COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12777

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12777

Learning How to Generalize

Joseph L. Austerweil,[a] Sophia Sanborn,[b] Thomas L. Griffiths[c]

[a]*Department of Psychology, University of Wisconsin-Madison*
[b]*Department of Psychology, University of California*
[c]*Department of Psychology, Princeton University*

Received 20 February 2018; received in revised form 10 June 2019; accepted 13 June 2019

Abstract

Generalization is a fundamental problem solved by every cognitive system in essentially every domain. Although it is known that how people generalize varies in complex ways depending on the context or domain, it is an open question how people *learn* the appropriate way to generalize for a new context. To understand this capability, we cast the problem of learning how to generalize as a problem of learning the appropriate hypothesis space for generalization. We propose a normative mathematical framework for learning how to generalize by learning inductive biases for which properties are relevant for generalization in a domain from the statistical structure of features and concepts observed in that domain. More formally, the framework predicts that an ideal learner should learn to generalize by either taking the weighted average of the results of generalizing according to each hypothesis space, with weights given by how well each hypothesis space fits the previously observed concepts, or by using the most likely hypothesis space. We compare the predictions of this framework to human generalization behavior with three experiments in one perceptual (rectangles) and two conceptual (animals and numbers) domains. Across all three studies we find support for the framework's predictions, including individual-level support for averaging in the third study.

Keywords: Generalization; Inductive inference; Bayesian modeling; Category learning

1. Introduction

Almost every two objects, events, or situations that we encounter are unique. Despite this fact, when people learn that one stimulus has a property, they reliably and systematically believe certain other stimuli have that property and others do not (Shepard, 1987). For example, if you learn that a dark, large circle is a *gnarble*, how likely is a dark, slightly smaller circle or a dark very small circle to be a *gnarble*? This is the problem of

Correspondence should be sent to Joseph L. Austerweil, Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson St, Madison WI, 53706. E-mail: austerweil@wisc.edu

generalization, which is pervasive across psychology. It occurs in many forms throughout psychology from inductive reasoning (Kemp & Tenenbaum, 2009) to concept learning (Tenenbaum & Griffiths, 2001) to word learning (Xu & Tenenbaum, 2007). How do people generalize a property from a group of stimuli observed to have the property to other stimuli?

One of the most celebrated results of cognitive psychology provides a seemingly simple answer to this question: We generalize a property from one stimulus to another stimulus when the two stimuli are close in an appropriate psychological space (Shepard, 1987). Further, this is equivalent to performing Bayesian inference over a hypothesis space of properties, where each hypothesis is a candidate set of stimuli that have the property (Shepard, 1987; Tenenbaum & Griffiths, 2001). Defining more complex hypothesis spaces has led to empirically successful extensions into more complex domains, such as discrete-valued stimuli (Russell, 1986; Shepard, 1989) and stimuli with richer structure (e.g., integers; Tenenbaum & Griffiths, 2001), where there is no simple method for formulating a psychological space.

However, the Bayesian generalization framework is only successful when the stimuli are represented in an appropriate psychological space (Shepard, 1987) or equivalently, using a psychologically valid hypothesis space (Austerweil & Griffiths, 2010) with appropriate sampling assumptions (Ransom, Hendrickson, Perfors, & Navarro, 2018). This equivalence between psychological space and a space of hypothesized properties reflects the fact that any set of stimuli relate to each other along a multitude of dimensions, and generalization patterns will differ dramatically depending on which features are used to represent these items (Austerweil & Griffiths, 2011, 2013). Typically, psychologists infer the psychological spaces of participants by performing multidimensional scaling or hierarchical clustering on participants' similarity judgments (Nosofsky, 1986; Shepard, 1980; Shepard & Arabie, 1979; Xu & Tenenbaum, 2007). However, the brain cannot use these methods to form its own representations of stimuli (what similarity judgments would it use?). As shown in previous work (Shepard, 1987; Tenenbaum & Griffiths, 2001), the Bayesian generalization framework provides a normative answer to the question of how people should extend a property, given a pre-defined hypothesis space. However, it leaves open the question of how people determine the appropriate hypothesis space for a domain —a problem the brain must solve. Thus, we are left with a new question: Which hypothesis space should and do people use to generalize properties over novel stimuli?

Previous empirical work has demonstrated that generalization patterns are highly sensitive to context—people generalize a property depending on complex interactions between how stimuli are categorized in the domain and the property type. People selectively attend to dimensions diagnostic for categorization (Aha & Goldstone, 1992; Kruschke, 1992; Nosofsky, 1986). People also project properties from one category member to other category members depending on the heterogeneity of members of the categories (Gelman, 1988; Gelman & Markman, 1986; Nisbett, Krantz, Jepson, & Kunda, 1983). Additionally, how people generalize a property can depend intimately on the interaction between the type of property and the domain (Heit & Rubinstein, 1994; Medin, Coley, Storms, & Hayes, 2003; Nisbett et al., 1983; Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008).

Previous computational work has shown that this domain-dependent generalization behavior can be captured by performing Bayesian inference with a different hypothesis space for each domain (Kemp & Tenenbaum, 2009; Shafto et al., 2008). When the domain and its related hypothesis space are known a priori, the previous results might provide a sufficient explanation of how people generalize properties in different domains. However, this is an implausible assumption for many real-world instances of generalization. In many contexts, people face unfamiliar stimuli possessing a multitude of features and are not told which features and/or hypothesis spaces should be used to generalize. Nonetheless, people are able to use the structure of stimuli in the current context to *learn how to generalize*.

The computational problem of learning how to generalize is equivalent to learning an *overhypothesis* (Goodman, 1955) that determines which hypotheses are “lawlike” in a domain or context. For example, consider finding a box filled with bags of marbles left in your attic by the previous occupant of your home. You look in two bags. The first bag contains all white marbles, while the second bag contains all black marbles. A hypothesis for a bag would be the probability a marble is white or black (e.g., it might be the case that 80% of the marbles are black and 20% are white in a bag). An overhypothesis would be the expectation you infer about hypotheses. In this case, after observing the two bags, you expect only two hypotheses: the hypothesis where all marbles are white and the hypothesis where all marbles are black. This aids future generalization. For the first marble in the third bag, you would be indifferent between it being white or black (across bags the probabilities are equal). However, once you learned the color of a single marble in this new bag, you would expect all the marbles in the new bag to be that color (Kemp, Perfors, & Tenenbaum, 2007).

Previous work has established the power of hierarchical Bayesian modeling as a possible explanation for how people could learn overhypotheses across a broad set of domains ranging from the visual images for a novel object (Salakhutdinov, Tenenbaum, & Torralba, 2012), to the visual relations between parts of characters in a novel alphabet (Lake, Salakhutdinov, & Tenenbaum, 2015), to other types of words and causal relationships (Kemp, Perfors, & Tenenbaum, 2007; Mansinghka, Kemp, Tenenbaum, & Griffiths, 2006; Perfors, Tenenbaum, & Regier, 2011). However, previous models are formulated for specific domains such as learning characters or words and have not been evaluated empirically through multiple, behavioral experiments across different domains. It remains an open question whether individuals use the same evidence and make the same inferences as these models. Further, previous work has not explored the implications of hierarchical Bayesian modeling in the context of generalization, where the relevant inference concerns the actual hypothesis space to adopt.

In this paper, we examine how people should and do learn to generalize in perceptual and conceptual domains. First, we build on the Bayesian generalization framework to show how an ideal learner should learn how to generalize. Our analysis indicates that the same framework with different sets of hypothesized patterns can explain how complex patterns of generalization can be learned. We test these predictions empirically in three domains: learning the dimensions to generalize properties over perceptual stimuli (rectangles) and performing property induction over two types of conceptual stimuli (animals and numbers).

2. Learning to generalize

Our mathematical framework builds directly on the Bayesian generalization model (Shepard, 1987), so we first summarize this approach. We then show how it produces different patterns of generalization behavior depending on its hypothesis space, and how to extend it to learn an appropriate hypothesis space for generalization.

2.1. Bayesian generalization framework

After observing that stimuli **x** has some property,[1] which other stimuli should have that property? For example, if a dark, large circle is a *gnarble*, which other stimuli are likely to be *gnarbles*? Shepard (1987) argued that this problem could be solved by assuming stimuli are points in a psychological space and *gnarbles* occupy a region in that space. More generally, we can imagine a set of hypotheses $\mathcal{H}$ about which objects are *gnarbles* and use Bayesian inference to evaluate the plausibility of the hypotheses (as reflected in the *posterior probability* of those hypotheses, $P(h|\mathbf{x})$; Tenenbaum & Griffiths, 2001). Assuming that stimuli are generated uniformly and independently from the true hypothesis ($P(\mathbf{x}|\mathrm{h}) = |h|^{-1}$, where $|h|$ is the number of stimuli having the property according to h) and some prior beliefs over hypotheses $P(h)$, the posterior probability that hypothesis h is the property that n given stimuli share is given by Bayes' rule.

$$P(h|\mathbf{x}) = \frac{P(h)\prod_{i=1}^{n} P(x_i|h)}{\sum_{h' \in \mathcal{H}} P(h')\prod_{i=1}^{n} P(x_i|h')} \tag{1}$$

This incorporates the prior plausibility of each hypothesis ($P(h)$) as well as the consistency of each hypothesis with the stimuli observed to be *gnarbles* so far ($P(\mathbf{x}|h)$). The probability of generalizing from **x** to another stimulus y is the sum of the posterior probabilities (Eq. 1) of all hypotheses under which the new stimulus would be a *gnarble.*

$$P(y|\mathbf{x}) = \sum_{h:y \in h} P(h|\mathbf{x}) \tag{2}$$

which constitutes a form of *hypothesis averaging* (Robert, 2007).

The predictions of the Bayesian generalization framework depend intimately on the nature of the hypotheses under consideration, with different hypothesis spaces leading to different generalization patterns. For example, Fig. 1(a and b) show that when generalizing over a two-dimensional space, a hypothesis space containing intervals that vary over one consequential dimension result in standard one-dimensional generalization gradients over that consequential dimension. As Fig. 1(b) shows, the consequential dimension and the "experimenter-defined" axes of the stimulus space are not necessarily aligned, which can result in a very different pattern of generalization behavior. Fig. 1(c–d) show that when generalizing what animals are likely to share a novel protein found in grizzly bears, which hypothesis space you use affects how likely different animals are to share the protein with grizzly bears. When a hypothesis space containing predator–prey pairs as hypotheses is used, salmon are likely to share the novel protein with grizzly bears (Fig. 1c). When a taxonomic hypothesis space is used, panda bears are likely to share the novel protein with grizzly bears (Fig. 1d).

