PHILOSOPHICAL PSYCHOLOGY
https://doi.org/10.1080/09515089.2020.1837358

Routledge
Taylor & Francis Group

ARTICLE

How do we know what babies know? The limits of inferring cognitive representations from visual fixation data

Isaac Davis

Philosophy, Carnegie Mellon University

ABSTRACT

Most infant cognitive studies use visual fixation time as the measure of interest. There are, however, some serious methodological and theoretical concerns regarding what these studies reveal about infant cognition and how their results ought to be interpreted. We propose a Bayesian modeling framework which helps address these concerns. This framework allows us to more precisely formulate hypotheses about infants' cognitive representations, formalize "linking hypotheses" that relate infants' visual fixation behavior with stimulus complexity, and better determine what questions a given experiment can and cannot answer.

ARTICLE HISTORY
Received 17 August 2019
Accepted 18 February 2020

KEYWORDS
Habituation; developmental psychology; rationalist modeling; bayesian modeling

1. Introduction

1.1. Studying infant cognition

Infant cognitive studies are always challenging, as there are relatively few ways to collect relevant data. Infants cannot tell us what they are thinking, and, for practical reasons, standard neuroimaging techniques are difficult to apply (Raschle et al., 2012). Thus, we are usually forced to rely on behavioral data alone. For this reason, the vast majority of infant cognitive studies use visual fixation time (i.e., the length of time that an infant visually attends to a stimulus) as the measure of interest. This leverages one of the few behaviors that infants of all ages regularly engage in: staring at things.

Interpreting visual fixation[1] data requires a *linking hypothesis* (Aslin, 2007; Teller, 1984), which relates fixation time to an underlying cognitive process. Most visual habituation paradigms rely on a linking hypothesis that relates an infant's fixation time to the novelty, complexity, or unexpectedness of the stimulus. The origins of this assumption are often attributed to a series of studies performed by Fantz (1961, 1964), which demonstrate that

CONTACT Isaac Davis isaacd@andrew.cmu.edu Department of Psychology Kirtland Hall 2 Hillhouse Ave New Haven, CT 06511

infants gradually shift their attention away from familiar stimuli and toward novel stimuli. *Habituation* refers to a steady decrease in fixation time as a stimulus is repeated, thereby becoming more familiar.

Visual habituation experiments have been used extensively to investigate how infants represent and understand the world. This includes object physics (e.g., Baillargeon, 1986; Leslie, 1984; Spelke et al., 1994), causation (e.g., Cohen & Oakes, 1993; Leslie & Keeble, 1987; Muentener & Carey, 2010; Oakes & Cohen, 1990), intentional behavior (e.g., Brandone & Wellman, 2009; Csibra & Gergely, 1998; Gergely et al., 1995; Phillips & Wellman, 2005; Woodward, 1998), and other agents' beliefs (e.g., Brooks & Meltzoff, 2002; Onishi & Baillargeon, 2005; Surian et al., 2007). Indeed, without the development of visual habituation experiments, we would know very little about infant cognition whatsoever.

There are, however, some concerns regarding what these experiments reveal and how their results ought to be interpreted. First, the conclusions drawn in a visual habituation experiment depend critically on the linking hypothesis, and it is crucial to precisely define and thoroughly test the linking hypothesis itself. To this end, several authors have called for an increased focus on modeling the habituation process itself (Aslin & Fiser, 2005; Colombo & Mitchell, 2009). Other authors have expressed more theoretical concerns about how hypotheses ought to be formulated and validated (Aslin, 2007; Oakes, 2010), and more practical concerns about the proper criteria for establishing habituation (Dannemiller, 1984; Thomas & Gilmore, 2004).

More generally, any study that asks what a subject knows or how a subject represents a stimulus faces a kind of underdetermination that can be difficult or even impossible to resolve. In particular, there are often multiple distinct accounts which result in identical behavior under the same experimental assumptions. To illustrate this, suppose that two young siblings, Ivan and Amos, go to an ice cream shop, and each one gets a cone. After leaving the shop, Ivan drops his cone and begins to cry. Now consider the following two descriptions of what transpires next:

(1) Amos sees that his brother is crying because he dropped his cone. He doesn't like it when his brother cries, so he gives Ivan his own cone, hoping that this will help Ivan to stop crying.
(2) Amos sees that his brother feels sad because he dropped his cone. He doesn't like it when his brother is sad, so he gives Ivan his own cone, hoping that this will help Ivan to feel better.

Both cases describe the same stimulus and response from an outside observer's perspective:

Ivan drops his cone and begins to cry, and then Amos gives Ivan his own cone. The reasoning underlying Amos' response in each scenario is very different, however. The first explanation is strictly behavioral: Amos reasons that Ivan's behavior (crying) is a direct reaction to a change in his environment (dropping his cone). The second explanation is mentalistic: Amos reasons that Ivan's behavior is a reaction to a hidden mental state (sadness) which is caused by dropping his cone. If Amos is sufficiently verbal, we might try to distinguish these two accounts by asking him to explain his reasoning; however, that is not an option when the subject is a preverbal infant. Thus, in order to avoid projecting on the infant a richer mental picture than is warranted by the data, it is important to precisely specify cognitive hypotheses and their corresponding behavioral predictions.

1.2. Contributions and overview

The main contribution of this paper is a computational modeling framework that allows us to do the following:

(1) Precisely specify hypotheses about how an infant represents a stimulus.
(2) Generate behavioral predictions from each hypothesis via a simulated version of habituation.
(3) Relate the design of a stimulus to the questions it could answer in a habituation experiment.

In Section 2, we start by reviewing visual habituation experiments in more detail, and identify the theoretical and methodological challenges which our framework addresses. We focus on the lack of formalizations of hypotheses about infants' cognitive representations, as well as the lack of formalizations of the linking assumptions through which experimental results are interpreted. In Section 3.1, we provide a conceptual overview of our framework and explain how it helps address these issues. Our framework takes a rationalist approach to cognitive modeling, and we illustrate a Bayesian implementation in Section 3.2. In Section 4, we validate our framework by replicating a seminal study on infants' understanding of intentional actions (Woodward, 1998). We demonstrate how our framework allows us to formalize the qualitative question posed in this study, how it enables a more precise interpretation of its results, and what further insights this interpretation suggests.

2. Background

2.1. Looking times and habituation experiments

The seminal demonstration of infant habituation is often attributed to a series of studies carried out by Robert Fantz in the early 1960s (1961, 1964). In the 1964 study, each infant was shown a sequence of stimuli paired side-by-side. In each trial, the stimulus on one side was held fixed, while the stimulus on the other side was novel. As the trials progressed, infants gradually shifted their visual attention away from the fixed stimulus and toward the novel one. *Habituation* refers to this progressive decrease in an infant's fixation time as a stimulus becomes more familiar.

A standard habituation experiment involves an initial habituation phase and a subsequent test or *dishabituation* phase. In the habituation phase, the infant is repeatedly shown the same stimulus over multiple trials. The experimenter records the infant's *fixation time* – that is, the duration for which the infant attends to the stimulus before looking away. As the stimulus is repeated, fixation time progressively decreases until a termination criterion is met. Typically, this occurs after a fixed number of trials, or once the fixation time reaches a sufficiently low threshold. It is critical to properly define and calibrate the termination criterion: if different infants habituate at different rates, they may be at very different stages of processing once the test phase begins. Infants who have not fully habituated may show fixation preferences which violate the linking hypothesis, such as preferring familiar stimuli over novel ones, and this can complicate our interpretation of the results (for an overview of termination criteria and associated methodological challenges, see Colombo & Mitchell, 2009).

