neural organisms: evidence from slime moulds. Proc. Biol. Sci. *283*, 20160446. https://doi.org/10.1098/rspb.2016.0446.

4. Wood, D.C. (1969). Parametric studies of the response decrement produced by mechanical stimuli in the protozoan, Stentor coeruleus. J. Neurobiol. *1*, 345–360. https://doi.org/10.1002/neu.480010309.

5. Rajan, D., Makushok, T., Kalish, A., Acuna, L., Bonville, A., Correa Almanza, K., Garibay, B., Tang, E., Voss, M., Lin, A., et al. (2023). Single-cell analysis of habituation in stentor

coeruleus. Curr. Biol. *33*, 241–251.e4. https://doi.org/10.1016/j.cub.2022.11.010.

6. Pfeffer, W. (1873). In Physiologische Untersuchungen (Engelmann). https://doi.org/10.5962/bhl.title.3758.

7. Applewhite, P.B. (1972). Behavioral plasticity in the sensitive plant, mimosa. Behav. Biol. *7*, 47–53. https://doi.org/10.1016/s0091-6773(72)80187-1.

8. Staddon, J.E.R. (1993). On rate-sensitive habituation. Adapt. Behav. *1*, 421–436. https://doi.org/10.1177/105971239300100402.

9. Staddon, J.E., and Higa, J.J. (1994). Multiple time scales in simple habituation. Psychol. Rev. *103*, 720–733. https://doi.org/10.1037/e537272012-064.

10. Uribe-Bahamonde, Y.E., Becerra, S.A., Ponce, F.P., and Vogel, E.H. (2019). A quantitative account of the behavioral characteristics of habituation: The sometimes opponent processes model of stimulus processing. Front. Psychol. *10*, 504. https://doi.org/10.3389/fpsyg.2019.00504.

11. Stanley, J.C. (1976). Computer simulation of a model of habituation. Nature *261*, 146–148. https://doi.org/10.1038/261146a0.

12. Wang, D.L. (1993). A neural model of synaptic plasticity underlying short-term and long-term habituation. Adapt. Behav. *2*, 111–129. https://doi.org/10.1177/105971239300200201.

13. Prescott, S.A. (1998). Interactions between depression and facilitation within neural networks: updating the dual-process theory of plasticity. Learn. Mem. *5*, 446–466. https://doi.org/10.1101/lm.5.6.446.

14. Dragoi, V. (2002). A feedforward model of suppressive and facilitatory habituation effects. Biol. Cybern. *86*, 419–426. https://doi.org/10.1007/s00422-001-0306-x.

15. Ramaswami, M. (2014). Network plasticity in adaptive filtering and behavioral habituation. Neuron *82*, 1216–1229. https://doi.org/10.1016/j.neuron.2014.04.035.

16. Anderson, B.D.O., and Moore, J.B. (2012). Optimal Filtering (Courier Corporation). https://doi.org/10.1109/tsmc.1982.4308806.

17. Rankin, C.H., Abrams, T., Barry, R.J., Bhatnagar, S., Clayton, D.F., Colombo, J., Coppola, G., Geyer, M.A., Glanzman, D.L., Marsland, S., et al. (2009). Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. Neurobiol. Learn. Mem. *92*, 135–138. https://doi.org/10.1016/j.nlm.2008.09.012.

18. Ladd Prosser, C., and Hunter, W.S. (1936). The extinction of startle responses and spinal reflexes in the white rat. Am. J. Physiol. *117*, 609–618. https://doi.org/10.1037/e562352011-114.

19. Thompson, R.F., and Spencer, W.A. (1966). Habituation: a model phenomenon for the study of neuronal substrates of behavior. Psychol. Rev. *73*, 16–43. https://doi.org/10.1037/h0022681.

20. Waters, W.F., and McDonald, D.G. (1975). Stimulus and temporal variables in the "below-zero"' habituation of the orienting response. Psychophysiology *12*, 461–464. https://doi.org/10.1111/j.1469-8986.1975.tb00024.x.

21. Rankin, C.H., and Broster, B.S. (1992). Factors affecting habituation and recovery from habituation in the nematode Caenorhabditis elegans. Behav. Neurosci. *106*, 239–249. https://doi.org/10.1037/0735-7044.106.2.239.

22. Konorski, J. (1948). Conditioned Reflexes and Neuron Organization (Cambridge University Press). https://doi.org/10.1002/1097-4679(195001)6:1<107::aid-jclp2270060132>3.0.co;2-0.

23. Groves, P.M., Lee, D., and Thompson, R.F. (1969). Effects of stimulus frequency and intensity on habituation and sensitization in acute spinal cat. Physiol. Behav. *4*, 383–388. https://doi.org/10.1016/0031-9384(69)90194-2.

24. Davis, M. (1970). Effects of interstimulus interval length and variability on startle-response habituation in the rat. J. Comp. Physiol. Psychol. *72*, 177–192. https://doi.org/10.1037/h0029472.