The hypothesis space used in Bayesian generalization models is typically fixed and specified a priori. Thus, generalization behavior for a particular stimulus and property is also fixed and specified a priori. However, as noted above, human generalization behavior is strongly influenced by context. Under the standard Bayesian generalization framework, such sensitivity falls out of the scope of normative predictions. However, we argue that certain contextual information, such as the structure of previously learned concepts in a domain, provides higher-level data that may be incorporated into future inferences to facilitate faster generalization, thus permitting transfer to new concepts. We now turn to the question of how an ideal learner should solve this problem of learning how to generalize in a domain based on previously learned concepts.

2.2. Learning a hypothesis space

We can extend the Bayesian generalization framework to learn what hypotheses are "lawlike" (in the sense of Goodman, 1955) by learning concepts in a domain, where a concept is a set of stimuli sharing a property. Given a set of possible hypothesis spaces, we formulate a hierarchical Bayesian model, where the appropriate hypothesis space for generalization is itself a higher-level random variable.[2] This higher-level variable is an "overhypothesis," representing our belief that each hypothesis space is appropriate for generalizing. The posterior probability of each hypothesis space given the observed concepts encodes how well the hypothesis space explains the concepts. To generalize, a Bayesian agent would take the weighted average over the generalization predictions resulting from using each hypothesis space, with weights corresponding to posterior probabilities. One can interpret this procedure as several Bayesian generalization models running in parallel that are then averaged together (where each generalization is weighted to the extent that its hypothesis space explained previously learned concepts in the domain). Learning an overhypothesis by this process results in learning how to generalize.

Formally, given a set of M possible hypothesis spaces $\mathcal{M} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, we define a hierarchical Bayesian model where the appropriate hypothesis space for generalization is itself a higher-level random variable. There are two ways that the model could generalize to a new stimulus y given an observed set of concepts $\mathbf{C} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and a set of stimuli (with the current property of interest) $\mathbf{x}_{n+1}$ with probability.

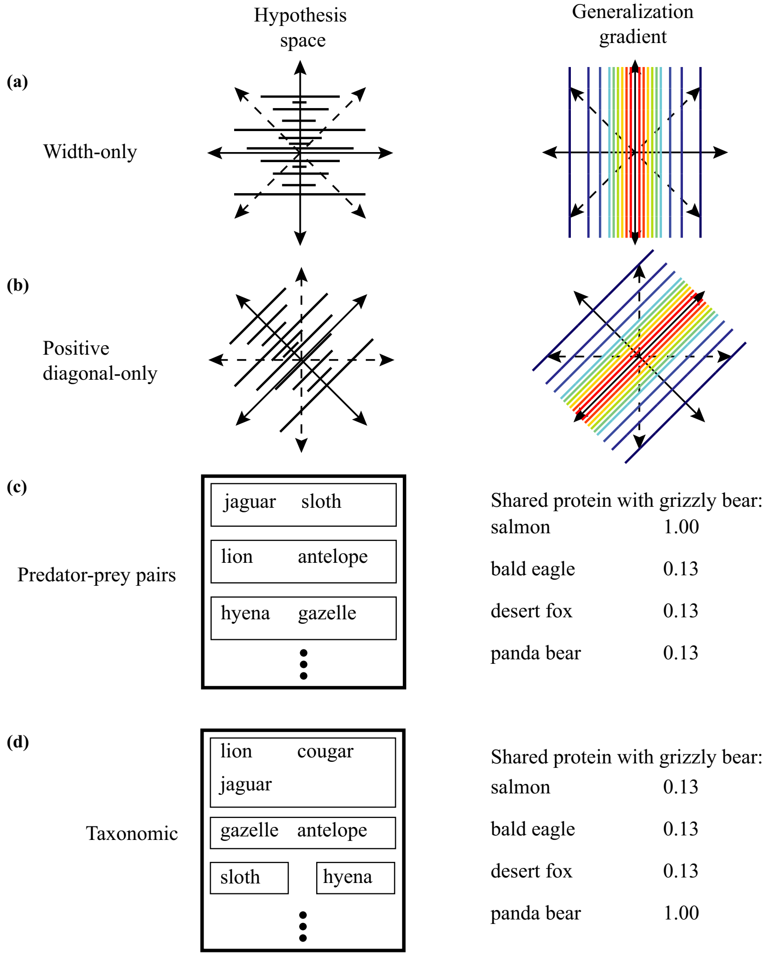

Fig. 1. Bayesian generalization. (a and b) Width-only and positive-diagonal-only hypothesis spaces and the resulting generalization gradients for two-dimensional perceptual stimuli. The generalization gradients are displayed as contour plots, where generalization probabilities are equal on a contour and warm colors represent greater probabilities. Each hypothesis is an interval in 2-D space, which is either aligned with the axes or diagonals (and has arbitrary extent in the orthogonal dimension). (c and d) Predator–prey pairs and taxonomic hypothesis spaces and their resulting generalization gradients for animal stimuli. Although the perceptual and animal stimuli are different domains, generalizations are made using the same underlying Bayesian framework, but with differing hypothesis spaces. It is important to note that the same mathematical procedure with different hypothesis spaces results in different generalization behavior.

$$P(\mathrm{y}|\mathbf{C}, \mathbf{x}_{n+1}) = \sum_{m=1}^{M} P(\mathrm{y}|\mathbf{x}_{n+1}, \mathcal{H}_m) P(\mathcal{H}_m|\mathbf{C}, \mathbf{x}_{n+1}), \tag{3}$$

where $P(y|\mathbf{x}_{n+1}, \mathcal{H}_m)$ is the probability of generalizing from the currently observed stimuli $\mathbf{x}_{n+1}$ to y under hypothesis space $\mathcal{H}_m$ (as specified by Eq. 2) and $P(\mathcal{H}_m|\mathbf{C}, \mathbf{x}_{n+1})$ the second term of Eq. 3 is the posterior probability of hypothesis space $\mathcal{H}_m$ given the previously learned concepts and the current stimuli. This approach is known as *model averaging*, and it is the standard method used to predict new items using a hierarchical Bayesian model. An alternative is *model selection*, which uses only the hypothesis space with maximum posterior probability for generalization, denoted $\mathcal{H}^*$.

$$P(y|\mathbf{C}, \mathbf{x}_{n+1}) = P(y|\mathbf{x}_{n+1}, \mathcal{H}^*), \text{where } \mathcal{H}^*_m = \arg\max_{\mathcal{H}\in M} P(\mathcal{H}|\mathbf{C}, \mathbf{x}_{n+1}) \tag{4}$$

Depending on the goal of the agent, both strategies can be viewed as normative and are used in Bayesian statistics when the model for a domain is uncertain (Clyde et al., 2007). This posterior probability can be computed by applying Bayes' rule.

$$P(\mathcal{H}_k|\mathbf{C}, \mathbf{x}_{n+1}) = \frac{P(\mathbf{C}, \mathbf{x}_{n+1}|\mathcal{H}_k)P(\mathcal{H}_k)}{\sum_{m=1}^{M} P(\mathbf{C}, \mathbf{x}_{n+1}|\mathcal{H}_m)P(\mathcal{H}_m)}, \tag{5}$$

where $P(\mathbf{C}, \mathbf{x}_{n+1}|\mathcal{H}_k)$ is the probability of observing a set of concepts $\mathbf{C}$ and the current stimuli in the new concept $\mathbf{x}_{n+1}$ under the hypothesis space $\mathcal{H}_k$ and $P(\mathcal{H}_k)$ is the prior probability of hypothesis space $\mathcal{H}_k$. The probability of the previously seen concepts and the currently observed stimuli $\mathbf{x}_{n+1}$ under hypothesis space $\mathcal{H}_k$ is.

$$P(\mathbf{C}, \mathbf{x}_{n+1}|\mathcal{H}_k) = \prod_{i=1}^{n+1} \sum_{h\in\mathcal{H}_k} P(h|\mathcal{H}_k) \prod_{x\in x_i} P(x|h), \tag{6}$$

where $P(h|\mathcal{H}_k)$ is the prior on h under hypothesis space $\mathcal{H}_k$ and $P(x|h)$ is the likelihood of stimulus x given h.

For example, consider the four hypothesis spaces for properties over a two-dimensional perceptual domain in Fig. 2(a). In this example, each point in the space corresponds to a rectangle and a concept is a group of rectangles that share a property. The rectangles in each concept share either the same width or height, which is more likely under the width-only and height-only hypothesis spaces (each concept fits in smaller intervals of the width-only and height-only hypothesis spaces). After observing these concepts, the model learns to generalize based on the width-only and height-only hypothesis spaces. Fig. 2(b) illustrates how the model learns to generalize in the animal domain. After observing two concepts (that jaguars and sloths share a novel protein and lions and antelopes share a different novel protein), the posterior probability of the *predator–prey pair* hypothesis space is the largest, though there is considerable probability for the *geographic* hypothesis space. This results in both approaches generalizing a novel property from grizzly bears to both salmon and bald eagles. The *model averaging* approach also predicts that the bald eagle is more likely to share the property with the grizzly bear than the other two animals, because it averages in generalizing under the *geographic* hypothesis space (grizzly bears and bald eagles both live in North America). This illustrates that model selection and model averaging can make different predictions for how people should learn to generalize novel concepts.

2.3. Testing the predictions of the model

The hierarchical Bayesian approach to learning how to generalize that we have presented makes clear predictions about how people should differ in their pattern of generalization based on their experience in a domain. In addition to testing these basic predictions, the experiments we present below examine the two strategies for learning to generalize: model averaging and model selection. Prior work in categorization and inductive inference has extensively explored a related problem over the last few decades: how do people predict unknown properties of a novel object when the category membership of that object is unknown?