In the test phase, the infant is shown two or more stimuli. Typically, the habituation stimulus depicts one or more salient features, and the test stimuli each differ with respect to one of these features. Dishabituation refers to a sharp increase in fixation time as the infant encounters a novel stimulus. If the infant dishabituates at a significantly higher rate with stimulus A than with stimulus B, this is interpreted as meaning that stimulus A appears more novel, unexpected, or complex to the infant, relative to the expectations formed during habituation. Using this reasoning, we can design experimental stimuli to determine which features are most integral in an infant's representation of a stimulus.

To better illustrate this, we briefly summarize a seminal habituation study of infants' understanding of goal-directed actions (Woodward, 1998). In this study, the habituation stimulus consisted of a stage with two platforms, each holding a visually distinct toy (see Figure 1). An actor stood to one side of the stage, so that only his or her arm was visible. In each habituation trial, the actor reached for and grasped one of the toys, targeting the same toy each time.

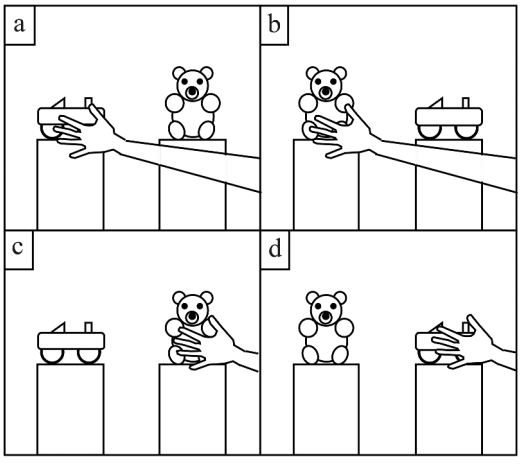

Figure 1. The four stimulus forms used in Woodward (1998). Each stimulus depicts an actor reaching for and grasping one of two visually distinct toys. Depending on the position of the toy relative to the actor (near or far), this results in one of two visually distinct reaching motions (long-reach or short-reach).

In the test phase, the toys were switched, and the infant was shown two test events. In the 'new goal"' event, the actor performed the same physical reaching motion as in the habituation stimulus, thereby grasping the opposite toy. In the 'new action' event, the actor reached for the opposite platform, thereby grasping the same toy as in the habituation stimulus. The underlying reasoning is that the stimulus depicts two salient features: *action* (the spatiotemporal profile of the actor's arm), and *outcome* (the toy which is grasped). If the infant encoded the habituation stimulus in terms of the

action feature, then, it was concluded, the test event which varies the action ('new action') appeared more novel to the infant, resulting in a higher rate of dishabituation. Conversely, if the infant encoded the habituation stimulus in terms of the outcome, then the 'new goal' test event appeared more novel. In this study, as well as in a subsequent panel of replications and control studies, 8- and 9-month-old infants consistently dishabituated to 'new goal' at higher rates, leading to the conclusion that outcomes are more salient than physical actions in infants' representations of reaching events.

The Woodward (1998) experiments illustrate nicely the core principles of habituation studies. Similar methods have been used extensively to explore how infants represent different aspects of the world, such as object physics (e.g., Baillargeon, 1986; Leslie, 1984; Spelke et al., 1994; Spelke et al., 1995), causation (e.g., Cohen & Oakes, 1993; Leslie & Keeble, 1987; Muentener & Carey, 2010; Oakes & Cohen, 1990), intentions (e.g., Brandone & Wellman, 2009; Csibra & Gergely, 1998; Gergely et al., 1995; Phillips & Wellman, 2005; Woodward, 1998), and beliefs (e.g., Brooks & Meltzoff, 2002; Onishi &

Baillargeon, 2005; Surian et al., 2007). However, some have expressed concerns about the design and interpretation of habituation experiments, calling for an increased focus on modeling the habituation process itself (Aslin, 2007; Aslin & Fiser, 2005; Colombo & Mitchell, 2009; Oakes, 2010; Thomas & Gilmore, 2004).

2.2. Theoretical models of habituation

Fixation data can only be interpreted through the lens of a *linking hypothesis*, which specifies the underlying cognitive or neurological processes driving visual habituation. There are multiple theoretical accounts of habituation, but most share the common view that fixation time reflects some combination of stimulus-driven attention, memory of previously encountered stimuli, and comparison between past and present stimuli (Kidd et al., 2012).

One of the most widely cited explanations for infant habituation is the Sokolov comparator model (Sokolov, 1963). This model is based on observations of an orienting reflex (OR): a response to nonthreatening, novel stimuli of moderate intensity (Colombo & Mitchell, 2009), which progressively decreases in magnitude as the stimulus is repeated. Sokolov theorized that as an organism repeatedly encounters a stimulus, the organism forms an internal representation or "cognitive schema" of that stimulus. Under this theory, the magnitude of the OR response – which, in the context of infant studies, is the infant's fixation time – is inversely proportional to the degree of similarity between the observed stimulus and the internal representation. Dual-process accounts of habituation (Groves & Thompson, 1970; Thompson & Spencer, 1966) add a separate "sensitization" process, which induces a transient spike in response strength at the onset of a new stimulus. More recent studies have revealed other factors which predict infant visual fixation and gaze shifts. These include saliency models (e.g., Mahdi et al., 2017), which predict fixation points by estimating the degree to which a stimulus component "stands out" from its background; contrast entropy models (e.g., Mahdi et al., 2015), which predict gaze shifts by estimating the visual information content of different fixation points; and saccadic models (e.g., Le Meur et al., 2017), which predict entire scan-paths or sequences of gaze-shifts.

While these studies explore how infants allocate their visual attention to different images or scenes, visual habituation experiments tend to involve a small set of repeated stimuli with similar visual features. Due to its explanatory power, relative conceptual simplicity, and plausible physiological underpinnings (Bernstein, 1979, 1981), the comparator theory (or comparator + sensitization) has remained the dominant account of infant visual habituation. Without formal models of certain aspects, however, these

theories provide only rough conceptual guidelines for designing habituation experiments, and they leave many critical questions unanswered.

2.3. Challenges and computational models

Here we review some often-raised concerns about infant habituation studies. These include practical concerns about experimental design and physiological concerns about the neural substrates underlying habituation. Several authors have proposed computational models to address these issues; however, our concerns relate more specifically to infants' "cognitive schema" and what we can infer about this schema from an infant's fixation behavior. Several authors have identified key challenges in making such inferences about infants' cognitive representations (e.g., Aslin, 2007; Oakes, 2010). We argue that the current literature lacks the computational models necessary to address these challenges.

The first concern is how an infant's intrinsic expectations – acquired through everyday experience prior to the experiment – affect the infant's performance in an experimental setting. The purpose of the habituation phase is to induce a novel expectation or to eliminate a prior expectation by repeatedly showing a "biasing stimulus." However, infants' intrinsic expectations can still "seep through" to the post-habituation phase, making it difficult to determine whether fixation times solely reflect the expectations formed during habituation (e.g., Quinn et al., 2002). It is therefore important to determine what preexisting expectations might influence an infant's performance and assess those expectations before interpreting experimental results.