25. Davis, M., and Wagner, A.R. (1968). Startle responsiveness after habituation to different intensities of tone. Psychon. Sci. *12*, 337–338. https://doi.org/10.3758/bf03331339.

26. File, S.E., and Russell, I.S. (1972). Specificity and savings of behavioural habituation over a series of intra-and inter-modal stimuli. Q. J. Exp. Psychol. *24*, 465–473. https://doi.org/10.1080/14640747208400306.

27. Wood, D.C. (1973). Stimulus specific habituation in a protozoan. Physiol. Behav. *11*, 349–354. https://doi.org/10.1016/0031-9384(73)90011-5.

28. Vogel, E.H., and Wagner, A.R. (2005). Stimulus specificity in the habituation of the startle response in the rat. Physiol. Behav. *86*, 516–525. https://doi.org/10.1016/j.physbeh.2005.08.042.

29. Pavlov, I.P. (1927). Conditioned Reflexes (Oxford University Press). https://doi.org/10.1136/bmj.2.4627.584.

30. Lehner, G.F.J. (1941). A study of the extinction of unconditioned reflexes. J. Exp. Psychol. *29*, 435–456. https://doi.org/10.1037/h0058702.

31. Hagbarth, K.E., and Kugelberg, E. (1958). Plasticity of the human abdominal skin reflex. Brain *81*, 305–319. https://doi.org/10.1093/brain/81.3.305.

32. Hagbarth, K.E., and Finer, B.L. (1963). The plasticity of human withdrawal reflexes to noxious skin stimuli in lower limbs. Prog. Brain Res. *1*, 65–81. https://doi.org/10.1016/s0079-6123(08)60589-0.

33. George, H.C., and Schwam, W.J. (1961). Effects of arousal reaction on nystagmus habituation in the cat. Am. J. Physiol. *200*, 29–33. https://doi.org/10.1037/e572322012-088.

34. Carew, T.J., Pinsker, H.M., and Kandel, E.R. (1972). Long-term habituation of a defensive withdrawal reflex in Aplysia. Science *175*, 451–454. https://doi.org/10.1126/science.175.4020.451.

35. Beck, C.D.O., and Rankin, C.H. (1997). Long-term habituation is produced by distributed training at long isis and not by massed training or short isis in caenorhabditis elegans. Anim. Learn. Behav. *25*, 446–457. https://doi.org/10.3758/bf03209851.

36. Gustav Theodor Fechner (1860). Element der Psychophysik (Breitkopf and Harterl). https://doi.org/10.1515/9783112579787-001.

37. Olsman, N., and Goentoro, L. (2016). Allosteric proteins as logarithmic sensors. Proc. Natl. Acad. Sci. USA *113*, E4423–E4430. https://doi.org/10.1073/pnas.1601791113.

38. Adler, M., and Alon, U. (2018). Fold-change detection in biological systems. Curr. Opin. Syst. Biol. *8*, 81–89. https://doi.org/10.1016/j.coisb.2017.12.005.

39. Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning (MIT Press). https://doi.org/10.7551/mitpress/3206.001.0001.

40. Colwill, R.M., Lattal, K.M., Whitlow, J.W., Jr., and Delamater, A.R. (2023). Habituation: It's not what you think it is. Behav. Processes *207*, 104845. https://doi.org/10.1016/j.beproc.2023.104845.

41. Courville, A.C., Daw, N.D., and Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. *10*, 294–300. https://doi.org/10.1016/j.tics.2006.05.004.

42. Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221. https://doi.org/10.1038/nn1954.

43. Nassar, M.R., Wilson, R.C., Heasly, B., and Gold, J.I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J. Neurosci. *30*, 12366–12378. https://doi.org/10.1523/jneurosci.0822-10.2010.

44. Gershman, S.J. (2015). A unifying probabilistic view of associative learning. PLoS Comput. Biol. *11*, e1004567. https://doi.org/10.1371/journal.pcbi.1004567.

45. Gardner, L.E. (1968). Retention and overhabituation of a dual-component response in Lumbricus terrestris. J. Comp. Physiol. Psychol. *66*, 315–318. https://doi.org/10.1037/h0026357.

46. Stephenson, D., and Siddle, D.A.T. (1976). Effects of "below-zero"' habituation on the electrodermal orienting response to a test stimulus. Psychophysiology *13*, 10–15. https://doi.org/10.1111/j.1469-8986.1976.tb03327.x.

47. Lewis, D.J. (1979). Psychobiology of active and inactive memory. Psychol. Bull. *86*, 1054–1083. https://doi.org/10.1037/0033-2909.86.5.1054.

48. Porter, J.M., Jr. (1938). Adaptation of the galvanic skin response. J. Exp. Psychol. *23*, 553–557. https://doi.org/10.1037/h0054236.

49. Sokolov, E.N. (1963). Higher nervous functions: The orienting reflex. Annu. Rev. Physiol. *25*, 545–580. https://doi.org/10.1146/annurev.ph.25.030163.002553.

50. Konorski, J. (1967). Integrative Activity of the Brain (University of Chicago Press). https://doi.org/10.1007/978-1-4899-6323-9_32.