Previous work has argued that a Bayesian agent should infer unknown properties of the novel object by averaging over hypotheses, taking the weighted average of the probability of the properties given each possible category where the weights are determined by how probable each category is given the observed properties of the known objects (Anderson, 1991; Murphy & Ross, 1994). Most empirical work contradicts this prediction: People tend to rate the likelihood that the novel object has the property according to the relative frequency of that property in the most probable category for the object (Hayes & Newell, 2009; Malt, Ross, & Murphy, 1995; Murphy & Ross, 1994). Based on this related previous empirical work, we should expect people to learn to generalize in our studies according to model selection. However, there are some cases where people do take into account multiple categories during induction. This usually requires special stimulus presentation or other experimental procedures or measures beyond what is typically used in inductive inference tasks (Chen, Ross, & Murphy, 2014, 2016; Papadopoulos, Hayes, & Newell, 2011; Verde, Murphy & Ross, 2005). Thus, it is possible that people might learn to generalize using model averaging instead.

Our analysis of learning how to generalize predicts that people should learn how to generalize in a domain based on the structure of the concepts observed in that domain. In Experiment 1, we test this prediction in a perceptual domain (rectangles) by teaching people to represent rectangles according to the appropriate set of dimensions for the concepts learned over rectangles. In Experiment 2, we investigate this prediction in a structured conceptual domain (animals) using a property induction task. In Experiment 3, we explicitly examine whether people perform model averaging or selection, in a different structured conceptual domain (numbers), using individual-level analyses. These three experiments across a broad range of domains provide a robust test of the model’s predictions.

3. Experiment 1: Learning dimensions to represent rectangles

The proposed model predicts that a learner should be able to infer the dimensions for representing rectangles in a novel domain from observed examples of concepts expressed

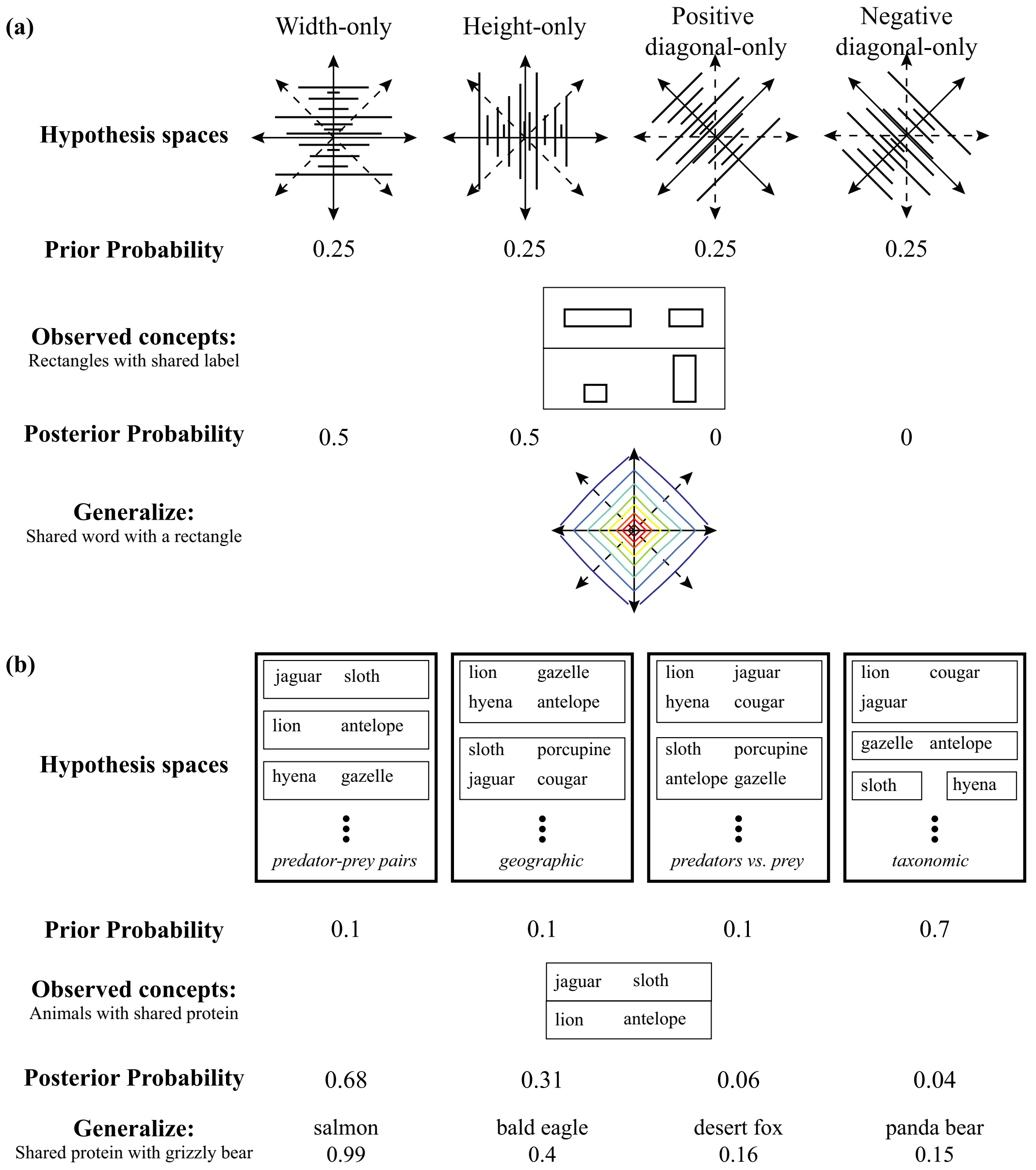

Fig. 2. Learning hypothesis spaces. (a) A two-dimensional perceptual domain. Each point in the space corresponds to a rectangle and a concept is a group of rectangles that share a property (in this case, being called a novel word). The rectangles in each concept share either the same width or height, which are more likely under the width and height hypothesis spaces. After observing these two concepts, the model learns to generalize based on the width and height hypothesis spaces. (b) An analogous example of how the model learns hypothesis spaces in a conceptual domain.

in that domain. Preliminary support for this prediction is provided by the results of Goldstone (1994), who showed that teaching people a novel axis-aligned concept could affect generalization along that axis. To perform a more thorough test of our predictions, we conduct an experiment in which we examined how learning concepts without feedback affects people's generalization judgments. We use rectangles varying in width and height as our set of stimuli and define concepts on two sets of experimentally manipulated dimensions: same width or height and same aspect ratio or area. These two sets of dimensions correspond to four hypothesis spaces for generalization shown in Fig. 1(a–b). Previously, Krantz and Tversky (1975) found that people weakly favor using area and aspect ratio as separable dimensions. However, people can use either pair of dimensions for generalizing depending on the context of previously learned concepts. This natural flexibility makes rectangles an ideal candidate for testing our predictions.

3.1. Methods

3.1.1. Participants

A total of 86 undergraduates from UC Berkeley participated for course credit.

3.1.2. Stimuli and procedure

The stimuli are rectangles varying in width and height (13–115 pixels in increments of approximately 25 pixels, with monitor dimensions of 18.75×10.5 inches with $1,920 \times 1,080$ resolution). The stimulus set is shown in Fig. 3. Depending on their assigned condition, participants learn 16 concepts that are either aligned with or orthogonal to the dimensions given by Fig. 4.[3]

Participants read the following "cover story," which sets the task in a more naturalistic context: "On an island in the Pacific Ocean, scientists found the ancient ruins of a small civilization. While excavating the ruins, they discovered symbols on the doors of the ruined houses. The scientists believe that the symbol on the door of the house carries information about the family that lived there. Some of the symbols were labeled, and each symbol was found with more than one label."

There are two phases to the experiment: training and test. For the training phase, there are two between-subjects conditions: the *axis-aligned* condition (n = 42), in which people learn the 16 axis-aligned concepts shown in Fig. 4(a), and the *diagonal-aligned* condition (n = 44), in which people learn the 16 diagonal-aligned concepts shown in Fig. 4(b). The conditions are matched such that every stimulus was presented the same number of times to the participants, there are the same number of objects in each concept, the concepts span the space of objects, and the variability of objects over each dimension is equal. This allows us to infer that any differences in generalization behavior must be due to differences in the structure of the concepts learned by participants in the conditions (as everything else was identical). Thus, our predictions are supported if we observe that the *axis-aligned* participants generalize novel properties more on the axes (constant width or height) than the *diagonal-aligned* participants, and conversely for the diagonals (constant aspect ratio or area).

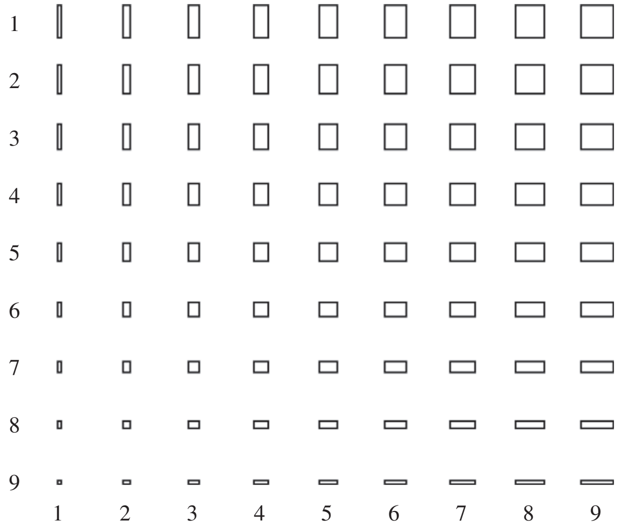

Fig. 3. The set of stimuli used in Experiment 1 (not to scale).

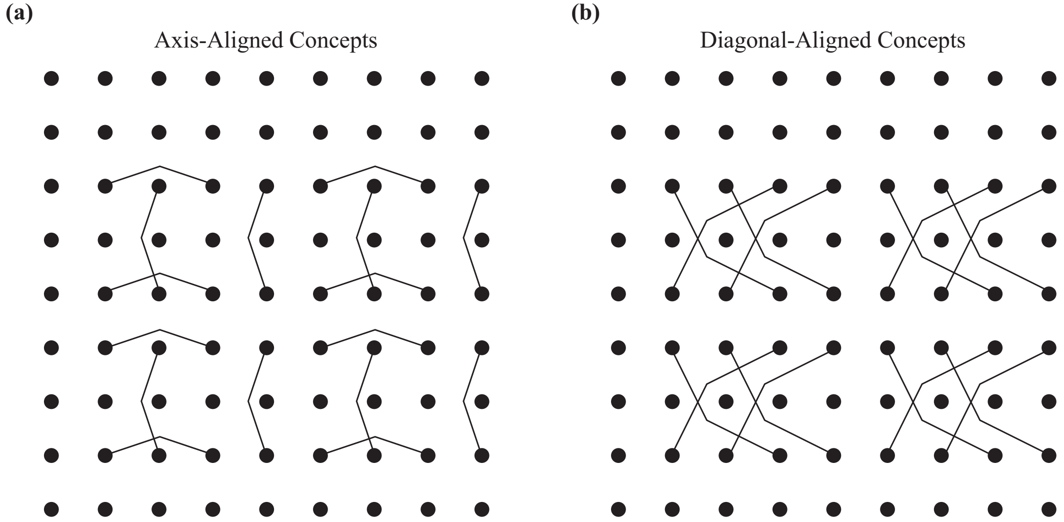

Fig. 4. The 16 concepts for the (a) *axis-aligned* and (b) *diagonal-aligned* conditions. Each concept is the collection of objects on a straight line on the grid. For the *axis-aligned* condition, all objects share either the same width or height. For the *diagonal-aligned* condition, all objects in a concept share either the same aspect ratio or area. Each object occurs the same number of times in each condition.