Another concern is what we can conclude about internal representations when the stimuli differ with respect to inferred features as well as observable features. Learning about cognition from fixation behavior is much more straightforward when the only salient dimensions are easily perceptible (e.g., color or shape). In many experiments, however, some stimuli differ in terms of an inferred feature, which has an observable effect but is not directly observable itself. A clear example is the Woodward (1998) experiment, wherein one of the test stimuli differed from the habituation stimulus in terms of the actor's goal. Infants consistently dishabituated more strongly to the 'new goal' event: this clearly shows that infants can detect the relevant physical features, but it does not directly show whether the infant represents those physical differences in the same way that an adult would – namely, as observable consequences of the actor's goal state. It is therefore critical to precisely specify the hypotheses that is being tested in order to avoid drawing stronger conclusions – that is, projecting on an infant a richer mental representation – than warranted by the data.

As we cannot directly observe an infant's cognitive representations and infants cannot tell us about their cognitive representations, addressing these challenges requires a computational framework that allows us to (a) formulate precise hypotheses about infants' cognitive representations and the background knowledge constraining these representations, and (b) connect hypotheses to behavioral predictions in an experimental setting. Because we cannot "observe" how an infant represents a stimulus, we must rely on principled counterfactual claims regarding how an infant *would* behave if they *did* represent a stimulus in a certain way, which we can then compare against observed behavior.

The current literature on formal models of habituation is largely characterized by two approaches, neither of which is well-suited for these tasks. The *regression analysis* approach (e.g., Ashmead & Davis, 1996; Dannemiller, 1984; Thomas & Gilmore, 2004) is generally used to perform robustness checks on certain experimental practices, such as calibrating termination criteria. These models abstract from the underlying representations and estimate fixation time as a direct function of trial-time. The other approach uses connectionist and dynamic systems models to investigate the neurological substrates underlying habituation (e.g., Elman et al., 1998; Sirois & Mareschal, 2002, 2004). These models directly represent the low-level neurological mechanisms involved in habituation and are similarly unsuited for answering questions at the level of cognitive representations. Thus, there is a clear gap in the relevant literature at the cognitive level, where our questions of interest reside.

3. Formal framework

Our framework needs to serve three main functions:

(1) Provide a way to formalize claims about how an infant represents a stimulus.
(2) Provide a way to connect these claims with predictions about fixation behavior.
(3) Provide a way to model the process through which these representations are acquired during habituation.

To this end, we adopt a rationalist approach to cognitive modeling; we treat the infant as an observer with some background expectations, carrying out (approximately) rational inferences in response to observed evidence. This is an increasingly common approach in cognitive science and psychology, most often realized with the machinery of Bayesian inference (e.g., Griffiths et al., 2008), though there are many ways to implement a rationalist framework.[2] We first present the core concepts of our framework at an

abstract level before presenting a Bayesian realization of these concepts in greater detail.

3.1. Conceptual overview

3.1.1. Schema space

A schema space specifies the set of "cognitive schema" which we the experimenters consider to be plausible candidates for how the subject represents a stimulus[3]. We may think of each schema as an "intuitive theory" through which the subject interprets a class of stimuli. This "intuitive theory," in conjunction with the subject's inference mechanism, determines the internal representations that are acquired during habituation. At this stage, we are intentionally vague about the nature of these representations. Representations may be pictorial, with structural and spatial properties analogous to the stimuli they represent (e.g., Kosslyn & Pomerantz, 1977); they may be discursive, characterized as logical propositions in a language of thought (e.g., Pylyshyn, 1981); or they may be hybrid constructions involving both analogue and logical elements (e.g., Tye, 1984). Their content may be causal (e.g., Dretske, 1981), functional (e.g., Block, 1987), or probabilistic (e.g., Chater et al., 2006). All that we require at this level of abstraction is an agreed-upon notion of representation.

Given our experimental stimuli and representation system, we can further constrain the schema space by appealing to what the subject is likely to already know. If, for example, our representations are causal graphical models, and we expect the subjects to understand the forward direction of time, we may omit any representation in which past features are causally dependent on present features. We may also derive constraints from other levels of analysis (e.g., behavioral, physiological, etc.), though we focus on constraints derived from expectations regarding the subject's background knowledge.

The final step is to interpret a qualitative claim or question as a subset of these representations. The details of this identification depend on our formalization; in general, we interpret a qualitative claim about the subject's knowledge as a set of defining properties over representations, and then we identify the claim with the subset of representations which satisfy those properties. We illustrate this process using probabilistic generative models in Section 3.2.

3.1.2. Linking hypothesis

The next component is a formalization of the linking hypothesis: that is, our assumptions connecting data (i.e., fixation behavior) to hypothesis (i.e., cognitive schema). Recall that nearly every account of habituation involves some notion of stimulus novelty, complexity, or unexpectedness. Intuitively,

the more complex or unexpected a stimulus appears to the infant, relative to the expectations formed during habituation, the longer we expect the infant to fixate on that stimulus. However, few habituation studies employ precise measures of stimulus complexity, relying instead on rough, qualitative assessments.

A linking hypothesis takes the form of an input–output map. Inputs are pairs of (a) a cognitive schema and (b) one or more test stimuli. The outputs are predictions about fixation behavior. There are several ways to formalize this, each with different applications. A *quantitative* linking hypothesis outputs a numerical prediction of the infant's fixation time on a test stimulus. We can compute this using a formal notion of complexity (e.g., *information content* or a measure of logical coherence) or expectedness (e.g., posterior likelihood) of the stimulus. We can then generate a numerical prediction of fixation time as a function of this complexity measure. The exact details of this computation depend on (a) how we formalize the hypothesis space, and (b) the exact linking hypothesis we are formalizing.[4] The core principles remain the same, however: we compute an objective measure of stimulus complexity, given an inferred schema, and we use this measure to predict the infant's fixation time.

A more qualitative approach is to compute the *ratio* of complexities, using a complexity measure of choice, to predict which test stimulus the infant will fixate on for longer. If stimulus s_1 is significantly more complex than s_2, we can predict that the infant will fixate on s_1 for longer, without us having to explicitly predict the fixation time itself. This method is generally easier to apply, as it requires significantly less numerical calibration.

3.1.3. Habituation

Following our rationalist assumptions, we model habituation as the subject's process of inferring a representation of the habituation stimulus, given their background knowledge and inference mechanism. How we characterize this search space depends on our formal representation system and our assumptions about the subject's background knowledge. At a high level, each schema determines the space of individual representations that the subject may acquire during habituation. Given the subject's search space and inference mechanism, we model habituation as an approximately rational search for a representation which "best" accounts for the habituation stimulus. The representation that results from this inference reflects the expectations that the subject acquires during habituation. We then apply our functionalized linking hypothesis to the test stimuli and inferred representations in order to generate predictions about the subject's test-phase fixation behavior. By comparing these predictions against observed fixation

behavior, we can invert this reasoning to make inferences about the infant's cognitive representations from experimental data.

3.2. Bayesian implementation

To illustrate our framework in greater detail, we provide a Bayesian realization of these concepts and demonstrate how this realization serves the core functions outlined in Section 3.1. We choose the Bayesian approach partially because of its growing popularity in cognitive science, but also because its formal machinery is particularly well-suited for the tasks we have outlined.

3.2.1. Overview

The basic claim underlying the Bayesian approach to cognition is that much of our cognitive behavior can be interpreted as approximately rational probabilistic inference. At its core is an observer model: we treat our subject as a rational observer with some prior knowledge or expectations and ask how that observer would optimally respond to a stimulus. We can then invert this analysis to ask what prior knowledge, representations, or expectations lead to behavior consistent with what we observe in our subject.

