Predicting graded dishabituation via perceptual stimulus embeddings in a rational learning model

Anonymous CogSci submission

Abstract

How do humans decide what to look at and when to stop looking? Many computational models have been proposed to address the mechanisms underlying such processes, but they are all limited in scope. In this paper, we present a version of the rational action, noisy choice for habituation (RANCH) model with an instantiated perceptual encoding hypothesis. We showed that the model can not only capture key looking time patterns such as habituation and dishabituation, but also make more fine grained predictions about dishabituation magnitude. We also investigated the generalizability and robustness of the RANCH parameters. In addition, we evaluated whether the assumptions in the model are critical. We found that RANCH's performance is relatively robust across parameters, and the parameters can be generalized to other tasks. However, the assumptions about the perceptual encoding process, the learning process and the decision process are critical for the model performance.

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

From birth, humans learn actively. Even before they can move on their own, infants can select information by deciding what to look at and when to stop looking (Haith, 1980; Raz & Saxe, 2020). Developmental psychologists have long leveraged this attentional decision-making to make inferences about the perceptual and cognitive abilities of infants by measuring how long infants look at certain stimuli (Aslin, 2007; Baillargeon, Spelke, & Wasserman, 1985; Fantz, 1963).

Two key phenomena are particularly critical for these inferences: habituation and dishabituation. Habituation refers to the decrease in looking time upon seeing the same or similar stimuli repeatedly; dishabituation refers to the increase in looking time following presentation of a novel stimulus after habituation. When participants dishabituate to a stimulus, they are thought to have distinguished between the novel stimulus and the stimulus they habituated to. Differences in dishabituation magnitude between different stimuli forms the basis of inferences about different levels of surprise that participants are experiencing. While habituation and dishabituation have been robustly documented, the underlying mechanisms of these looking time changes remain poorly understood. In this paper, we address this gap by presenting a rational model that provides principled predictions on the magnitude of dishabituation. Critically, this model can be applied generally to make predictions about looking time for arbitrary stimuli by using embeddings derived from a convolutional neural network.

The dominant framework in explaining changes in looking time proposes that habituation and dishabituation are driven by the amount of information to be encoded in the stimulus (Hunter & Ames, 1988). Observers look longer at a stimulus if the stimulus has a lot of unencoded information, and as exposure to the stimulus accumulates, less information is left unencoded, leading to shorter looking time. While this theory has been highly influential, the lack of formal details about what is meant by "encoding" opens the door for post-hoc interpretation of looking time measurements. A stimulus could be argued to be novel because it has distinct perceptual features, but it could also be familiar because of its conceptual characteristics. In part as a result of this interpretive ambiguity, concerns have been raised about looking time measurements should be the foundation for central claims in developmental psychology (Blumberg & Adolph, 2023; Haith, 1998; Paulus, 2022).

Computational models provide an important tool for formalizing the details of the habituation and dishabituation process. One set of models describe infants' looking behaviors with information-theoretic measures derived from ideal observer models (Kidd, Piantadosi, & Aslin, 2012; Francesco Poli, Ghilardi, Mars, Hinne, & Hunnius, 2023; F. Poli, Serino, Mars, & Hunnius, 2020). For example, Francesco Poli et al. (2023) developed a model that calculates the Kullback–Leibler (KL) divergence of each event infants saw in an experiment. This measure was shown to predict infants' looking time, with higher KL links to longer looking time.

While these models provided quantitative accounts of the habituation process, they do not model the infant's information sampling process directly. Instead, they describe trial-level correlations between these information-theoretic measures and measured looking times. In other words, these models do not explain how information theoretics measures are causally related to the sampling decision at any given moment. Furthermore, these prior models presuppose an abstracted representation of the stimuli, and do not instantiate a precise hypothesis about how visual encoding occurs during attentional decision-making. This limits the ability of these models to make principled predictions on new, unseen stimuli.

To address these issues, Rational Action, Noisy Choice for

Habituation (RANCH) was developed (Cao, Raz, Saxe, & Frank, 2023; Raz, Cao, Saxe, & Frank, 2023). RANCH described an agent's looking behavior as a rational exploration of noisy perceptual samples. This model construes the looking time paradigm as a series of binary decisions: to keep sampling from the current stimulus, or to move on to the next stimulus. This model makes sampling decisions based on the Expected Information Gain (EIG) of the perceptual samples, and therefore can be seen as a rational analysis of looking behavior (Anderson, 1991; Lieder & Griffiths, 2020; Oaksford & Chater, 1994).