The 16 concepts are presented to participants in a random order as examples of objects that are called different nonsense names randomly chosen from a standardized list. For each trial in the training phase, the objects in each concept are displayed on the screen and participants are asked whether they thought the current object shown individually below the objects in the concept could be called that name. For each concept, this is done with every object with dimension values in $\{1,3,5,7,9\} \times \{1,3,5,7,9\}$ (we only used this subset of the stimuli to keep the experiment relatively short). The test phase of the experiment is identical to the first phase except participants generalizations are tested for concepts consisting of single objects ($\{(2,2),(2,8),(5,5),(8,2),(8,8)\}$ were tested) over the total 9 × 9 set of objects. No feedback is ever given in either condition.

3.2. Results and discussion

Fig. 5(b) shows participant generalization responses for the test phase in the two conditions. The responses were aligned and then averaged over the five concepts per participant and then over participants in the condition. We then took the difference between the responses for the two conditions and compared them to the difference between the generalization predictions produced by the Bayesian model shown in Fig. 5(a).[4] Participants in the *diagonal-aligned* condition generalized more on the diagonals than those in the *axis-aligned* condition (and vice versa), supporting the model predictions (16/24 predicted to be larger by the *axis-aligned* condition, $p = .08$, and 27/32 for the *diagonal-aligned* condition, $p < .001$).[5] As the differences in generalization between the two groups of participants in Fig. 5(b) is not limited to the axes and diagonals, our results are not consistent with the "coincidence effect" of Tversky and Gati (1982), which predicts generalization only when the dimension values are precisely equal. Overall, these results are consistent with the predicted change in the pattern of generalization that is indicated by our hierarchical Bayesian model.

4. Learning hypothesis spaces and selective attention

The results of Experiment 1 show a clear effect of prior experience on people's patterns of generalization for novel concepts: The axes along which people generalize depend on the kind of concepts that they had been exposed to. This phenomenon has some parallels with previous work on dimensionally selective attention in the categorization literature, which has shown that if a dimension is more relevant for accurately categorizing a set of stimuli, participants will rely on that dimension more than other dimensions to generalize, a phenomenon known as *selective attention* (Kruschke, 1992; Nosofsky, 1986; Shepard, 1964). When categories are defined as regions in some geometric space, the result of selectively attending to a useful dimension is to "stretch" it relative to other dimensions such that changes in the attended dimension affect generalization more than changes in the other dimensions (which selective attention "shrinks"). For example, in the Generalized Context Model and its variants (Kruschke, 1992; Nosofsky, 1986), the calculation of distance, b_{ij}, between stimuli $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is affected by the selective attention to each dimension w_d

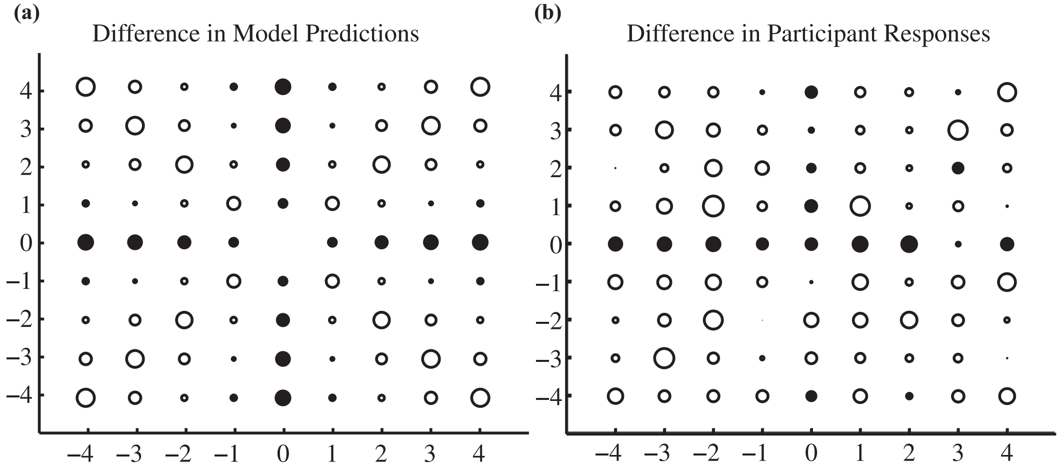

Fig. 5. Model predictions and results for Experiment 1. (a) Predicted differences in generalization produced by Bayesian models learning from the *axis-aligned* and *diagonal-aligned* concepts. (b) Differences between participant responses in the *axis-aligned* and *diagonal-aligned* conditions. The results are presented as bubble plots where the size of the bubble represents the degree of generalization. Solid and open bubbles represent positive and negative values, respectively. The generalizations people produced were aligned and then averaged over the five concepts per participant and over participants. Participants in the *diagonal-aligned* condition generalized more on the diagonals than those in the *axis-aligned* condition (and vice versa).

$$b_{ij} = \left(\sum_{d=1}^{D} w_d \left| x_{id} - x_{jd} \right|^r \right)^{1/r},$$

where w_d is non-negative and the amount of attentional weight sums to one. This is then used to specify the similarity between pairs of stimuli s_{ij} by defining a generalization gradient, for example $s_{ij} = \exp(-b_{ij})$.

Our results in Experiment 1 can be interpreted in similar light: if we characterize the different dimensions along which rectangles can be represented as being given different weights in people's generalizations, then our results can be taken as showing that people can shift those weights from favoring one set of dimensions (area and aspect ratio) to another (height and width). Indeed, this shouldn't come as a surprise: there is a natural parallel between this kind of shift of dimensional attention and our hierarchical Bayesian approach to generalization.

The two hypothesis spaces contrasted with one another in Experiment 1 can be interpreted as picking out different axes along which rectangles should be represented for generalization. Within each hypothesis space, we might imagine that hypotheses are axis-aligned regions—the multi-dimensional analogue of the one-dimensional intervals that Shepard (1987) used in his original analysis of generalization. For the original case, Tenenbaum (1999, Appendix B) showed that the Bayesian generalization model with a

Gamma distribution for the prior on the size of each dimension of the regions produces patterns of generalization given by.

$$P(y|x) \propto \exp\left(-\sum_{d=1}^{D} w_d|x_d - y_d|\right),$$

where w_d depends on the parameters of the prior: If the prior favors smaller regions along that dimension, then w_d is greater, consistent with an increased cost of generalizing along that dimension. This is proportional to the weighted Minkowski distance formula when $r = 1$. This is appropriate given that selective attention focuses on the case of stimuli with separable dimensions which is typically modeled by a weighted Minkowski distance with $r = 1$.

Using this result, we can characterize the Bayesian hypothesis space learning approach as adjusting the relative contribution of two different generalization gradients, each assigning different weights to the underlying dimensions. One hypothesis space assigns zero weight to height and width and non-zero weight to area and aspect ratio. The other assigns zero weight to area and aspect ratio and non-zero weight to height and width. The average of the generalization gradients that result from these different hypothesis spaces can thus be approximated by a single generalization gradient in which the weights of all four dimensions are adjusted.

Despite these parallels, there are significant ways in which the previous empirical and theoretical treatment of selective attention relates to our results. First, selective attention is typically implemented using error-driven learning to adjust the weights assigned to different dimensions of a psychological space prior to applying a simple generalization function (e.g., Kruschke, 1992). Typical studies examining category learning and selective attention highlight the importance of corrective feedback (e.g., Rehder & Hoffman, 2005). Further, recent work has found differences in supervised human category learning depending on whether participants guess and are given feedback when taught exemplars from each category, or simply observe labeled examples from each category (Kurtz, 2015; Levering & Kurtz, 2015). Thus, to the best of our knowledge, it is unknown whether people learn to selectively attend to dimensions during supervised category learning in the absence of feedback. Our studies examine this type of context and find that people can change the dimensions they use in generalization without corrective feedback.

Second—and perhaps more important—under our approach, the change in the way that dimensions are treated is a result of a *qualitative* shift of representation rather than a *quantitative* shift of attention. In the hierarchical Bayesian model, this shift is a result of a discrete change in the hypothesis space being used. In the case of model selection, generalization *only* depends on the dimensions consistent with that hypothesis space. Variation along other dimensions is ignored, assuming that the variation is not large enough to change which dimension is most diagnostic. Even in model averaging, we expect systematic changes in which the dimensions associated with specific hypothesis spaces come to dominate over time while the contribution of others diminishes to zero.

Because our framework provides a computational-level account, we anticipate that there are (algorithmic-level) process models that could capture the pattern of behavior shown in our experiment through appropriately tuned mechanisms for selective attention based on distributional information within each concept (though it would be difficult to capture these results with purely feedback-driven selective attention mechanisms). However, a shift in which continuous dimensions are used to support generalization is only one aspect of the predictions of our framework. To demonstrate that our approach provides a broad account of learning how to generalize, we conducted two experiments with more structured concepts, showing that similar results hold in domains where it is harder to pick out stimulus dimensions and apply a selective attention account, and where we can test some of the more specific predictions of the framework.

5. Experiment 2: Learning a hypothesis space for animals

Our extension to the Bayesian generalization framework for learning how to generalize predicts that people should use the concepts observed in a given context to generalize new concepts in that context. In Experiment 2, we test this prediction in a conceptual domain by exploring how participants generalize properties in a property induction task in which we demonstrate that certain animals have some property and ask participants to evaluate which other animals share that property. With this task, we target the question of whether the context of learned concepts affects how people generalize properties in property induction.