An observer model consists of two components. The first is the observer's *hypothesis space*: this is the set of possible representations that the observer may consider for a class of stimuli. At an abstract level, each representation is a probability distribution over possible observations and enables probabilistic inference over the relevant domain. If, for example, the stimuli consist of two features $s = (f_1, f_2)$, then a representation in this domain is just a joint probability distribution over pairs $P(f_1, f_2)$. Using this representation, the observer can compute the likelihood of a particular observation and predict the value of one feature by observing the other.

The second component is a prior distribution over representations. This reflects the observer's prior degree of belief in each representation in the absence of evidence. When presented with some evidence E, the observer updates the degree to which they believe each representation according to Bayes' theorem:

$$P(t|E) = P(E|t)P(t)/P(E)$$

Here, $P(t|E)$ denotes a rational Bayesian observer's posterior degree of belief in representation t, given evidence E and prior beliefs $P(t)$. The denominator $P(E)$ is a normalizing term; in many cases, as in ours, we can ignore this term, as we are only interested in a ratio of posteriors, and the normalizing terms cancel out.

In recent years, Bayesian cognitive scientists have developed observer models to account for a wide variety of cognitive functions, including object perception (Kersten & Yuille, 2003), prediction of object trajectories (Weiss & Adelson, 1998), visual feature inference (Griffiths & Austerweil, 2009), and belief–desire inference (C. Baker et al., 2011; C. L. Baker et al., 2009). While these models are often technical and complex, they are all grounded in these basic components: a hypothesis space of probability distributions, a prior distribution over hypotheses, and learning and inference using Bayes' Theorem.

3.2.2. Formalizing the hypothesis space

Most often, a Bayesian observer's hypothesis space is defined using a generative model. A probabilistic generative model consists of a structural model and a parameter vector, which jointly define a probability distribution over stimulus values. The structural model consists of a variable set and a dependency relation among those variables which is typically but not necessarily causal. The variable set may include any observable features of the stimulus as well as latent features posited by the observer. The dependency relation determines an efficient way to parameterize the joint probability distribution induced by the model. If we let $X = (x_1, x_2, \ldots, x_n)$ denote the model's variable set and $\mathrm{par}(x)$ denote the set of variables in X on which x is directly dependent, then the joint distribution $\mathrm{P}(X)$ factors as follows:[5]

$$\mathrm{P}(X) = \mathrm{P}(x_1|\mathrm{par}(x_1)) * \mathrm{P}(x_2|\mathrm{par}(x_2)) * \ldots * \mathrm{P}(x_n|\mathrm{par}(x_n))$$

Each term $\mathrm{P}(x|\mathrm{par}(x))$ corresponds to a vector of parameters, and the set of all such terms constitutes the parameter space for the generative model. We can think of the structural model as the observer's "intuitive theory" for a domain, and each parameterization as a representation of a single stimulus in that domain.

To better illustrate such a hypothesis space, recall the stimulus forms used in the Woodward (1998) experiments (Figure 1). We can represent the salient features of these stimuli with three variables $\mathbf{s} = (e_0, \boldsymbol{a}, e_1)$:

(1) e_0 denotes the *initial state* of the stimulus, that is, the positions of the toys, the position of the actor's arm and hand, and so on.
(2) a denotes the *action*, that is, the physical motion profile of the actor's arm.
(3) e_1 denotes the *outcome*, which encodes the same information as e_0 after the action is performed.

A structural model in this domain contains the variables (e_0, a, e_1), as well as any latent variables posited by the observer. For the purpose of this illustration, we restrict the latent variables to at most one hidden feature g and one

bias parameter for each feature. The bias parameters encode the actor's preferences or attitudes corresponding to each feature, for example, the default likelihood of the actor choosing "long-reach" or "short-reach."

The hidden feature *g* corresponds to a *goal* and allows us to distinguish mentalistic theories of agent-actions from purely behavioral theories. Intuitively, a model without this feature explains the action in terms of contingencies between observable inputs (environment) and observable outputs (behavior), while a model with this feature explains the action in terms of a goal-driven agent.

Figure 2 illustrates four examples of structural models for the Woodward (1998) stimuli in this domain. Each corresponds to an "intuitive theory" of reaching actions; each allowable parameterization of a model corresponds to one particular representation that the observer may consider for a single instance of reaching. We can therefore define our schema space as the set of structural models over this variable set which are consistent with any constraints derived from the subject's presumed background knowledge.

With a formalized schema space, we can more precisely characterize the sort of qualitative claims about infants' representations that are tested in habituation experiments. The Woodward (1998) experiments, for example, explore whether infants encode reaching events in terms of the actor's arm-motion or the target object. We can interpret each of these possibilities in terms of a schema's structural model. In particular, if a model includes one or more trainable parameters corresponding to the action feature and no

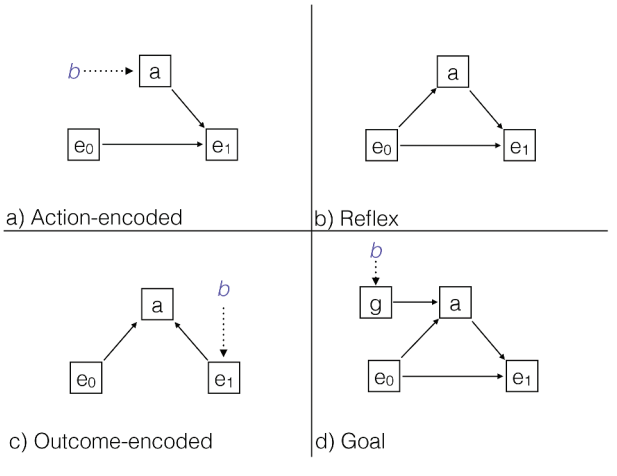

Figure 2. Examples of structural models for the Woodward (1998) stimuli. (a) Under this model, the actor has a bias for certain arm motions, and the outcome results from the actor's chosen motion. (b) Under this model, the action is a direct response to the initial configuration of the environment. (c) This denotes a "teleological" model (Csibra & Gergely, 1998), under which the actor has a bias for achieving a certain outcome and selects the action that achieves that outcome. (d) This is a "goal-model," which is similar to c but includes a latent goal feature distinct from the outcome.

trainable parameters corresponding to the goal or outcome feature (e.g., Figure 2(a,b)), then the infant would have to attend primarily to the arm-motion in order to be able to infer the appropriate representation. We can therefore identify these as the "motion-encoded" models. Similarly, if the structural model includes a trainable parameter corresponding to the goal or outcome (e.g., Figure 2(c,d)), then the infant would have to attend to these features in order to infer a representation. We can therefore identify these as the "outcome-encoded" models. We illustrate this identification more exhaustively in Section 4.

3.2.3. Observer models and fixation times

The common assumption underlying most habituation experiments is that an infant will fixate longer on the stimulus that is more complex or unexpected, given the expectations acquired during habituation. A formalized linking hypothesis therefore requires a precise measure of complexity or unexpectedness. The Bayesian framework provides a natural analogue to these notions in the likelihood function.