Newer versions of RANCH also incorporate recent progress in convolutional neural networks, which have offered insights into how the visual system encodes objects (Doshi & Konkle, 2023; Hebart, Zheng, Pereira, & Baker, 2020; Yamins et al., 2014). The activations of these brain-inspired neural networks form embedding spaces, each of which can be seen as a quantitative hypothesis about the representations that humans form for visual stimuli (Schrimpf et al., 2020). For example, Lee (2022) projected the final layer of a ResNet50 into a "perceptually-aligned" space, by making its representations match dissimilarity matrices derived from human adult reaction times in a 2-AFC match-to-sample task. Passing new stimuli through this perceptual alignment yields a possible representation of how humans embed different visual stimuli in a low-dimensional space. Using this perceptually-aligned embedding space as a model of perceptual encoding, RANCH can generate fine-grained predictions about how long observers will look at stimuli sequence previously unseen by the model in the training.

Previously, Cao et al. (2023) and Raz et al. (2023) have shown that RANCH can successfully model habituation and dishabituation in adults and infants. Here, we test RANCH's ability to (1) predict responses in new data, and (2) predict a key phenomenon in qualitative accounts of habituation. For the first test, we tune its parameters to behavior on a generic habituation and dishabituation task, and test its performance on a previously unseen behavioral dataset. For the second test, we focus on RANCH's ability to reproduce a prediction from Hunter & Ames (1988) model of habituation and dishabituation: that observers' dishabituation magnitude should be related to the similarity between the habituated stimulus and the novel stimulus. The more dissimilar two stimuli are, the more one should dishabituate to the novel stimulus.

To conduct these tests, we first fit RANCH's parameters to a habituation-dishabituation experiment in which participants saw sequences of monsters which were either familiar or novel (dataset reported in Cao et al., 2023, Fig 1, left). Then, we use the best-fitting parameters to generate predictions for a new experiment that measure the subtle differences in dishabituation magnitude based on stimulus similarity (Fig 1, right). In this experiment, we systematically varied the similarity between habituation and dishabituation stimuli such that dishabituation stimuli differed in their pose angle, their number, their identity, or their animacy.

To preview our results, we show that RANCH can predict looking time responses in new data just by transferring model parameters fit from previous data, with marginal differences in performance compared to completely refitting to the new data. RANCH also captured the particular ordering of the dishabituation magnitude as a function of stimulus dissimilarity, thereby predicting a novel qualitative phenomenon (graded dishabituation) without ever being trained on it. Finally, we show that RANCH is relatively robust across parameter settings, but the assumptions about its perceptual representation, learning process, and the decision process are all critical to its performance.

Behavioral Experiment

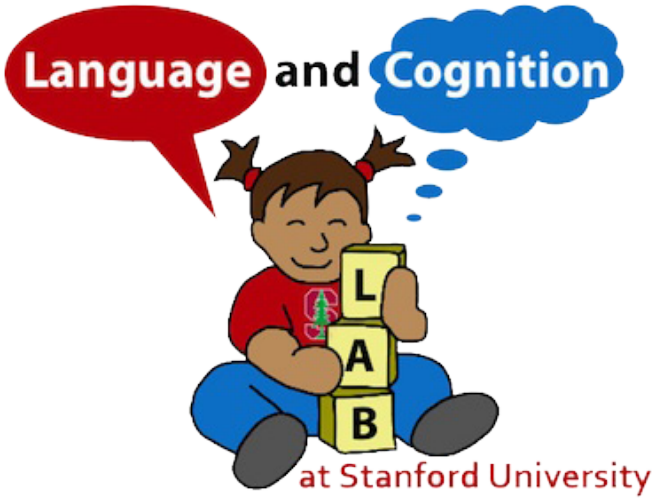

Figure 1: Experimental design of behavioral task: Adults watched sequences of stimuli consisting a familiarization to a stimulus (left), following by one of five trial types, which violates one property of the familiar stimulus: In the violation trials, participants saw either (1) the same exact stimulus (Background), (2) the same stimulus in a different orientation (Pose), (3) the same stimulus, but duplicated (or single if it was already duplicated; Number), (4) a different stimulus but in the same animacy class (Identity), or (5) a different stimulus of the opposite animacy (Animacy).