5.1. Methods

5.1.1. Participants

At UC Berkeley, 752 undergraduates participated for course credit. There were four between-subjects conditions, with 189, 188, 190, and 185 participants in the *predator-prey pairs*, *geographic*, *predators vs. prey*, and *taxonomic* conditions, respectively.

5.1.2. Stimuli and procedure

After reading a story about proteins found in the blood of eight animals (cougar, porcupine, lion, antelope, jaguar, sloth, striped hyena, and gazelle), participants rate how likely four other animals (salmon, desert fox, bald eagle, and panda) are to share a protein with a grizzly bear. The groupings of animals sharing the same protein are determined by the participant's condition as shown in Fig. 6.

For example, participants in the *predator-prey pairs* condition read the following instructions (other conditions used the same cover story except animals shared proteins according to the appropriate conceptual relationship given in Fig. 6):

"Imagine that you are a scientist, trying to learn about animal physiology. You hear that other scientists have recently discovered some proteins in the blood of different animals that protect them against Toma Disorder. Here are their findings so far:

	Protein A	Protein B	Protein C	Protein D	Protein E
Predator-prey pairs	cougar porcupine	lion antelope	jaguar sloth	hyena gazelle	
Geographic	antelope lion hyena gazelle	sloth jaguar porcupine cougar			
Predators vs. prey	cougar lion hyena jaguar	sloth gazelle porcupine antelope			
Taxonomic	cougar jaguar lion	gazelle antelope	hyena	sloth	porcupine

Fig. 6. Assignments of animals to proteins for the conditions in Experiment 1 with appropriate groupings of animals being used in each condition. The assignments determine the context of previous concepts that are used by participants and the model to learn how to perform property induction.

Protein A has been found in cougar and porcupine blood.
Protein B has been found in lion and antelope blood.
Protein C has been found in jaguar and sloth blood.
Protein D has been found in striped hyena and gazelle blood.
Scientists have recently found Protein E in grizzly bear blood. You want to figure out which other animals might have Protein E in their blood."

After reading the cover story, participants respond on a 1-7 Likert rating scale, ranging from 1 (*Very Unlikely*) to 7 (*Very Likely*), how likely they think it was for Protein E to be found in the blood of the four test animals (*salmon*, *desert fox*, *bald eagle*, and *panda*).

5.2. Results and discussion

Fig. 7(a) shows the averaged participant judgments. Each group of bars depicts how generalization judgments change for the same test item depending on the context. Participants generalize a protein found in grizzly bear blood differently depending on the context of which animals shared other proteins. We conducted an analysis of deviance on a mixed-effects ordinal logistic regression (following Liddell & Kruschke, 2018), which resulted in a significant effect of context, $\chi^2\ (3, N = 752) = 25.24, p < .001$, and an interaction of context with judgment $\chi^2\ (9, N = 752) = 161.02, p < .001$. In particular, participants generalize the protein in grizzly bear blood to a prey, salmon, more when other predator–prey pairs shared a protein ($t(2981) = 4.10, p < .001$). Further, they generalize the protein in grizzly bear blood to a biological relative, panda, more when other taxonomic relatives shared a protein ($t(2981) = 6.38, p < .001$). Thus, participants extend properties based on the conceptual relations they observed in the context.

To evaluate the model, each conceptual structure (condition) is represented as a hypothesis space. The prior distribution over them has one parameter for the weight of the taxonomic hypothesis space, with the remaining mass distributed unfiormly over the other hypothesis spaces. The parameter was set to 0.9999 by minimizing the distance between the model results and the participant responses. This large value is consistent with prior research suggesting that taxonomic relationships are salient for induction of biological properties (Heit & Rubinstein, 1994).

The hypothesis spaces were formulated prior to looking at the participant responses. The *predator–prey pairs* hypothesis space contains typical pairings of predator-prey relationships. For example, in this hypothesis space, the cougar is in two hypotheses (with porcupine and sloth), and the grizzly bear is in one hypothesis (with salmon). The *geographic* hypothesis space has three hypotheses, each containing the animals typically found on the continent (South America, Africa, and North America). The *predators vs. prey* hypothesis space has two hypotheses, preys (porcupine, antelope, sloth, gazelle, and salmon) and predators (cougar, lion, jaguar, hyena, grizzly bear, desert fox, and bald eagle). The *taxonomic* hypothesis space contains 10 hypotheses: the big cats, bovids (antelope and gazelle), bears, and single member hypotheses for the other animals. To ensure that a hypothesis space never has zero probability, a "catch-all" hypothesis containing every animal was also added to each hypothesis space. The size of each hypothesis is the number of animals it contains and the prior distribution over hypotheses in each hypothesis space was uniform.

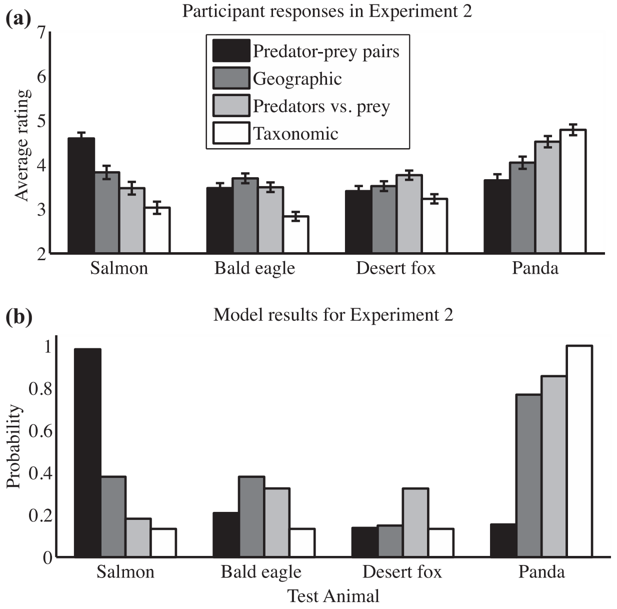

Fig. 7. (a) Participant responses and (b) model averaging results for Experiment 2. Participants and the model change how they extend a property (having a protein that is found in the blood of grizzly bear) appropriate to the observed conceptual structure of the context (assignment of animals to proteins).

Fig. 7(b) shows the model results for Experiment 2. The model qualitatively captures how people change their judgments for each test animal in the four contexts. Both people and the model generalize from the grizzly bear to each test animal the most in its favored context. Also, both give the largest generalization judgments to the salmon in the *predator-prey pairs* context and the panda in the *taxonomic* context. Additionally, model averaging accounts for participant responses better quantitatively ($r = .93$, Pearson's and Spearman's correlations), then model selection ($r = .86$ and $r = .28$, Pearson's and Spearman's correlations, respectively).

6. Experiment 3: Model averaging and model selection with number concepts

Experiments 1 and 2 support that people can learn to generalize in a context based on the concepts observed in that context, as predicted by our extension to the Bayesian generalization framework. The results of Experiment 2 also provide weak support for people performing model averaging rather than model selection. However, as we only have a single data point from each participant, it is difficult to determine whether participants generalize using model averaging or model selection. In Experiment 3, we expand on Experiment 2 by asking each participant to generalize multiple times given different information, thus enabling us to identify which strategy each participant is using. To provide converging evidence for the framework, we used a different domain for the experiment: numbers between 1 and 100.

6.1. Methods

6.1.1. Participants

A total of 188 adults residing in the United States were recruited through Amazon Mechanical Turk and received $2.00 for their participation. Participants were excluded from analyses if they gave the same response to every question throughout the experiment (four participants) or if they had previously participated (four participants). Eight participants in total were excluded for such reasons, resulting in a final sample of 180 participants. Participants were assigned to either a *magnitude* or *mathematical* condition. Three different between-subjects sets of stimuli were used for each condition, which resulted in six groups of 30 participants each.

6.1.2. Stimuli and procedure

We use a variant of the number game task described by Tenenbaum (2000). Participants are informed that they would learn about four simple computer programs, each of which accepts a certain set of numbers between 1 and 100. The participants' goal is to determine which numbers each program accepts. Participants are told that they would see several random examples of accepted numbers from each program.

The experiment consists of a learning phase and test phase. In the learning phase, participants are introduced to the first three "programs. "For each program, participants are shown three accepted numbers one at a time, which together imply a particular number concept. For participants in a *magnitude* condition, the concepts are defined by a shared property of their magnitudes, such as being “between 10 and 20” or “between 72 and 90.” For participants in a *mathematical* condition, the concepts are defined by a mathematical property shared by the three numbers, such as being “multiples of 5” or “even numbers.” We use three different stimulus sets of paired *magnitude* and *mathematical* conditions. Participants in the *magnitude* and *mathematical* conditions within each stimulus set were shown the same numbers over the course of the study, but the numbers were arranged into different concepts in the two conditions. See Fig. 8 for full details of the stimuli.

Example numbers are shown to participants one at a time. After each new example is displayed, participants are given a set of 16 response numbers, displayed on one page, in random order. Participants provide responses on a 1–7 Likert rating scale, ranging from 1 (*Very Unlikely*) to 7 (*Very Likely*), for how likely they think it is that the program would accept each number. On each trial, participants were shown a single number and asked to provide Likert ratings for a set of response numbers. The same sets of response numbers are used across conditions, within a stimulus set. Each program had a different response set. In the learning phase, this set is comprised of sixteen numbers that fall into four categories: Magnitude (four numbers), Mathematical (four numbers), and Random (four numbers). The Magnitude response numbers have high posterior probability in the magnitude condition, and the Mathematical response numbers have high posterior probability in the mathematical condition. Random numbers are unrelated to the example numbers in both conditions.

The learning phase provides a context within which participants can learn a hypothesis space for generalization. In the test phase, we examine whether the context changes participants’ generalization patterns for a novel, ambiguous example. Participants are introduced to the final program and are shown a single example number (the test number). The test number was chosen to have minimal difference in a priori probability under each hypothesis space. In the test phase, participants provide ratings for 20 response numbers. Eight of these numbers are unrelated to the example number (Random type), four are related by magnitude (Magnitude type), four are related mathematically (Mathematical type), and an additional four are related by both mathematical properties and magnitude (Both type).

