Learning Concepts: A Learning-Theoretic Solution to the Complex-First Paradox

Nina L. Poth and Peter Brössel*†

Children acquire complex concepts like DOG earlier than simple concepts like BROWN, even though our best neuroscientific theories suggest that learning the former is harder than learning the latter and, thus, should take more time (Markus Werning). This is the complex-first paradox. We present a novel solution to the complex-first paradox. Our solution builds on a generalization of Fei Xu and Joshua B. Tenenbaum's Bayesian model of word learning. By focusing on a rational theory of concept learning, we show that it is easier to infer the meaning of complex concepts than that of simple concepts.

1. Introduction. The complex-first paradox described by Werning (2010) is as follows: substance concepts like DOG or MILK (typically associated with concrete noun words like 'dog' and 'milk') "are semantically more complex and their neural realizations more widely distributed in [the] cortex than those expressed by the other word classes in question" (1097), that is, attributive concepts like BROWN and WARM, which are typically expressed by adjectives

Received May 2018; revised December 2018.

*To contact the authors, please write to: Nina L. Poth, Department of Philosophy, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, United Kingdom; e-mail: nina.poth@ed.ac.uk. Peter Brössel, Department of Philosophy and Center for Mind, Brain, and Cognitive Evolution, Ruhr-University Bochum, Germany; e-mail: peter.broessel@rub.de.

†We are particularly grateful for the valuable comments of two anonymous reviewers. Earlier versions of this article were presented at the Salzburg Conference for Young Analytic Philosophy 2015, the German Society for Analytic Philosophy GAP.9 conference, and the Conceptual Spaces at Work 2016 conference. We would like to thank the participants at these events for their helpful feedback. Special thanks to the members of the Emmy Noether Research Group, From Perception to Belief and Back Again, and the graduate students in the Philosophy Department at Ruhr-University Bochum who gave critical feedback on an earlier draft of this article. In addition, we want to thank Ben Young for proofreading the manuscript. Research on this article has been generously supported by an Emmy Noether Grant from the German Research Council (BR 5210/1-1).

Philosophy of Science, 87 (January 2020) pp. 135–151. 0031-8248/2020/8701-0007$10.00
Copyright 2020 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

or abstract noun words: 'brown(ness)', 'warm(th)'.[1] Thus, in the light of our best neuroscientific theories, one would expect that learning substance concepts would be more difficult than learning attributive concepts and thus take not only more effort but also more time. Yet, surprisingly, we observe that young concept learners first understand the meanings of concrete noun words and then subsequently acquire the meanings of adjectives or abstract nouns (Nelson 1973; Jackson-Maldonado et al. 1993; Sandhofer and Smith 2007). This finding presents a puzzle for cognitive scientists and philosophers of mind and language: How is it possible that children learn the meanings of concrete nouns like 'dog' earlier than the meanings of adjectives like 'brown' that denote abstract attributes?

In this article we present a novel philosophical solution to the complex-first paradox. Gärdenfors (2019) also outlines an idea for a solution to the complex-first paradox based on his theory of covariance detection mechanisms in conceptual spaces. In section 2 we introduce Werning's complex-first paradox and clarify what kind of solution we want to provide in this article. In particular, we advocate a change of perspective—instead of focusing on the neuroscientific perspective of concept acquisition, we focus on theories of rational word and concept learning. Thus, the explanation we want to provide is on the computational level instead of the implementation level (Marr 1982) and concentrates on the mechanism for learning concepts, instead of the outcomes of this learning mechanism. That is to say, instead of focusing on the implementation (i.e., the neural realization) of concepts, we focus on the workings of the learning mechanism from an algorithmic/computational-level perspective and ignore the implementational level of this learning mechanism. (In this article, in general we ignore questions concerning the implementation of the underlying learning mechanism.) In section 3, we introduce one specific computational-level theory of rational word learning introduced by Xu and Tenenbaum (2007). In section 4 we demonstrate how on the basis of Xu and Tenenbaum's model the paradox can be solved. In section 5 we conclude with a summary and a critical perspective outlining future directions of research.