The likelihood term $P(s|t)$ denotes the probability of observing stimulus s given representation t. If this is very low, then an instance of s would be highly unexpected for an observer with representation t. We can leverage this interpretation of posterior likelihood to formalize both qualitative and quantitative linking hypotheses. If $P(s_1|t)/P(s_2|t)$ is significantly lower than 1, we predict that an observer with representation t will fixate on test stimulus s_1 for significantly longer than s_2, and vice versa if $P(s_1|t)/P(s_2|t)$ is significantly larger than 1. Kemp and Xu (2009) use a similar approach to connect a generative model of object trajectories with fixation predictions in object-perception experiments.

A more quantitative approach is to compute an objective measure of stimulus-complexity using posterior likelihood. Kidd et al. (2012) use this approach to test a particular linking hypothesis. They define a generative model of their experimental stimuli and equate the complexity of a stimulus with its *surprisal* $\log P(s|t)$ under the generative model. Surprisal is often used in statistics and information theory as a proxy for information content, and with a given model, it quantifies the memory cost of encoding a stimulus for an ideal observer. We can therefore use the surprisal of a stimulus under a given representation as a basis for quantitative predictions about an infant's fixation time.

Note that this points to two different ways one can use this reasoning to test hypotheses about habituation. The first is to apply a fixed linking hypothesis to a set of generative models, to determine which models induce predictions consistent with observed behavior. This is useful for testing hypotheses about how infants represent stimuli. The second is to assume a fixed generative model of a class of stimuli and generate looking-time

predictions under multiple complexity-dependent linking hypotheses, which is applicable for testing the linking hypothesis itself. For our present purposes, we focus on the first application.

3.2.4. Modeling habituation

In the Bayesian framework, the observer's habituation-phase inference process is realized as posterior inference using Bayes' theorem. Given a habituation stimulus s, let S_n denote a sequence of n stimuli identical to s. A rational Bayesian observer interprets this evidence using Bayes' theorem, updating their posterior degree of belief $P(t|S_n)$ in each representation. As n increases, we can identify the increasing familiarity of the habituation stimulus with the increasing posterior likelihood $P(s|S_n)$ under the subject's inferred distribution. This is obtained by integrating $P(s|t)$ over all values of t (i.e., all representations compatible with the observer's schema), weighted by $P(t|S_n)$.

If necessary, we can formulate explicit habituation criteria in terms of the posterior likelihood. If our criterion is reached after the n-th habituation trial, then we simulate the observer's test-phase performance by computing the likelihood of each test stimulus under the posterior distribution $P(t|S_n)$. This formalizes the notion that an infant interprets the test stimuli with respect to the representation inferred during habituation. By computing the likelihoods of the two test stimuli under this representation, we can apply our linking hypothesis to predict how the observer will allocate her visual attention.

Alternatively, we can abstract away from methodological concerns regarding habituation criteria by simulating the observer's performance in the test phase after each habituation trial. Unlike a real-world habituation experiment, we do not have to wait for the observer to reach a pre-defined threshold before applying this computation. We can therefore obtain simulated curves plotting the degree to which each test stimulus would be unexpected for an observer habituated to n habituation stimuli, for any value of n[6]. This allows us to generate simulated habituation curves as well as simulated plots of the observer's test-phase performance after each habituation trial. We show examples of simulated habituation and test-response curves in Appendix C.

3.3. Interpreting the Bayesian framework

Rationalist approaches have become increasingly common in cognitive science, most frequently using the formal machinery of Bayesian inference. There are, however, different perspectives regarding the proper application and interpretation of such models (for a more thorough review of these perspectives, see Chater et al., 2006; Griffiths et al., 2008;

Jones & Love, 2011; Lee, 2011). To understand these perspectives, it is important to understand the "three levels of analysis" typically identified in cognitive modeling (Marr, 1982). At the *computational* level, we characterize the abstract problem solved by some cognitive function, as well as the information involved in this problem. At the *algorithmic* level, we characterize the process through which a cognitive agent might solve that problem, including the representations involved and how those representations are manipulated. At the *implementation* level, we identify how these processes might be implemented in physiological substrates. Of course, these levels are not completely independent; knowledge at lower levels of analysis can provide constraints on hypotheses at higher levels.

Many rationalists in cognitive science restrict their interpretation to the computational level of analysis; they present a model as a useful way to characterize the cognitive problems being solved, rather than as a literal claim about the processes through which they are solved. Other papers adopt a more realist perspective, treating the model as a hypothesis about the cognitive representations involved in such processes. Our perspective in developing this framework is similar in spirit to the latter, though somewhat different in application. Much of the work in explaining human cognition with generative models focuses on "existence demonstrations"; that is, demonstrations of a certain generative model which, when appropriately parameterized, approximately replicates human performance in some cognitive task (e.g., categorizing novel objects, learning novel words, etc.). Our framework is similar in that we interpret the model as a candidate hypothesis about how infants represent stimuli. However, rather than focusing on individual plausible models, we characterize a broader space of possible models, and we interpret qualitative claims regarding infant cognition as being subsets of these models. This, we argue, provides a more precise way of specifying qualitative hypotheses about infant cognition, and it can assist us both in answering questions stemming from data and identifying tractable questions to ask.

Our interpretation of rationalist assumptions in this paper is similar to the use of utility functions in economics. In particular, economists are interested in the preferences people have and how they act on those preferences to make economic decisions. We cannot directly observe a preference, nor can we directly observe the cognitive processes underlying decision making, so any attempt to study this subject requires some assumptions regarding what an agent's behavior reveals about their preferences. The overwhelmingly common approach is to model economic decision-makers as rational agents optimizing a personal utility function. This assumption is flexible enough to capture nearly

any pattern of behavior, and it provides economists with a unified language for formulating hypotheses and generating predictions. We view the role of rationalist assumptions in cognitive science similarly: we are interested in infant cognitive representations and how they act on their representations, but we cannot observe these processes directly. We therefore assume that an infant's behavior reflects an approximately rational inference process about a representation of a relevant stimulus. This provides a unified language for formulating cognitive hypotheses and connecting those hypotheses to behavioral predictions.

4. Case study and simulations

In this section, we illustrate our framework by replicating the Woodward (1998) experiments. Note that the main purpose of this initial demonstration is to validate our framework against existing data and a qualitative interpretation of that data. We discuss other potential applications of the framework more extensively in the next section.

4.1. Setting up the simulations

The first step is constructing our schema space, which we briefly outlined in Section 3.2. This construction has three parts: first, we identify the potentially salient features of our experimental stimuli. We shall use the representation described in Section 3.2, which consists of three observable features, one hidden feature g, and one bias parameter β for each feature (as explained in Section 3.2.2). Second, we identify constraints on schemas based on what we can reasonably assume about the observer's background knowledge. Finally, we define our schema space as the set of all structural models consistent with these constraints. This construction leaves us with 14 possible structural models (see Appendix A for a full specification and explanation of constraints and models). These correspond to the 14 "schemas" that we consider to be plausible candidates for the infant's "intuitive theory" of reaching. Our replication will therefore involve 14 sets of simulations, one for each candidate schema.

The next step is to define our question of interest as a subset of this schema space. For this replication, our question of interest is the following: do infants encode reaching events in terms of arm-motion or outcome? To formalize this, we must identify subsets H_1 and H_2, which correspond to these two possibilities. This identification can be defined in terms of a model's dependency relation: an H_1 or "motion-encoded" model contains at least one of $e_0 \to a$ or $\beta_a \to a$ and cannot contain either $e_1 \to a$ or $g \to a$. That is, actions may directly depend on external

circumstances and/or the actor's internal biases, but not on outcomes or goals. An H_2 or "outcome-encoded" model contains at least one of the arrows $e_1 \rightarrow a$ or $g \rightarrow a$, meaning that actions directly depend on outcomes or goals.