Although the test number does not by itself imply a particular concept or concept type, we predicted that participants in the magnitude condition would generalize more to Magnitude response numbers and participants in the mathematical condition would generalize more to Mathematical response numbers in accordance with the predictions of the hierarchical model. Further, we predicted minimal difference across conditions in generalization to Random and Both numbers.

6.2. Results and discussion

Fig. 9(a) shows average participant responses for the test case across all stimulus sets, aggregated by response number types. Fig. 10(a) and (c) show averaged participant responses for a particular test case (stimulus set 1), not aggregated by number types,

		Program A	**Program B**	**Program C**	**Test Number**
Set 1	Magnitude	*Interval 93…96* [96, 95, 93]	*Interval 60…65* [65, 66, 60]	*Interval 10…15* [15, 12, 10]	[32]
	Mathematical	*Multiples of 12* [96, 12, 60]	*Multiples of 5* [65, 95, 10]	*Multiples of 3* [15, 66, 93]	
Set 2	Magnitude	*Interval 40…49* [40, 49, 48]	*Interval 6…10* [7, 6, 10]	*Interval 70…77* [72, 77, 70]	[18]
	Mathematical	*Multiples of 10* [40, 10, 70]	*Multiples of 7* [7, 49, 77]	*Multiples of 6* [72, 6, 48]	
Set 3	Magnitude	*Interval 22…27* [22, 27, 24]	*Interval 32…36* [32, 33, 36]	*Interval 80…88* [81, 80, 88]	[48]
	Mathematical	*Multiples of 11* [22, 33, 88]	*Multiples of 8* [32, 24, 80]	*Multiples of 9* [81, 27, 36]	

Fig. 8. Concepts and example numbers used in each condition. Within each stimulus set, the same nine example numbers are shown to participants during the learning phase, but they are arranged into different types of concepts (*magnitude* or *mathematical*) in the two conditions. Within each stimulus set, the same test number is used across conditions.

illustrating the different generalization patterns seen in the different conditions. Participants generalized program acceptance differently depending on the property type (*magnitude* or *mathematical*) that defined previously observed acceptance. An Analysis of Deviance on Ordinal Linear Regression revealed main effects of condition $F(1, 2880) = 66.59, p < .0001$, and response number type $F(3, 522) = 58.56, p < .0001$, an interaction effect of condition and type $F(3, 2880) = 39.64, p < .0001$, and no effect of stimulus set $F(2, 174) = 1.92, p = .15$.

Further, participants generalized to Magnitude numbers more when observed concepts were defined by magnitude properties $(M = 5.32, SD = 1.21)$ than when they were defined by mathematical properties ($M = 2.82$, $SD = 1.65, t(178) = 11.58, p < .0001$), and participants generalized to Mathematical numbers more when observed concepts were defined by mathematical properties ($M = 4.53$, $SD = 1.47$) than when they were defined by magnitude properties ($M = 2.43$, $SD = 1.36$, $t(178) = 9.97$, $p < .0001$). Thus, participants extended properties according to the type of relations that were observed to define concepts in the context.

There were marginal but not significant differences across conditions in generalization to Both numbers; participants generalized to Both numbers slightly more in the magnitude condition ($M = 4.49, SD = 1.20$) than in the mathematical condition ($M = 4.10, SD = 1.59$, $t(178) = 1.86, p = .06$), which is a pattern also observed in our model; the model assigns higher probability to Both numbers in the magnitude condition ($M = 0.44$) than in the mathematical condition ($M = .30$). There was also a marginal but not significant difference across conditions in generalization to Random numbers; participants generalized to Random numbers slightly more in the mathematical condition ($M = 2.31$, $SD = 1.26$) than in the magnitude condition ($M = 1.98$, $SD = 0.98$, $t(178) = 1.97$, $p = .05$), which may reflect a tendency to assume uncertainty about the extent of a number's mathematical properties.

To form model predictions, each property type (magnitude or mathematical) is a hypothesis space. The prior distribution over the two spaces contains a parameter λ for the weight of the mathematical hypothesis space, with the remaining mass $(1 - \lambda)$. ing assigned to the magnitude space. We used 5,075 hypotheses adapted from the hypothesis space used in Tenenbaum et al. (2011), which consists of 24 mathematical concepts (even numbers, odd numbers, squares, cubes, primes, multiples of 3, 4,..., 12, powers of 2, 3,..., 10) and 5,050 magnitude concepts (every contiguous interval between 1 and 100). Both hypothesis spaces contained a "catch-all" hypothesis that contained all numbers between 1 and 100, to ensure that a hypothesis space never had zero probability. Tenenbaum

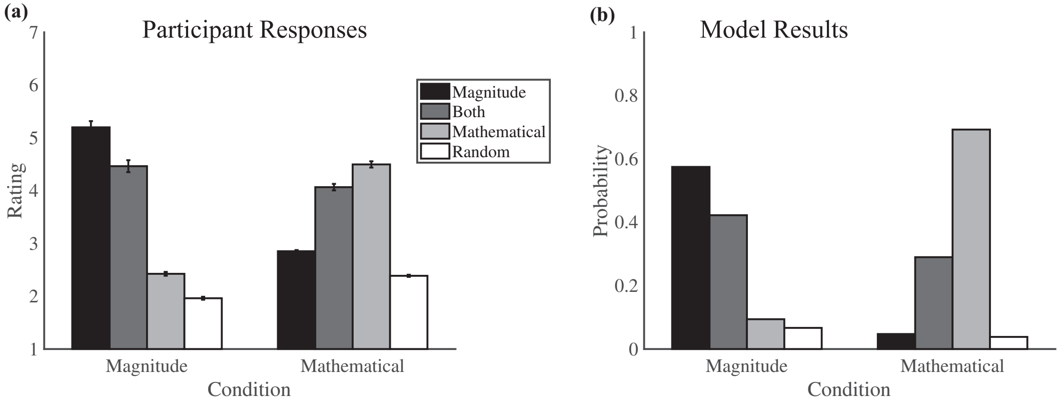

Fig. 9. (a) Average participant responses and (b) model averaging results for Experiment 3 across all three stimulus sets, aggregated by response number type. Participants and the model extend concept membership differentially, depending on which feature types (magnitude or mathematical) defined concepts in the observed context. Error bars show standard error.

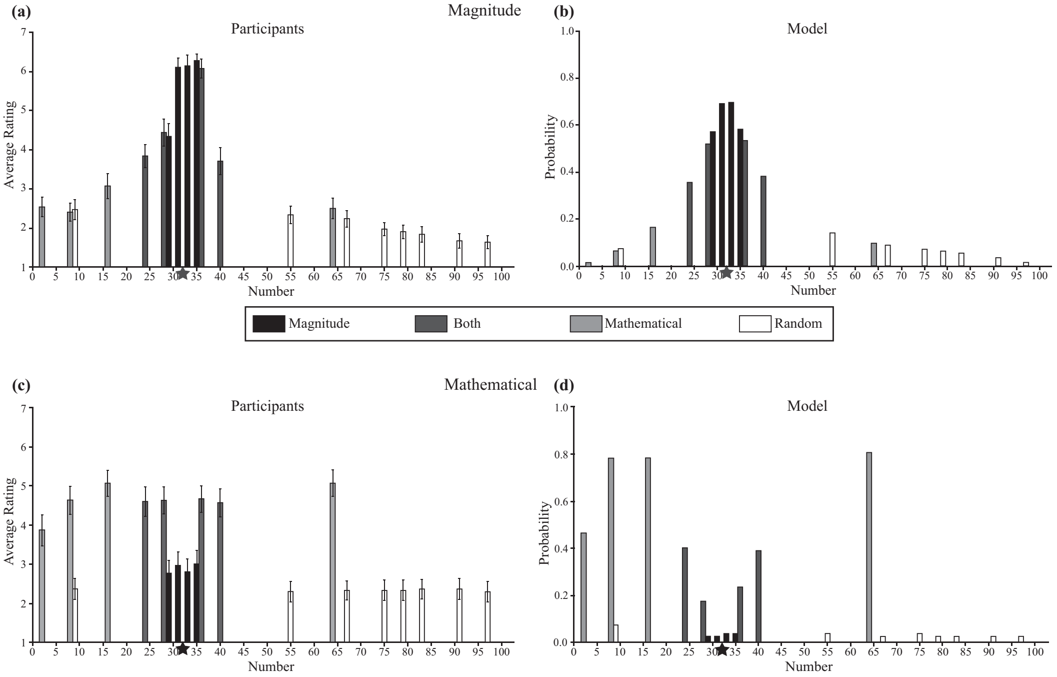

Fig. 10. Participant and model averaging responses to the test number (32, starred) in Stimulus Set 1. (a and b) Magnitude condition. (c and d) Mathematical condition. Participants' generalization patterns change depending on the previously observed concepts. The model provides a good qualitative fit to these patterns. Error bars show standard error.

(1999) originally included a third category of hypotheses: numbers that end in 0–9. We excluded these hypotheses for simplification. The prior over the mathematical hypothesis space was uniformly distributed. Following Tenenbaum et al. (2011), the prior over the magnitude hypothesis space was distributed according to an Erlang distribution, $p(h) \propto (|h|/\sigma^2)e^{-|h|/\sigma}$, where $\sigma = 10$, to capture the intuition that intervals of intermediate size are more likely concept candidates than very large or very small intervals. The size of each hypothesis is the number of numbers it contains.

Additionally, we incorporated a noise parameter, ε, which accounts for participants' uncertainty about the properties of numbers and is incorporated into the likelihood such that:

$$\begin{aligned} P(d|h) &= \frac{(1-\varepsilon)}{|h|} + \frac{\varepsilon}{100} \qquad \textit{if } d \in h \\ &= \frac{\varepsilon}{100} \qquad \textit{otherwise} \end{aligned}$$

Fig. 9(b) shows model results, averaging over number types. Fig. 10(b) and (d) show model predictions for a magnitude and mathematical test trial, not aggregating by number type. The model qualitatively captures participants' different patterns of generalization in the two conditions. Both participants and the model strongly favor generalization to numbers that relate to the example along the same type of feature as that which defined previously observed concepts—despite the fact that the single example number does not by itself imply any particular concept or concept type. Additionally, the model accounts for participant responses well quantitatively on the test trial ($r = .72$. Sarman, $r = .85$. arson), as well as on participant responses throughout the entire experiment, including both learning and test phase ($r = .83$. Sarman, $r = .87$. Prson).