2. The Complex-First Paradox and the Different Levels of Explanation.

According to Werning (2010), the divergence between the theoretical predictions and children's actual word- and concept-acquisition behavior can

1. Following Werning, as well as philosophers such as Millikan, we use the term 'substance concepts' simply to refer to those concepts that represent "(1) stuffs (gold, milk), (2) real kinds (cat, chair), and (3) individuals (Mama, Bill Clinton, the Empire State Building)" (Millikan 1998, 55). No commitment to a specific metaphysics comes with using the term. When we refer to terms, we follow the usual convention and use single quotation marks. When we refer to concepts we use small capitals.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

be traced back to a set of propositions, which are each individually plausible but, taken together, lead to a paradox in the sense that they are "apparently inconsistent" and, thus, point at "an explanatory deficit in linguistic theory" (1096). The propositions that lead to the paradox concern concepts, their neural realization, and the time and effort it takes to acquire them.

1. The meanings of concrete nouns are substance concepts.
2. Substance concepts are semantically more complex and their neural realizations more widely distributed in the cortex than those expressed by abstract attributive concepts like GREEN.
3. For a cortically implemented syntax-semantics interface, the more widely distributed a concept's neural realization is, the more effort it takes to establish a link between the concept and some lexical expression thereof.
4. In ontogeny and phylogeny, capabilities demanding more effort, all other things being equal, can be expected to develop and, respectively, evolve later than those demanding less effort.
5. The meanings of concrete nouns, in ontogeny and (probably) phylogeny, are acquired earlier than those of many—perhaps even all—other word classes, in particular, attributive concepts such as GREEN.[2]

Werning discusses each of the above propositions in detail. For each of them he provides ample empirical evidence that supports the respective empirical claim, even though he is, of course, aware that the available empirical evidence alone cannot establish those propositions beyond reasonable doubt. Our problem does not concern the credibility of these propositions but the perception of them as a paradox. Of course, Werning is also well aware that strictly speaking 1–4 do not imply the falsity of 5 (and thus the word 'puzzle' would perhaps have been more appropriate than 'paradox'). However, they lead us, according to Werning, to expect that the complex substance concepts, which we express by concrete nouns, would be acquired later than abstract attributive concepts. Once we acknowledge that these propositions are not inconsistent, we recognize this puzzle as what it is: a challenge. The challenge is to provide an explanation for why, despite propositions 1–4, substance concepts are learned earlier than attributive concepts.

To answer that challenge, we propose adopting a new perspective in two respects. First, Marr (1982) distinguishes three levels of analysis in cognitive science: the computational, the algorithmic/representational, and the implementational. On the computational level we describe what function a cognitive

2. Although nouns are commonly learned earlier than many other word classes, the nouns-prior-to-adjectives phenomenon has received particular attention in the literature on word learning (see, e.g., Gasser and Smith 1998).

https://doi.org/10.1086/706077 Published online by Cambridge University Press

mechanism computes (i.e., the input-output function characterizing the mechanism). On the algorithmic level we describe how exactly this input-output function is computed, that is, which representations are involved and how these representations are processed by the mechanism. Finally, on the implementational level we describe how the cognitive mechanism and its processes are physically realized. As one can see from the premises, Werning's paradox is obtained by focusing on the level of implementation of our conceptual capacities. Our proposal is to focus on the computational level and to study the input-output function that characterizes the learning mechanism.

Second and more importantly, Werning ignores the mechanism that enables us to learn words and concepts. In particular, we are still missing a computational-, algorithmic-, or implementational-level account of how the brain and the mind learn concepts. Werning concentrates instead on the effort and time it takes to physically implement the neural realizations corresponding to individual concepts. Thus, he concentrates on the implementation of the outputs of the learning mechanism, but he does not discuss the implementation of the learning mechanism. He argues that, in light of all we know about the implementation of our concepts, establishing the neural realizations of substance concepts should take more time and effort in comparison to the establishment of the neural realizations of abstract attributive concepts. Werning ignores the time and effort it takes to learn these words and concepts, and he ignores the learning mechanism. We propose to study exactly this. However—and this follows from our earlier constraint—we will also discuss the learning mechanism from the computational-level perspective.