Methods

Stimuli All stimuli were created using images selected from Unity assets "Quirky Series - Animals" and "3D Prop Vegetables and Fruits". We added the same minor shaking animation to each image to increase interest. For each image from the animal set, we created a mirrored version to create an image with a different pose. Since most stimuli from the vegetable set were symmetrical, the images were slightly tilted before mirroring.

Procedure The experiment was a self-paced web-based experiment. Participants saw 24 blocks that consisted of either two, four, or six trials. On each trial, a schematic screen

would rise up to reveal a stimulus behind it. Participants pressed the spacebar to go on to the next trial after a minimum viewing time of 500 ms, triggering the schematic screen to drop and raise again to reveal the next stimulus.

Each participant saw eight types of repeating stimuli. The eight types of stimuli included all combinations of the three features that each included two levels: animacy (e.g. animate or inanimate), number (singleton or pair), and pose (facing left or right).

Eight blocks consisted of one stimulus being repeatedly presented throughout the block (i.e. background trials). The remaining sixteen blocks included two stimuli, including one that was repeatedly present and one that deviated from the trial. The deviant trial was always different from the repeating stimulus in one of the four dimensions: animacy violation, identity violation, number violation, and pose violation. The deviant trial always showed up in the last trial of the block. In the first three violation types, the feature would be switched to the previously unseen level (e.g. an inanimate deviant after an animate repeating stimulus, keeping the number and the pose the same). For identity violation, the participants would see a different, but within-category, exemplar from the repeating stimuli type. Among the sixteen blocks, the four violation types each appeared four times. After each block, participants performed a simple memory task as a filler task.

To control the distribution of background blocks and deviant blocks, we grouped the 24 blocks into four groups. Each group consisted of two background blocks, and one deviant block from each violation type. The order of blocks within each group was randomized.

Participants We recruited 550 adult participants on Prolific. Participants were excluded if either (1) the standard deviation of log-transformed of their reaction times on all trials is less than 0.15 (indicating key-smashing); (2) spent more than three absolute deviations above the median of the task completion time as reported by Prolific, or (3) provided the incorrect response to more than 20% of the memory task. After the participant-level exclusion, we also applied trial-level exclusion. A trial was excluded from final analysis if it is three absolute deviations away from the median in the log-transformed space across all participants. The final sample included 468 participants.

Results

The sample size and analysis plan are all pre-registered and can be found here. All analyses scripts are publicly available and can be found here [LINK].

We were primarily interested in (1) whether our experimental paradigm captured habituation and dishabituation and (2) whether the magnitude of dishabituation was influenced by the type of violation. We tested these two hypotheses in a linear mixed-effect model with maximal random effect structure that predicts log-transformed looking time with the following specification on the fixed effect: `log(total_rt) ~ trial_number + is_first_trial + (trial_number + is_first_trial) * stimulus_number + (trial_number + is_first_trial) * stimulus_pose +(trial_number + is_first_trial) * stimulus_animacy + (trial_number + is_first_trial) * violation_type + log(block_number)`. The violation_type has five levels, including the background trial and four types of violation. [1].

Through different parameterization of the model (i.e. setting different baseline levels for violation type), we found evidence for habituation and graded dishabituation. When background was treated as the baseline, there was a significant effect of trial number (β = -0.02, SE = 0, $p < .001$). Moreover, looking time to animacy violations was significantly longer than to number violations (β = 0.17, SE = 0.04, $p < .001$) and pose violations (β = 0.18, SE = 0.04, $p < .001$), so are identity violations (cf. number: β = 0.18, SE = 0.04, $p < .001$; cf. pose: β = 0.19, SE = 0.04, $p < .001$). But animacy violation was not different from the identity violation, nor was the number violation different from the pose violation (all p > 0.1)

We also pre-registered a qualitative prediction on the ordering of the dishabituation magnitude (i.e. category $>$ number $>$ identity $>$ pose). However, we did not find evidence consistent with this prediction. The qualitative ordering in our data was animacy (M = 2694.74), identity (M = 2489.2), pose (M = 2201.44) and number (M = 2117.63).

Model

We next wanted to ask whether looking time in our task can be seen as rational information acquisition, using a "rational analysis" approach previously described for a variety of aspects of human behavior (Dubey & Griffiths, 2020; Oaksford & Chater, 1994). Our goal was to develop a model which formalizes the entire process underlying looking time: from perception of a stimulus to deciding how long to look at it. To do so, our model has three separate components describing 1) the perceptual embeddings RANCH learns from 2) how RANCH learns a concept over this perceptual space and 3) how RANCH makes decisions about how long to sample from a stimulus based on its expected information gain. Here, we describe these three components in turn.