To model habituation, we simulate an observer with schema T inferring a representation t of repeated stimulus s via Bayesian posterior inference. Each schema consists of one of the 14 structural models and its corresponding parameter space, and each schema corresponds to a row in our results table. For each simulation, we plot the observer's "habituation" rate by plotting, for each trial, the posterior likelihood of the next stimulus according to the observer's posterior distribution.

For the test phase, we apply a qualitative linking hypothesis: given the observer's posterior distribution $P(t|S_n)$ after the n-th habituation trial, we predict that the observer will fixate on test stimulus s_1 longer than s_2 if and only if the posterior likelihood of s_1 is significantly lower than that of s_2, that is, if $P(s_1 |S_n) \square P(s_2 |S_n))$. This connects a hypothesis about the observer's schema with a prediction about the observer's relative fixation times during testing. Note that if our goal were to replicate quantitative predictions or trends, we would need to perform a more rigorous parametric analysis and comparison against existing results. However, given the qualitative nature of the predictions we seek to replicate – that is, preference for one stimulus over another – little analysis is needed to perform the current validation of our framework. Additionally, we can abstract away from methodological concerns about termination criteria by simulating the test phase after each habituation trial, rather than waiting until a termination criterion is reached. We plot the observer's predicted test-phase performance after each of a large but fixed number of habituation trials (see Appendix B for simulation specifications and Appendix C for examples of habituation and test curves).

Table 1. Simulation results.

Model	Hypothesis	Preference
h_1^1	H_1	New action
h_1^2	H_1	None
h_1^3	H_1	None
h_1^4	H_1	None
h_2^1	H_2	New goal
h_2^2	H_2	None
h_2^3	H_2	None
h_2^4	H_2	New goal
h_2^5	H_2	None
h_2^6	H_2	New goal
h_2^7	H_2	New goal
h_2^8	H_2	New goal
h_2^9	H_2	None
h_2^{10}	H_2	None

4.2. Results and analysis

Table 1 shows the compiled results from our 14 simulations. A 'new goal' preference indicates that the posterior likelihood of 'new action' reaches at least 50% higher than the posterior likelihood of 'new goal', and vice versa for 'new action'. While we do not use explicitly coded termination criteria (as we can simulate test results after any number of habituation trials), the posterior likelihood consistently reached a 150% threshold of initial likelihood after 6–9 trials across all simulations. These results show only one model (h_1^1) which prefers 'new action', while five models prefer 'new goal'. Based on this table, we see that every model which develops a preference for the 'new goal' event belongs to H_2. This validates the experimenter's assumption that an observer who attends longer to the 'new goal' test event encodes the habituation event in terms of its outcome. Similarly, the only model which results in a preference for the 'new action' event belongs to H_1.

Beyond this validation, these simulations help address some of the concerns raised in Section 2.3. First, we noted the difficulty of drawing conclusions when experimental stimuli differ along inferred features such as a goal. Fixation data can reveal which stimuli appear more unexpected to the infant, but this does not directly tell us how the infant represents the stimulus. Replicating an experiment in this framework helps us determine what distinctions we can and cannot infer among candidate representations from a given stimulus design.

In this case, we can rule out an H_1 (motion-encoded) model for any infant who attends significantly longer to 'new goal'. However, among the models which develop a preference for 'new goal', one model (h_2^1) does not involve a latent 'goal' feature. Instead, this model identifies the physical outcome that follows the action as the main determinant of the action itself. This reflects a *teleological* model of actions (Csibra & Gergely, 1998), according to which the physical outcome explains or justifies (and temporally follows) the actor's movement. This is distinguished from a *causal-mentalistic* model, according to which a latent goal feature causes (and temporally precedes) the actor's movement. The results of these simulations demonstrate that the Woodward (1998) stimuli cannot be used to distinguish between these two possibilities. In order to make this distinction, one would need stimuli in which the actor's goal differs in some way from the physical outcome that follows (for example, see Brandone & Wellman, 2009). Thus, replicating the experiment in this framework helps us clarify the questions that a given stimulus can be used to answer, as well as the kind of stimuli needed to answer a given question.

A second challenge is distinguishing the expectations an infant acquires during habituation from the infant's intrinsic expectations. This

framework helps us separate these two kinds of expectations. To this end, note that our simulations were performed with uniform prior distributions over all parameters. Intuitively, this encodes an assumption that the observer has no prior expectations regarding the actor's biases. These results reflect solely the expectations acquired by the observer during habituation, and therefore, they illustrate the baseline expectations that an observer would form in the absence of any prior expectations. However, we can further determine the influence that an observer's prior beliefs would have by simply changing the prior distributions for each parameter. We can therefore use our framework to generate predictions of the form "an observer with model M and prior expectations P should prefer stimulus x over y after habituation to z." By holding M fixed and varying P, we can predict the observer's post-habituation preferences for different prior expectations. Vice versa, we can infer properties of the observer's prior expectations by holding some hypothesized M fixed and observing a subject's post-habituation preferences.

This basic example illustrates some of the main applications of our framework. By replicating habituation experiments in simulations like these, we can validate the reasoning underlying an experiment – more precisely, we can relate the design of a stimulus to the questions it can help answer, and we can better distinguish a subject's prior expectations from those acquired during habituation.

5. Conclusions and future work

Because there are so few ways to obtain data relevant to infant cognition, most of our knowledge comes from fixation-time experiments. There are, however, some serious concerns regarding their proper design and interpretation. As we have argued, there is a gap in the relevant literature at the cognitive level: there are regression-analysis models for assessing practical questions of experimental design, and connectionist models for exploring the neurological substrates underlying habituation. However, in order understand how an infant represents a stimulus and know what kind of inferences we can make about this representation from habituation experiments, we need an explicit model of the representations in question.

We believe that our framework helps fill this gap by serving three main functions. First, it helps us more precisely formulate hypotheses about infants' cognitive representations, allowing us to interpret qualitative questions as sets of generative models. Second, it allows us to formalize linking hypotheses that depend on stimulus complexity or unexpectedness and thereby connect hypotheses to behavioral predictions. Finally, we can integrate these components to replicate and

analyze fixation experiments via a simulated version of habituation. As we saw in our case study, this helps us determine what questions an experiment can answer, what inferences are justified from a particular body of data, and what prior expectations or knowledge may influence an infant's performance in the test-phase.

There are several different applications of this framework to be explored in future work. While the case-study in Section 4 applies the framework retroactively to existing data, we can also use the framework more constructively: that is, by formalizing a space of representations for given stimuli, and a linking hypothesis connecting each representation to a behavioral prediction, we can determine which representations can and cannot be distinguished through behavioral data before an experiment is performed. We can then invert this reasoning to design stimuli that will be most useful for answering a given question about infants' cognitive representations. In future work, we will explore the constructive potential of our framework for assisting with the design of experimental stimuli.