We are interested in whether hypothesis space learning is better characterized by *model averaging* or *model selection*. In model averaging, inferences are made by taking a weighted average of the probabilities of each property under each hypothesis space. In model selection, only the hypothesis space with the maximum probability is used to draw inferences. We compare human generalization behavior to these two models. We also compare to a null, *non-learning* model, which performs full hierarchical Bayesian generalization with model averaging but does not update its estimate of the probabilities of each hypothesis space over the course of the experiment; the probability of each space is fixed from the start by a prior.

We compare models with Bayes Factor estimates obtained through grid estimation. We consider a grid of values for epsilon and lambda, with values bounded at 0 and 1 and intervals of 0.02. This yields a set of 2,500 parameterizations per model. We obtain the likelihood of each model and parameterization per participant with ordinal logistic regression, which permits the prediction of the participants' discrete Likert scale ratings from the model's continuous posterior probability estimates. The Bayes Factor is estimated as the ratio of the model likelihoods, averaged over all parameter combinations. We interpret Bayes Factors with the scheme of Kass and Raftery (1995) and bin Bayes Factors according to the interpreted strength of evidence.

Fig. 11 shows the percentage per participant of Bayes Factors in each bin, with comparisons between (a) model selection and non-learning, (b) model averaging and non-learning, and (c) model selection and model averaging. We first compare model selection and model averaging to the null non-learning model. In line with the interpretation of our experimental manipulations, the majority of participants are better fit by model selection or model averaging than non-learning (Fig. 11a & b), with this difference considerably stronger for model averaging. This supports the interpretation that participants are learning which hypothesis space to use in generalization through the course of the experiment.

We can now ask whether participants' hypothesis space learning behavior is better described as a model averaging or model selection strategy. We find that 66% of participants are better fit by model averaging than model selection. The calculated Bayes Factors suggest very strong evidence for model averaging over model selection for 48.89% of participants, strong evidence for 4.44%, positive evidence for 7.78%, and weak evidence for 4.44%. There is very strong evidence for model selection over model averaging for 1.67% of participants, positive evidence for 19.44% of participants, and weak evidence for 13.33% of participants. Our results show an overall preference for a model averaging strategy, with a smaller proportion of participants producing judgments consistent with a model selection strategy.

7. General discussion

Forming appropriate generalizations requires learning appropriate hypothesis spaces for generalization, regardless of whether generalization occurs across stimuli that have continuous dimensions or discrete features. In this article, we outlined two proposals for how people should learn to generalize by extending the Bayesian generalization framework: (a) *model selection*, where the hypothesis space most consistent with previously observed properties in that context is used for future generalizations, and (b) *model averaging*, where people perform a weighted average over the generalization behavior of each hypothesis space, with weights given by how consistent each hypothesis space is with previously observed concepts in the current context. We then conducted three behavioral experiments to test these computational proposals. In Experiment 1, people learned about novel words for rectangles that were consistent with selectively attending to two of four (correlated) dimensions. In Experiment 2, people learned to generalize a novel animal property in a novel context. We found that aggregate predictions supported the model averaging hypothesis. However, it was possible that this was an artifact produced by averaging over individuals that each generalize by sampling a single hypothesis space (Estes, 1956). We only received one judgment from each participant and so examining their inferences at the individual level was not possible. In Experiment 3, we tested these possibilities in detail using novel number concepts in a novel context and found converging individual-level support that people perform model averaging when they learn to generalize.

Our empirical results complement previous work by exploring how people learn the appropriate pattern of generalization for a novel blank property in a domain using the previously learned properties in that domain, and our computational work extends the Bayesian generalization framework to show how a learner can learn how to generalize. In the remainder of the article, we discuss the implications of our work for understanding

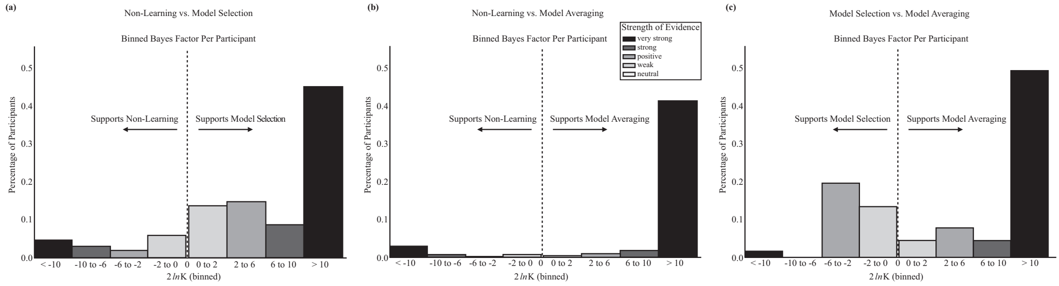

Fig. 11. Model comparisons for (a) model selection vs. non-learning, (b) model averaging vs. non-learning, and (c) model averaging vs. model selection.

human inductive reasoning, process-level concerns for the extended Bayesian generalization framework, and limitations and future directions.

7.1. Broader implications for inductive reasoning

The problem of induction has long puzzled philosophers (Goodman, 1955; Hume, 1748; Quine, 1960). Inductive inference is essential in environments that provide noisy, unreliable data. Yet it is—by definition—an ill-posed problem, and thus requires significant constraints on the learning system. The constraints required to make inductive inferences in many evolutionarily ancient cognitive systems may be hardwired. However, humans are remarkable for their ability to flexibly and reliably make inductive inferences in arbitrary domains for which the requisite constraints could not have been entirely built in, such as in science and medicine. A growing body of literature suggests that humans, including infants as young as 9 months, are able to learn certain inductive biases from limited data (Dewar & Xu, 2010; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Our work contributes to this line of research and shows that, across multiple domains, concept learning occurs simultaneously at multiple levels of abstraction. Abstract information, such as the *kind* of features (e.g., color) that labels are distributed over, is relevant across a broader array of contexts than specific information, such as the particular color (e.g., red) that correlates with a label. Thus, what is learned at this abstract level can be transferred across learning episodes, providing constraints to those new contexts and substantially accelerating the learning process. Our results show that humans are not only sensitive to this abstract information, but additionally possess an understanding of how this information constrains lower-level possibilities, thus enabling inferences that are consistent with optimal Bayesian inference over hierarchically structured hypothesis spaces. Our model synthesizes the hierarchical Bayesian approach to overhypothesis learning with the Bayesian generalization framework, providing a domain-general computational framework for learning how to generalize. Further, we present a correspondence between learning a hypothesis space and learning to selectively attend to particular stimulus dimensions. This suggests a role for learned selective attention in inductive inference: Learned selective attention may function as a filter that provides Bayesian optimal inductive biases on a learning problem. Our results emphasize the flexibility of human inductive inference and suggest that this flexibility may be due to powerful meta-learning mechanisms.

7.2. Psychologically plausibility and process-level concerns

The results of Experiments 2 and 3 support individuals learning to generalize in a domain by averaging over hypothesis spaces. From a computational standpoint, this is a remarkable feat. The extended Bayesian generalization framework, defined by Eqs. 1, 2, 3, 5, and 6, contain several summations over all hypotheses in multiple hypothesis spaces, and then a summation over the set of hypothesis spaces itself. Due to the computational complexity of the standard Bayesian generalization model, researchers have developed psychologically plausible approximations for it (Shi, Griffiths, Feldman, & Sanborn, 2010). The standard model is defined by Eqs. 1 and 2, which is only one component of the extended framework. If this component is computationally complex enough to warrant the development of sophisticated, yet psychologically valid, approximation techniques (Shi et al., 2010), how could people learn to generalize according to the extended Bayesian generalization framework? Further, previous work in category-based inductive reasoning has found robust support for people selecting a category and predicting based on it, rather than averaging (e.g., Murphy & Ross, 1994). Yet our experimental results provide support for people learning to generalize in a manner consistent with the extended Bayesian generalization framework. How do we reconcile our results with previous computational and empirical results?

First, it is important to note that our implementations of the extended Bayesian generalization framework are not intended as process-level models of human generalization. Rather, they are models within a computational-level framework. The framework is a language for specifying the goal of human generalization behavior and its normative solution (Marr, 1982). It is not a proposal of how the mind achieves this goal. Second, recent work on category-based inductive reasoning has expanded the set of circumstances where people's inferences are consistent with averaging rather than selecting (Chen et al., 2014, 2016; Hayes & Newell, 2009; Konovalova & Le Mens, 2018). For example, Konovalova and Le Mens (2018) analyzed previous experiments in this literature and found that they violated a central assumption of the tested averaging models: The features of exemplars were not conditionally independent given their category labels. When tested with categories where the features were conditionally independent given the exemplars' category labels, people act in accordance with averaging. Thus, people averaging over possibilities when performing inductive inferences may not be as psychologically implausible as previously thought.

Regardless of whether category-based induction is a selection or averaging process, a full explanation of human generalization based on the extended Bayesian generalization framework will need a process-level implementation. Although we leave a full formulation and evaluation to future work, a promising direction for developing a process-level account is to build it from a psychologically plausible, process-level model of the original Bayesian generalization framework (Shi et al., 2010). Here is a sketch of a process-level account. Consider M approximations to the Bayesian generalization framework, one for each hypothesis space, using Shi et al. (2010)'s process model. Their process model provides approximations for the generalization gradient and posterior probability of hypotheses after observing a set of objects that have some property. Thus, it already provides an account for the first term for generalizing according to the extended Bayesian generalization model (either averaging or selection, Eqs. 3 and 4, respectively). The generalization models are all independent of each other and thus can be calculated in parallel without much additional cost.

The weights for each generalization model, the second term in Eq. 5, is more complicated to calculate. Note that this is the posterior probability of a hypothesis space after observing a set of concepts in the current context. No other posterior probability is needed to calculate the term given by Eq. 5. It is plausible that this could be approximated and sequentially updated in a similar manner as Shi et al. (2010). They used an importance sampling scheme, where a set of exemplars are used to approximate the posterior, where each exemplar corresponds to a hypothesis and its fit of the current observations. Rather than hypotheses, the exemplars in the process model for the extended Bayesian generalization framework would correspond to hypothesis spaces. To calculate the fit of a *hypothesis space* to a set of observations requires summing the fit of each hypothesis within the hypothesis space. This adds an extra order of complexity to the potential process model. These calculations would suffice to be a process-level implementation of generalizing according to the extended Bayesian generalization framework. One additional possible concern is the summation over generalization gradients for each hypothesis space. This summation can be approximated by sampling hypothesis spaces from the posterior probability over hypothesis spaces and then averaging over the generalization gradients of each sampled hypothesis space. It would be challenging to dissociate this possibility from explicit model averaging, but it may be possible to do so using a cognitive load manipulation. This would be a *rational* process model, meaning that it was directly derived to be an approximation to the computational-level model (Sanborn, Griffiths, & Navarro, 2010). Note that this is just one possibility of how a psychologically plausible process-level implementation of a model in this framework could be defined. Making a concrete implementation of this model, as well as other possible process models, and empirically testing them is an important direction for future research.