Perceptual representation To allow RANCH to operate on raw images, we used the perceptually-aligned embeddings obtained from a model presented recently by Lee (2022). We use these projections into a perceptually-aligned embedding space as a principled low-dimensional representation of stimuli, over which our learning model can form perceptual concepts. We used the first three principal components of the embedding space. 57.9% of the variance was captured by these first three PCs. A visualization of experimental stimuli in the embedding space can be seen in Figure 2A.

[1] In our preregistered model, we specified an interaction between trial_number and is_first_trial that is automatically removed in the final model.

Learning model RANCH's goal is to learn a concept in the perceptual embedding space described above, through noisy perceptual samples from a stimulus. The concept is parameterized by a 3D Gaussian (μ, σ), which represents beliefs about the location and variance of the concept in the embedding space. This concept (μ, σ) generates exemplars (y): exemplars of the concept. RANCH observes repeated noisy samples $(\bar{z})$ from each exemplar. For any sample (z) from an exemplar (y), the model expects the observation to get corrupted by zero-mean gaussian noise with standard deviation (ε). A plate diagram is shown in Figure 1B. We used a normal-inverse-gamma prior on the concept, the conjugate prior for a normal with unknown mean and variance, on the concept parameterized as μ_p,ν_p,α_p, β_p. Still, applying perceptual noise to y breaks the conjugate relation, so we computed approximate posteriors using grid approximation over (μ, σ) and (ε). This computationally expensive approximation was accomplished through a pytorch implementation and distributed GPU computation.

Decision model To decide whether to take an additional sample from the same stimulus, RANCH computes expected information gain (EIG) of the next sample. EIG is computed as the product of the posterior predictive probability of the next sample and the information gained conditioned on that next sample, by iterating through a grid of possible subsequent samples. RANCH then makes a softmax choice (with temperature = 1) between taking another sample and looking away. We assumed that participants expect a constant information gain from looking away (the "world EIG"). Therefore, as EIG from the stimulus drops below world EIG, it becomes increasingly likely that RANCH will look away.

Simulating the experiment To model the behavioral experiment, we first extracted the first three principal components of the embeddings of all the stimuli used in the experiment. We assembled the stimuli into sequences following the stimuli sequence participants saw in each block. For the block with deviating stimuli, we sampled deviant stimuli from corresponding violation categories. We sampled 23 stimuli pairs for each combination of violation type and deviant position. Since the model makes stochastic sampling decisions, we conducted 400 runs for each stimuli sequence for each parameter setting.

Alternative models To test the importance of each of RANCH's components for its performance, we defined three lesioned models, in which one key feature of the model was removed. First, to test the importance of the perceptual embeddings, we ran a version of RANCH in which the mappings from stimulus labels to embeddings were permuted, such that which embedding was associated with which violation type was randomized ("Random embeddings"). Second, we ran a version in which RANCH assumes that each perceptual sample in the learning process is noiseless, rather than corrupted by ε ("No noise").Third, we ran a version in which RANCH made decisions randomly rather than based on the learning model ("No learning").

Training Data We used the behavioral dataset reported in Cao et al. (2023) as the training data. In this prior experiment, 449 adults participated in a self-paced viewing of sequences of simple stimuli. The procedure was similar to the present experiment with two key differences: First, stimuli used in this previous experiment were animated monsters. Second, unlike the current study, the previous experiment only contained dishabituation stimuli that varied in the identity of the monster (comparable to the "identity violation" stimuli in the current experiment) There was no manipulation on the dissimilarity between the repeating stimulus and the deviant stimulus in a block.

Model Evaluation

Discussion

In this paper, we report on a novel experiment in which participants are familiarized to sequences of animations, and we measure habituation and their dishabituation to different types of violations of familiar stimuli. We find that adults' dishabituation is graded by the type of violation they see, and that the magnitude of dishabituation is predicted by a rational model which takes noisy samples from perceptual embeddings of the same stimuli that participants saw.

Our model, RANCH, through its use of perceptual embeddings, operates directly on raw images and therefore can generate predictions for previously unseen stimuli or even tasks. Making use of this property, we tested RANCH's performance on our graded dishabituation task using parameters fit to behavior on a completely different self-paced looking timefree-viewing task {raz2023modeling} another behavioral task.