Our twofold shift of focus comes at a price: our approach will be silent on which kind of concept is learned first in phylogeny. To determine which kind of concept is learned first in phylogeny, we would need to study the phylogeny of the human sensory system and the phylogeny of the learning mechanism for learning concepts. But this is not what we are doing. We want to meet the challenge of the complex-first paradox by focusing on the computational aspects of the mechanism of concept learning as we currently find it in infants.

To sum up, we seek to resolve the complex-first paradox, not by arguing directly against one of the five propositions but by providing a computational-level account of concept learning. This novel perspective will allow us to discover a more promising alternative, avoiding the complex-first paradox by explaining why children learn substance concepts faster than abstract attributive concepts.

3. Concept Learning as Bayesian Inference

3.1. Xu and Tenenbaum's Account of Concept Learning. It is helpful to understand word and concept learning as an induction task. The child receives

https://doi.org/10.1086/706077 Published online by Cambridge University Press

evidence of how a specific word or label is used to refer to certain objects that it can visually perceive.[3] On this evidential basis the child has to form a hypothesis about what concept is associated with that word or label, and on the basis of such a hypothesis children can then categorize further objects as falling under that label. On the picture we propose, concept learning goes hand in hand with word learning. Both word and concept learning are crucial steps to linguistic communication, and both are completed if and only if an appropriate concept has been linked with the given label. More specifically, an agent has linked an appropriate concept *C* with a label *L* (as used by the majority of English speakers) if and only if on the basis of *C* the agent is able to solve the corresponding categorization task: to correctly subsume a new object under the label *L* if and only if the majority of English speakers also categorizes the object under *L*.[4]

Before introducing the framework in a more formally rigorous way, let us introduce the key idea at hand with an illustrative example taken from Xu and Tenenbaum (2007). Suppose a father wants to teach his child the concept of DALMATIAN. He does so by pointing at instances of Dalmatians while uttering the label 'fuxur'. This is the evidence the child receives. Then in a second step the child has to infer a hypothesis about what concept is linked with the label 'fuxur' on the basis of this evidence. Especially since the child only receives positive evidence, that is, positive instances for the given label, there are many hypotheses that are consistent with that evidence. The label 'fuxur' might mean DALMATIAN, DOG, MAMMAL, BLACK-WHITE DOTTED, and so on. The child's task is now to determine the most likely or plausible of these hypotheses, and via trial and error (with the help of corrections by the father) the child will learn that the label 'fuxur' is paired with the concept DALMATIAN.

3. In this article we concentrate exclusively on visual perception. We want to study the categorization of external world objects on the basis of perceiving these objects and this is best understood for visual perception. In particular, whether the objects of auditory and olfactory perception are external world objects (instead of only their smell and their sound) is often disputed in the literature (O'Callaghan 2016). We do not want to burden the current investigation with philosophical and psychological discussions about the objects of olfactory and auditory perception. Nevertheless, we believe that a similar story can be told for learning concepts with the help of all forms of perception. To tell this story, we would need to introduce our theory of the content of perceptual experiences first, and because of space limitations, we cannot do this in the current article. For an outline of our theory of the content of perceptual experiences, see Brössel (2017).

4. Our assumption that one has learned and thus possesses a concept if and only if one is able to solve a corresponding categorization task is not undisputed in philosophy. It is inadequate for more abstract concepts like 'war' and 'love'. For this reason we restrict our account of concept learning to perceptual concepts. In this respect, we also assume that children first have to learn the perceptual concept of a dog before they can modify it to also contain conceptual truths about the ancestry and diet of dogs.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

More formally, we can frame the induction task that children are confronted with as follows:[5]

1. *Evidence*: A single piece of evidence for a child consists of an observed object o and the (typically uttered) label L, with which it has been tagged (T). The complete available evidence concerning some label L then consists of an ordered sequence of so-far-observed object-label pairs:

$$e_i^L = T(o_i, L)$$
$$E_n^L = (e_1^L \wedge \ldots \wedge e_n^L).$$

2. *Hypotheses*: The set of hypotheses $\mathcal{H}_L$ among which the child has to choose. The hypotheses have the form $H_{(C_i,L)}$ and state that some specific concept C_i is paired with the label L:[6]

$$\mathcal{H}_L = \{H_{\langle C_i,L\rangle} | \text{where } C_i \text{ is a concept}\}.$$

As is standard in Bayesian cognitive science, Xu and Tenenbaum (2007) treat the relevant learning mechanism (on the computational level) as a Bayesian learning mechanism. On Xu and Tenenbaum's account, children learn the meaning of labels (i.e., concepts) as if they were Bayesian agents, and thus as if they were using Bayes's Theorem to choose the winning hypothesis.