51. Wagner, A.R. (1979). Habituation and memory. In Mechanisms of Learning and Motivation (Lawrence Erlbaum Associates), pp. 53–82. https://doi.org/10.1016/b978-0-12-549860-9.50010-9.

52. Mackintosh, N.J. (1987). Neurobiology, psychology and habituation. Behav. Res. Ther. *25*, 81–97. https://doi.org/10.1016/0005-7967(87)90079-9.

53. Blumstein, D.T. (2016). Habituation and sensitization: new thoughts about old ideas. Anim. Behav. *120*, 255–262. https://doi.org/10.1016/j.anbehav.2016.05.012.

54. Sherratt, T.N., Dewan, I., and Skelhorn, J. (2023). The optimal time to approach an unfamiliar object: a Bayesian model. Behav. Ecol. *34*, 840–849. https://doi.org/10.1093/beheco/arad032.

55. Kidd, C., Piantadosi, S.T., and Aslin, R.N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PLoS One *7*, e36399. https://doi.org/10.1371/journal.pone.0036399.

56. Ueda, K., Sekoguchi, T., and Yanagisawa, H. (2021). How predictability affects habituation

to novelty. PLoS One *16*, e0237278. https://doi.org/10.1101/2020.07.24.219253.

57. Cao, A., Raz, G., Saxe, R., and Frank, M.C. (2023). Habituation reflects optimal exploration over noisy perceptual samples. Top. Cogn. Sci. *15*, 290–302. https://doi.org/10.31234/osf.io/jb7qy.

58. Bonzanni, M., Rouleau, N., Levin, M., and Kaplan, D.L. (2019). On the generalization of habituation: how discrete biological systems respond to repetitive stimuli: a novel model of habituation that is independent of any biological system. Bioessays *41*, 1900028. https://doi.org/10.1002/bies.201900028.

59. Sarkka, S., Solin, A., and Hartikainen, J. (2013). Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at gaussian process regression through Kalman filtering. IEEE Signal Process. Mag. *30*, 51–61. https://doi.org/10.1109/msp.2013.2246292.

60. Vincent, C., Pinsker, H., Kupfermann, I., and Kandel, E.R. (1970). Neuronal mechanisms of habituation and dishabituation of the gill-withdrawal reflex in Aplysia. Science *167*, 1745–1748. https://doi.org/10.1126/science.167.3926.1745.

61. Fotowat, H., and Engert, F. (2023). Neural circuits underlying habituation of visually evoked escape behaviors in larval zebrafish. Elife *12*, e82916. https://doi.org/10.7554/elife.82916.sa2.

62. Glanzman, D.L. (2009). Habituation in Aplysia: the Cheshire cat of neurobiology. Neurobiol. Learn. Mem. *92*, 147–154. https://doi.org/10.1016/j.nlm.2009.03.005.

63. McFadden, P.N., and Koshland, D.E., Jr. (1990). Habituation in the single cell: diminished secretion of norepinephrine with repetitive depolarization of PC12 cells. Proc. Natl. Acad. Sci. USA *87*, 2031–2035. https://doi.org/10.1073/pnas.87.5.2031.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Simulation code for reproducing all the figures	The author	https://github.com/sjgershm/habituation

RESOURCE AVAILABILITY

Lead contact

Samuel Gershman, gershman@fas.harvard.edu.

Materials availability

No materials were generated as part of this study.

Data and code availability

- No data were generated or analyzed as part of this study.
- Simulation code for reproducing all the figures is available at https://github.com/sjgershm/habituation.
- Any additional information required to run the simulations reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Covariance function

While there is a wide range of possible covariance functions, we assume stationarity (i.e., the covariance structure depends only on $t - t'$) and smoothness (i.e., at least first-order differentiability). A conventional choice of covariance function satisfying these assumptions is the squared exponential:

$$k_{t,t'} = \exp\left(-\frac{|t - t'|^2}{2\lambda^2}\right), \tag{Equation 8}$$

where $\lambda > 0$ is the *length-scale*, which determines characteristic timescale over which fluctuations tend to occur. In the multiple stimulus case with $\mathbf{z} = [t, \mathbf{s}]$, the squared exponential covariance function becomes:

$$k_{\mathbf{z},\mathbf{z'}} = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z'}\|^2}{2\lambda^2}\right) \tag{Equation 9}$$

In the simulations reported above, we assume that the stimulus space is 1-dimensional; this assumption is sufficient to capture the relevant empirical phenomena. However, more generally it is plausible that the stimulus space is multi-dimensional.

Model parameters

All simulations are based on a fixed set of parameters: $\alpha = 0.3, \lambda = 1, \psi = 0.5$. In principle, the parameters could be adapted based on the stimulus history. However, this is unnecessary for the key results, so we leave the topic of parameter adaptation to future work.

Response normalization

Following a standard normalization procedure in the analysis of animal habituation data,[19] model responses were multiplied by 100 and divided by the response to an isolated stimulus. This allows us to interpret a normalized response of 100 as an unhabituated reference point.