Additionally, while this paper focuses on evaluating hypotheses about infants' representations under fixed experimental assumptions, we can also use the framework to test the experimental assumptions themselves. We can, for example, fix a single observer model of a stimulus and generate behavioral predictions under multiple linking hypotheses. By comparing these predictions against human infants' responses to the same stimuli, we can assess which linking hypothesis best fits experimental data (similar to the approach in Kidd et al., 2012). Finally, we can perform robustness checks against variations in subjects' intrinsic expectations by simulating a population of infants with a distribution of different prior expectations and comparing habituation performance at individual and aggregate levels. This is similar in application to the regression analysis framework outlined by Thomas and Gilmore (2004). Thus, we believe our proposed framework can help to address many of the methodological and theoretical challenges inherent to studying infant cognition.

Acknowledgments

Open Access funding provided by the Qatar National Library.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Isaac Davis received his PhD in Logic, Computation, and Methodology from Carnegie Mellon University, and is currently a postdoctoral fellow in the Department of Psychology at Yale University.

References

Ashmead, D. H., & Davis, D. L. (1996). Measuring habituation in infants: An approach using regression analysis. *Child Development*, *67*(6), 2677–2690. https://doi.org/10.2307/1131746

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *Trends in Cognitive Sciences*, *9*(3), 92–98. https://doi.org/10.1016/j.tics.2005.01.003

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, *23*(1), 21–41. https://doi.org/10.1016/0010-0277(86)90052-1

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Bernstein, A. S. (1979). The orienting response as novelty and significance detector: Reply to O'Gorman. *Psychophysiology*, *16*(3), 263–273. https://doi.org/10.1111/j.1469-8986.1979.tb02989.x

Bernstein, A. S. (1981). The orienting response and stimulus significance: Further comments. *Biological Psychology*, *12*(2–3), 171–185. https://doi.org/10.1016/0301-0511(81)90010-7

Block, N. (1987). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy*, *10*(1), 615–678. https://doi.org/10.1111/j.1475-4975.1987.tb00558.x

Brandone, A. C., & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science*, *20*(1), 85–91. https://doi.org/10.1111/j.1467-9280.2008.02246.x

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*(6), 958. https://doi.org/10.1037/0012-1649.38.6.958

Chater, N., Tenenbaum, J. B., & Yuille, A. L. (2006). Probabilistic models of cogntion: Conceptual foundations. *Trends in Cognitive Science*, *10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, *29*(3), 421. https://doi.org/10.1037/0012-1649.29.3.421

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*(2), 225–234. https://doi.org/10.1016/j.nlm.2008.06.002

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, *1*(2), 255–259. https://doi.org/10.1111/1467-7687.00039

Dannemiller, J. L. (1984). Infant habituation criteria: I. A Monte Carlo study of the 50% decrement criterion. *Infant Behavior & Development*, *7*(2), 147–166. https://doi.org/10.1016/S0163-6383(84)80055-7

Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, *66*, S424–S435. https://doi.org/10.1086/392743

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Elman, J. L., Bates, E. A., & Johnson, M. H. (1998). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

Fantz, R. L. (1961). A method for studying depth perception in infants under six months of age. *The Psychological Record*, *11*(1), 27–32. https://doi.org/10.1007/BF03393383

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670. https://doi.org/10.1126/science.146.3644.668

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. https://doi.org/10.1016/0010-0277(95)00661-H

Goodman, N. D., & Stuhlmuller, A. (2018). (electronic). *The Design and Implementation of Probabilistic Programming Languages*. http://dippl.org

Griffiths, T. L., & Austerweil, J. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In *Advances in neural information processing systems* (pp. 97–104).

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge University Press.

Groves, P. M., & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review*, *77*(5), 419. https://doi.org/10.1037/h0029810

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*(1–3), 189–209. https://doi.org/10.1016/0010-0277(94)90028-0

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134

Kemp, C., & Xu, F. (2009). An ideal observer model of infant object perception. In *Advances in neural information processing systems* (pp. 825–832).

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158. https://doi.org/10.1016/S0959-4388(03)00042-4

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. https://doi.org/10.1371/journal.pone.0036399

Kosslyn, S. M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology*, *9*(1), 52–76. https://doi.org/10.1016/0010-0285(77)90004-4

Le Meur, O., Coutrot, A., Liu, Z., Rämä, P., Le Roch, A., & Helo, A. (2017). Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing*, *26*(10), 4777–4789. https://doi.org/10.1109/TIP.2017.2722238

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology*, *2*(1), 19–32. https://doi.org/10.1111/j.2044-835X.1984.tb00531.x

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288. https://doi.org/10.1016/S0010-0277(87)80006-9

Mahdi, A., Schlesinger, M., Amso, D., & Qin, J. (2015, August). Infants gaze pattern analyzing using contrast entropy minimization. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (pp. 106–111). IEEE.

Mahdi, A., Su, M., Schlesinger, M., & Qin, J. (2017). A comparison study of saliency models for fixation prediction on infants and adults. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 485–498. https://doi.org/10.1109/TCDS.2017.2696439

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*(2), 63–86. https://doi.org/10.1016/j.cogpsych.2010.02.001

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development*, *11*(3), 255–268. https://doi.org/10.1080/15248371003699977

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*(2), 193–207. https://doi.org/10.1016/0885-2014(90)90026-P

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Phillips, A. T., & Wellman, H. M. (2005). Infants' understanding of object-directed action. *Cognition*, *98*(2), 137–155. https://doi.org/10.1016/j.cognition.2004.11.005

Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review*, *88*(1), 16. https://doi.org/10.1037/0033-295X.88.1.16

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*(9), 1109–1121. https://doi.org/10.1068/p3331

Raschle, N., Zuk, J., Ortiz-Mantilla, S., Sliva, D. D., Franceschi, A., Grant, P. E., & Gaab, N. (2012). Pediatric neuroimaging in early childhood and infancy: Challenges and practical guidelines. *Annals of the New York Academy of Sciences*, *1252*, 43. https://doi.org/10.1111/j.1749-6632.2012.06457.x

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences*, *6*(7), 293–298. https://doi.org/10.1016/S1364-6613(02)01926-5

Sirois, S., & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience*, *16*(8), 1352–1362. https://doi.org/10.1162/0898929042304778

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3. https://doi.org/10.1037/0033-2909.119.1.3

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology*, *25*(1), 545–580. https://doi.org/10.1146/annurev.ph.25.030163.002553

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition*, *51*(2), 131–176. https://doi.org/10.1016/0010-0277(94)90013-2

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Teller, D. Y. (1984). Linking propositions. *Vision Research*, *24*(10), 1233–1246. https://doi.org/10.1016/0042-6989(84)90178-0

Thomas, H., & Gilmore, R. O. (2004). Habituation assessment in infancy. *Psychological Methods*, *9*(1), 70. https://doi.org/10.1037/1082-989X.9.1.70

Thompson, R. F., & Spencer, W. A. (1966). Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, *73*(1), 16. https://doi.org/10.1037/h0022681

Tye, M. (1984). The debate about mental imagery. *The Journal of Philosophy*, *81*(11), 678–691. https://doi.org/10.2307/2026175

Weiss, Y., & Adelson, E. H. (1998). *Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision. Technical report A.I. Memo No. 1624*, MIT.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Appendix A

Derivation of schema space

We define our schema space as the set of all directed graphical models over the variables $X = (e_0, a, e_1, g, \beta_a, \beta_g, \beta_{e1})$ consistent with the following constraints:

(1) Acyclicity.
(2) No directed arrows into e_0. This encodes the intuition that the initial state of the environment is fixed prior to the start of the trial and cannot be influenced by any feature of the trial itself.
(3) Bias parameters β_a, β_g, and β_{e1} may only have arrows into their respective features, which follows from their definition as bias parameters.
(4) If any of a, e_1, or g have no other parent in the structural model, they must have the corresponding bias parameter as a parent. Technically, bias parameters should be present for all features, even if they have another parent. However, these parameters would only be relevant for comparing the behavior of one actor against the behavior of another. As our current study involves observation of only one actor, we may omit the bias parameters for features with other parents.
(5) The induced probability distribution $P(e_0, a, e_1)$ must be consistent with the transition distribution $P(e_1|e_0, a)$. This parametric assumption encodes infants' knowledge of object physics and the physical principles of reaching (see Leslie, 1984).
(6) If a model includes the goal variable g, then g must be sufficiently strongly correlated with e_1. This parametric constraint encodes the knowledge that goals track outcomes (for infants who interpret the action in terms of a goal-driven agent).