7.3. Limitations and future directions

In this article, we provided a computational framework for understanding how people learn to generalize that is empirically validated across three separate experiments. However, this is only one of the first steps in a full explanation of how people learn to generalize. Within the proposed computational framework, we demonstrated that learning how to generalize is formally equivalent to learning a hypothesis space, which is a fundamental and relatively unexplored issue for Bayesian models of cognition. One assumption of our framework is that the set of hypothesis spaces is known a priori. Although it may make sense to assume that some hypothesis spaces are innate (e.g., generalizing over a one-dimensional continuous dimension), some hypothesis spaces are clearly learned (e.g., an adult's concept of numbers). We leave this question for future work.

Following other work in the Bayesian literature (including this article), one possibility would be to include a higher, more abstract level in the hierarchy. This would enable sets of hypothesis spaces to be inferred. But then how would the sets of sets of hypothesis be learned? There must be some highest level to the hierarchical model. Although this may be possible, spelling out the model and testing it empirically is a substantial undertaking. Interestingly, there are some efforts in statistics to formulate such a model (e.g., the Automatic Statistician; Ghahramani, 2015). This type of model, at least as presently implemented, would be unable to learn to generalize as people did in Experiments 2 and 3 from only the observations given to people because it lacks appropriate hypothesis spaces to evaluate, but perhaps it would if it was given the same information and observations about animals and numbers as people experience throughout their lives. Characterizing "naturally observed evidence" for domains such as animals and numbers is a difficult undertaking. An alternative approach would be to use a domain where most people are novices (e.g., radiology) and "teach" these models with the same evidence as people are given in these domains as they become experts. Would the set of hypothesis spaces of the trained model match that of an expert radiologist? When the expert radiologist learns about a new type of tumor, she would be likely to bring many of the previously learned concepts in the radiology domain to speed her generalization of the new type of tumor to other images. Would the model also learn to generalize new types of tumors in this manner?

A more concrete direction for future research is testing the extended Bayesian generalization framework's account of selective attention to other models that learn selective attention (e.g., ALCOVE; Kruschke, 1992). Although no existing results are likely to dissociate the two accounts, they are dissociable using carefully designed experiments. The Bayesian generalization framework is only sensitive to the number and range of exemplars observed in a concept, but exemplar-based approaches are sensitive to the distribution of exemplars within this range. Thus, they should make different predictions for how category learning affects selective attention when exemplars differ in how they vary within the range (e.g., uniform vs. only at the edges), which we plan to test in future work (including converging tests, such as the Garner Interference Task; Garner & Felfoldy, 1970). In addition, our approach predicts that changes in the attention given to particular dimensions should be correlated, insofar as those dimensions correspond to a particular hypothesis space. For example, in the case of rectangles, we should expect to see the weights of height and width increase or decrease together, likewise area and aspect ratio. This kind of correlated shift in selective attention is not captured by most existing models (see Navarro, 2010, and Heller et al., 2009, for extensions to the Rational Model of Categorization that also could capture correlated shifts in selective attention).

8. Conclusions

Generalization is a fundamental problem solved by every cognitive system in essentially every domain. Previous analyses of the generalization problem (Shepard, 1987; Tenenbaum & Griffiths, 2001) indicated how an ideal learner should act assuming that an appropriate hypothesis space for generalizations is known. However, how people arrive at an appropriate hypothesis space has been left as an open question. For some cognitive systems that have been fine-tuned over the course of evolution, it may be conceivable that people are born with appropriate constraints for performing generalization in that domain. However, people are capable of performing rapid generalization in arbitrary novel domains, suggesting that they are able to infer these constraints from their observations of the properties of stimuli. Focusing on the problem of learning how to generalize, our analysis shows how an ideal learner would infer hypothesis spaces from the structure of learned concepts. Our experimental results suggest that people do so in a way that is consistent with our framework, which provides a novel explanation of how people learn to generalize.

In addition to providing and empirically testing an explanation of how people learn to generalize, our results also serve an important role for Bayesian models in psychology. Bayesian models have been used to explain a range of different cognitive phenomena (Chater, Tenenbaum, & Yuille, 2006; Tenenbaum et al., 2011), but the hypothesis spaces used in the models are often hand-picked by the modeler and usually specific to the particular investigated phenomenon. This leaves open the question of how people choose the hypotheses for a given set of observed stimuli. Our framework presents an answer to this problem—a hypothesis space is used for a set of observed stimuli depending on its prior probability and how well it explains the observed stimuli. For the moment, we leave open the problem of where the hypothesis spaces come from, but we are in the process of constructing hypotheses to answer this question.

Acknowledgments

We thank Stephen Palmer, Karen Schloss, Tania Lombrozo, the Berkeley Computational Cognitive Science Lab, and the Austerweil Lab for insightful discussions, and Christina Vu, Benj Shaprio, and Julia Ying, for help administering the experiment. A preliminary version of the mathematical framework was previously presented at the 32nd Annual Meeting of the Cognitive Science Society (Austerweil & Griffiths, 2010). This research was supported by grant number FA9550-07-1-0351 from the Air Force Office of Science Research and the Office of the VCRGE at UW-Madison with funding from the WARF.

Notes

1. The following convention is used: A lowercase italicized letter denotes a single stimulus or hypothesis, an uppercase italicized letters denotes the upper limit of a dimension, a bold lowercase letter denotes a vector, an uppercase bold letter denotes a set of vectors or matrices, and a curly uppercase letter denotes a space or a set of spaces.
2. The set of hypothesis spaces may be innately endowed or constructed from a probabilistic grammar. The framework is agnostic to where the set of hypothesis spaces comes from, but it is an important question for future research.
3. The stimuli were generated in Matlab and presented to the participants using the Psychophysics Toolbox (Brainard, 1997).
4. The model's hypothesis spaces consisted of intervals over four one-dimensional hypothesis spaces (width, height, aspect ratio, and area), which are defined by Equations 6 and 7. Predictions were made by taking the difference between the

generalization probability (Eq. 3) given the concepts in the axis-aligned and diagonal-aligned conditions.

5. Only the absolute differences in generalization predicted by the model greater than 0.06 were included. Both statistical tests are one-sided Binomial sign tests.

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In J. K. Kruschke (Ed.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 534–549). Hillsdale, NJ: Lawrence: Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Austerweil, J. L., & Griffiths, T. L. (2010) In Learning hypothesis spaces and dimensions through concept learning. S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, *63*, 173–209.

Austerweil, J. L., & Griffiths, T. L. (2013). Constructing flexible feature representations using nonparametric Bayesian inference. *Psychological Review*, *120*(4), 817–851.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Special issue on "probabilistic models of cognition." *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General*, *143*(1), 227–246.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2016). Eyetracking reveals multiple-category use in induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(7), 1050–1067.

Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jeffreys, W. H., Luo, R., Paulo, R., & Loredo, T. (2007). Current challenges in Bayesian model choice. In G. F. Babu & E. D. Feigelson (Eds.), *Statistical challenges in modern astronomy IV*. (Vol 371) (pp. 224–240). San Francisco, CA: Astronomical Society of the Pacific.

Dewar, K. M., & Xu, F. (2010). Induction, Overhypothesis, and the Origin of Abstract Knowledge: Evidence From 9-Month-Old Infants. *Psychological Science*, *2*(12), 1871–1877.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*, 225–241.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65–95.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178–200.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, *37*, 730–743.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(2), 411–422.

Heller, K. A., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* 22 (pp. 727–735). San Francisco: Neural Information Processing Systems Corporation.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Konovalova, E., & Le Mens, G. (2018). Feature inference with uncertain categories: Re-assessing Anderson's rational model. *Psychonomic Bulletin & Review*, *25*, 1666–1681.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*, 4–34.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266–282.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 646–661.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. In R. Dechter & T. Richardson (Eds.), *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 324–331). Arlington, VA: AUAI Press.

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

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychological Bulletin & Review*, *10*, 517–532.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Navarro, D. (2010). Learning the context of a category. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing Systems*. Vol. *23* (pp. 1795–1803). San Diego, CA: Neural Information Processing Systems Foundation.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2011). Noncategorical approaches to feature prediction with uncertain categories. *Memory & Cognition*, *39*, 304–318.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntatic principles. *Cognition*, *118*(3), 306–338.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Ransom, K., Hendrickson, A. T., Perfors, A., & Navarro, D. J. (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 931–935). Austin, TX: Cognitive Science Society.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41.

Robert, C. P. (2007). *The Bayesian choice: A decision-theoretic motivation*. New York: Springer.

Russell, S. J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the national conference on artificial intelligence* (pp. 284–288). Philadelphia, PA: AAAI.

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2012). One-shot learning with a hierarchical nonparametric Bayesian model. Presented at the Proceedings of ICML

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*, 175–192.

Shepard, R. N. (1964). Attention and the Metric Structure of the Stimulus Space. *Journal of Mathematical Psychology*, *1*, 54–87.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shepard, R. N., et al. (1989). Internal representation of universal regularities: A challenge for connectionism. In L. Nadel (Ed.), *Neural Connections, Mental Computation* (pp. 104–134). Cambridge, MA: The MIT Press.

Shepard, R. N., & Arabie, P. (1979). Additive clutering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87–123.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13–19.

Tenenbaum, J. B. (1999). A Bayesian framework for concept learning. Ph.D. Thesis. Massachussetts Institute of Technology. Cambridge, MA.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advancesin Neural Information Processing Systems 12* (pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Tversky, A., & Gati, I. (1982). Similarity, separability and the triangle inequality. *Psychological Review*, *89*, 123–154.

Verde, M. F., Murphy, G. L., & Ross, B. H. (2005). Influence of multiple categories on the prediction of unknown properties. *Memory & Cognition*, *33*, 479–487.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.