Lesioning RANCH by removing key components caused its performance to drop significantly relative to its original implementation. This suggests that the aspects that we lesioned - noisy perception, connecting sampling to concept learning and a psychologically-plausible embedding space - are essential for explaining behavior in our task.

Our work is limited in a variety of ways: First, in the current work, we implemented a version of RANCH which takes a specific form for each of its components: the perceptual representation, the learning model and the linking hypothesis between learning and attentional sampling. However, RANCH's modular and interpretable structure allows researchers to adjust its components according to the population or task for which predictions are being generated. For example, the perceptual embedding space used in this paper was aligned to adult behavior @lee2022rapid, but infants likely represent visual objects differently from adults. Using perceptual representations based on visual input experienced by infants may provide a better fit to infant data {Zhuang et al. (2021), Orhan & Lake (2023)}. Similarly, task settings in which there was hierarchical structure to the stimuli sequences would call for more complicated learning models. In

previous work we have also explored the effect of varying the linking hypothesis by replacing the rational, but computationally expensive, expected information gain (EIG) with easier-to-compute information-theoretic quantities such as surprisal or KL-divergence (Cao et al. (2023), Raz et al. (2023)).

Second, while inspired by infant looking time research, our current work only has adult participants. Beyond encoding stimuli differently from infants, adults may conceptualize our task differently from infants, and experience different task demands. In particular, infants are quite sensitive to changes in the number of objects that are displayed (REFS), but in the current study, adults dishabituated to number violations as little as to pose violations, the sublest violation in our task. Conducting this study in infants may reveal qualitative differences between infant and adult dishabituation. Furthermore, given the interpretability of the model parameters we fit to behavior, conducting the same experiment with infants may lead to interpretable developmental differences in the model parameters, such as priors on perceptual noise and prior uncertainty about the mean and standard deviation of perceptual concepts.

Overall, our work presents a rational model, RANCH, which describes how humans decide how long to look at stimuli. Using a psychologically motivated visual encoding model allows RANCH to operate on raw images, and generate predictions for previously unseen stimuli or tasks. We think that the generality and interpretability of our model framework constitutes a significant step towards predictive modeling of adult, and eventually infant, looking time, thereby putting the field on firmer ground.

References

10 Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

Blumberg, M. S., & Adolph, K. E. (2023). Protracted development of motor cortex constrains rich interpretations of infant cognition. *Trends in Cognitive Sciences*.

Cao, A., Raz, G., Saxe, R., & Frank, M. C. (2023). Habituation reflects optimal exploration over noisy perceptual samples. *Topics in Cognitive Science*, *15*(2), 290–302.

Doshi, F. R., & Konkle, T. (2023). Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, *9*(25), eade8187.

Dubey, R., & Griffiths, T. L. (2020). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455.

Fantz, R. L. (1963). Pattern vision in newborn infants. *Science*, *140*(3564), 296–297.

Haith, M. M. (1980). *Rules that babies look by: The organization of newborn visual activity*. Lawrence Erlbaum Associates.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, *21*(2), 167–179.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*.

Kidd, C., Piantadosi, S. T., & 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*(5), e36399.

Lee, M. J. (2022). *Rapid visual object learning in humans is explainable by low-dimensional image representations* (PhD thesis). Massachusetts Institute of Technology.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Orhan, A. E., & Lake, B. M. (2023). What can generic neural networks learn from a child's visual experience? *arXiv Preprint arXiv:2305.15372*.

Paulus, M. (2022). Should infant psychology rely on the violation-of-expectation method? Not anymore. *Infant and Child Development*, *31*(1), e2306.

Poli, Francesco, Ghilardi, T., Mars, R. B., Hinne, M., & Hunnius, S. (2023). Eight-month-old infants meta-learn by downweighting irrelevant evidence. *Open Mind*, *7*, 141–155.

Poli, F., Serino, G., Mars, R., & Hunnius, S. (2020). Infants tailor their attention to maximize learning. *Science Advances*, *6*(39), eabb5053.

Raz, G., Cao, A., Saxe, R., & Frank, M. (2023). Modeling habituation in infants and adults using rational curiosity over perceptual embeddings. In *Intrinsically-motivated and open-ended learning workshop@ NeurIPS2023*.

Raz, G., & Saxe, R. (2020). Learning in infancy is active, endogenously motivated, and depends on the prefrontal cortices. *Annual Review of Developmental Psychology*, *2*, 247–268.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.