Figure A1 illustrates three models which fail to meet these criteria for different reasons and are therefore omitted from our simulations. Figure A2 illustrates the 14 models which do meet these criteria and constitute the basis for our simulations. We omit the bias variables from these figures in order to save space.

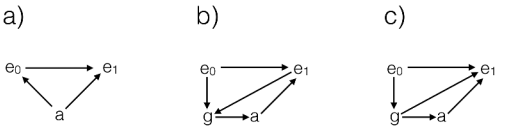

Figure A1. Three examples of structural models which are not consistent with our constraints. 3a) violates constraint #2, as it contains an arrow from a into e_0. 3b) violates constraint #1, as it contains a cycle. 3 c) violates the parametric constraint #5. In particular, this model allows the probability of e_1 to vary even when the values of e_0 and a are fixed, which violates the requirement that the joint distribution over (e_0, a, e_1) be consistent with the transition distribution P(e_1|a, e_0).

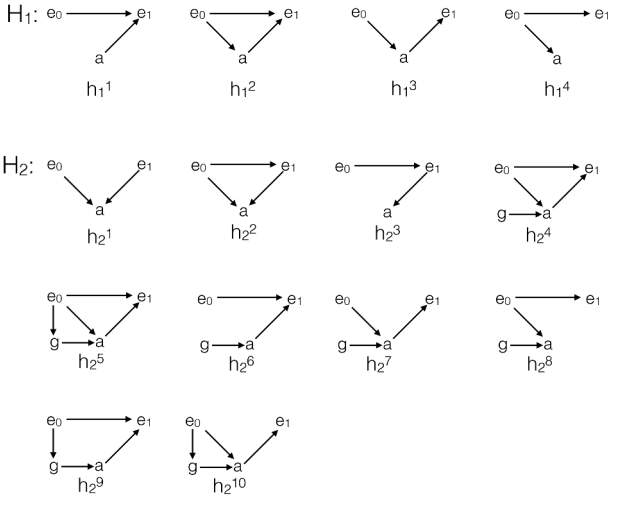

Figure A2. The 14 structural models consistent with our list of constraints.

Appendix B

Simulation specifications

All simulations were coded in WebPPL, a probabilistic programming language for generative models (Goodman & Stuhlmuller, 2018). For each model M, we compute the posterior likelihood of the habituation event s, given the observer's prior beliefs (i.e., model M and prior distributions) and n observations of the habituation event. We denote this likelihood as P(s|S_n, M). We equate the increase in posterior likelihood with the observer's familiarization to the habituation stimulus. For these simulations, all parameters are drawn from uniform priors.

In general, computing the posterior likelihood exactly may be intractable, so we approximate the posterior using a Markov Chain Monte Carlo (MCMC) sampling method, which generates a set of 10,000 samples that approximates the true posterior. This step is solely to improve the tractability of the simulations and does not reflect an assumption regarding the observer's cognitive processes. We perform these computations for n = 0, 1, …, 20. In addition to the habituation event likelihood, we compute,

for each n, the posterior likelihood of each test event s_1 ('new goal') and s_2 ('new action') under the distribution $P(s|S_n, M)$. The ratio of these likelihoods reflects whether the observer shows a preference for s_1, s_2, or neither. In particular, we assume that an observer who attends longer to test event s_1 does so because s_1 is significantly more unexpected than s_2. Thus, we report that the observer "prefers" s_1 if and only if the posterior ratio $P(s_2|S_n, M)/P(s_1|S_n, M)$ is sufficiently larger than 1. We use a threshold of 1.5 for our results table.

Appendix C

Simulated habituation and test curves

To better illustrate the outputs of each simulation, Figure C1 shows habituation and test-event response curves for the action, reflex, and outcome models (h_1^1, h_1^2, and h_2^1 in Appendix A1)

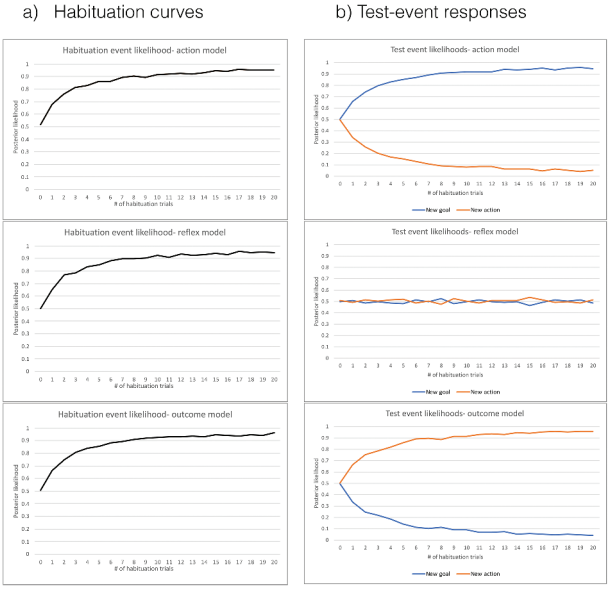

Figure C1. Simulated habituation and test-event response curves for action, reflex, and outcome models. Panel 5a illustrates the steady increase in posterior likelihood, corresponding to the observer's increasing familiarity with the habituation event. This occurs across all models. Panel 5b illustrates the posterior likelihood of both test events, given n observations of the habituation event. If the observer starts with an action model, the posterior likelihood of 'new action' drops significantly as habituation proceeds, while the posterior likelihood of 'new goal' increases significantly. Thus, as n increases, an action-model observer develops a strong preference for 'new action'. On the bottom of Panel 5b, the response graph illustrates that an outcome-observer would instead develop a strong preference for 'new goal', while the middle panel demonstrates that a reflex-observer would develop no preference for either event, both events being equally unexpected.

Notes

1. Throughout this paper, we use 'fixation' to refer to 'visual fixation' exclusively. Similarly, we use 'habituation' to refer to 'visual habituation' exclusively.
2. For example, mental models (Johnson-Laird, 1994), inference to the best explanation (Douven, 1999), and dual-process accounts of inductive and deductive reasoning (Sloman, 1996).
3. In a more general setting, we would call this the experimenter's 'hypothesis space'. However, this term has a more specific meaning in the context of Bayesian observer models. To avoid confusion, we use 'hypothesis space' to denote the standard Bayesian concept, and 'schema space' to denote the set of cognitive schema which the experiment considers as plausible hypotheses.
4. For example, under a dual-process linking hypothesis, the predicted fixation time would be a function of stimulus complexity *and* a sensitization rate, the latter being estimated through some other process.
5. This is the standard Markov property for causal graphical models.
6. If $n = 0$, this simulates the observer acting on prior expectations alone.