Theorem 1. Bayes's Theorem in the context of concept learning:

$$\Pr(H_{\langle C_i,L\rangle} | E_n^L) = \frac{\Pr(E_n^L | H_{\langle C_i,L\rangle}) \Pr(H_{\langle C_i,L\rangle})}{\sum_{C_j} \Pr(E_n^L | H_{\langle C_j,L\rangle}) \Pr(H_{\langle C_j,L\rangle})}. \quad (1)$$

As a result, the most probable hypothesis $H_{\langle C_i,L\rangle}$ will be the one that exhibits a good mixture between a high prior probability, $\Pr(H_{\langle C_i,L\rangle})$, and a high likelihood

5. For simplicity we assume the concept learner acquires evidence with respect to only one label.

6. Nothing here commits us to the view that the hypothesis space and concepts available to an agent are innate. This is a computational-level account of word and concept learning, and we only model the task of concept learning and the solutions to which the children come. In this article we remain silent on how children solve the task, i.e., on how exactly the brain calculates the computational function we are describing and how it is implemented. Colombo (2018) is one of the best papers available on the question of what role Bayesian modeling can play for the empiricism-nativism debate.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

of the evidence given the hypothesis, $\Pr(E_n^L|H_{\langle C_i,L\rangle})$. How, though, can these probabilities be determined for each of the hypotheses in $\mathcal{H}_L$?

Xu and Tenenbaum (2007) concentrate on providing an account of how the likelihoods are to be determined in the given setting. For determining $\Pr(E_n^L|H_{\langle C_i,L\rangle})$ we need to explain how likely the observational evidence of n objects falling under the label L is, under the assumption that the label L is associated with the concept C_i. Accounting for the likelihoods is the key step in explaining word and concept learning in general and resolving the complex-first paradox in particular. The likelihoods encode how children learn words and concepts; after all, the priors are fixed before any learning takes place, and they are not revised in light of the incoming evidence. Only the likelihoods change when the evidence comes in. In order to understand the basic idea of their proposal, let us return to the above example of how the child learns that the concept DALMATIAN is to be linked with the label 'fuxur'. Suppose a child has the following evidence, $E_4^{\text{fuxur}} = \wedge_1^4 e_i^{\text{fuxur}}$, where each of the e_i^{fuxur} has the form $\langle o_i, \text{fuxur}\rangle$ and the observed objects o_i are as follows:

o_1 a black-on-white spotted Dalmatian
o_2 a black-on-white spotted Dalmatian
o_3 a brown-on-white spotted Dalmatian
o_4 a black-on-white spotted Dalmatian

Such evidence already excludes many hypotheses in the hypothesis space. In particular, it excludes hypotheses that pair the label 'fuxur' with a concept that does not include Dalmatians, such as CAR, CAT, BLACK, and so on. However, there are still many candidate concepts that might be paired with the label, for example, the hypotheses:

H_{Dal} 'fuxur' means DALMATIAN: ⟨DALMATIAN, 'fuxur'⟩
H_D 'fuxur' means DOG: ⟨DOG, 'fuxur'⟩
H_M 'fuxur' means MAMMAL: ⟨MAMMAL, 'fuxur'⟩
H_A 'fuxur' means ANIMAL: ⟨ANIMAL, 'fuxur'⟩
H_b 'fuxur' means BLACK OR BROWN SPOTS ON WHITE: ⟨BLACK OR BROWN SPOTS ON WHITE, 'fuxur'⟩

Intuitively, the child should believe most strongly that 'fuxur' means DALMATIAN (i.e., the hypothesis H_{Dal}). This is certainly the hypothesis that most adults would accept upon getting the evidence E_4^{fuxur}. Given our assumption that children learn as if they were Bayesians, we have to ask how to set likelihoods in such a way that children can learn that 'fuxur' means DALMATIAN. To answer that question Xu and Tenenbaum (2007) propose what they call the size principle.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

The Size Principle:[7]

$$\Pr(E_n^L | H_{\langle C_i, L \rangle}) = \left(\frac{1}{\text{size}(C_i)} \right)^n, \tag{2}$$

if all n objects labeled L fall under the concept C_i and 0 otherwise.

As already explained, the evidence E_n^L specifies for n objects that they have been labeled L. The size principle states that the probability of this observation being true given that the label L is associated with the concept C_i is, if all n objects fall under that concept, proportional to the size of the concept C_i. As a first pass at understanding the size principle, the size of a concept is the number of objects falling under the concept, that is, the cardinality of the concept's extension. The size principle is supported by the assumption that if the father wants to teach the use of the label L, which is associated with the concept C_i, he draws the n examples randomly from the set of all objects in the concept's extension. For one object, the probability of E_1^L given that C_i is associated with L is $1/\text{size}(C_i)$. For two objects o_1 and o_2, the relevant probability is $(1/\text{size}(C_i))^2$, because the two draws are independent of each other. With the help of this size principle, Xu and Tenenbaum (2007) can explain what they term the "suspicious coincidence" effect. To explain this, let us return to our example of the four observed Dalmatians. If the four exemplars are a random sample, it is highly unlikely that we will observe all and only Dalmatians being labeled L if in fact the label is associated with the concept 'dog' or the concept 'black or brown spots on white'. Therefore, as a general rule of inference, given few and very similar exemplars for a label L, it is more plausible to infer that the label is associated with a concept with a smaller (as opposed to larger) extension.

Of course, taken literally this story must be false, as children do not have access to the true size of the extension of a given concept. However, as Xu and Tenenbaum (2007) elaborate, children do have access to a psychological proxy, the similarity between the observed and labeled objects. (Assuming that the size principle is grounded in similarity judgments, which are ultimately grounded in our perceptual system, could also serve as an explanation for why, in concept learning, all or almost all humans learn as if the size principle were true.) Concerning this issue there are certainly more questions that have to be asked and answered. Unfortunately, Xu and Tenenbaum

7. We make some changes to the original formulation from Xu and Tenenbaum (2007). In particular, Xu and Tenenbaum use the following formulation: $(1/\text{size}(h))^n$. It seems to us instead that eq. (4) is more accurate because hypotheses are indexes for concept-label pairs and thus cannot have a size themselves. It seems correct to us to take the size to be instead about the narrowness of a set of possible objects that is indicated by a concept, and replace size(h) with size(C_i), for the ith concept.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

(2007) do not address these questions, but we ask and answer some of them in Poth (2016), Brössel (2017), and Brössel and Poth (2019). Roughly the idea that we present there is that—following philosophers such as Gärdenfors (2000) and psychologists such as Shepard (1987)—we assume that perceptual concepts correspond to regions in a phenomenal similarity space. On this basis, we spell out a second understanding of Xu and Tenenbaum's size principle. Instead of identifying the size of a perceptual concept with its extension, we propose to identify the size of a perceptual concept with the size of the corresponding region in such a phenomenal similarity space. In line with Gärdenfors's and Shepard's interpretations, we interpret the size of such a region to be a proxy for how perceptually similar, on average, a concept's instances are. We refer to this definition of 'size' in the current context as the narrowness of a set of possible examples for a concept.

This second understanding of the size principle obviously suggests predictions concerning concept learning different from the first. Concepts such as DICE and FAIR DICE have different extensions, but they are perceptually almost indistinguishable. Thus, learning DICE and FAIR DICE from positive examples alone should, according to our preferred understanding of the size principle, result in almost identical learning behavior. More generally, such a reading of the size principle predicts a similar learning behavior for both concepts because DICE and FAIR DICE correspond to roughly the same regions in phenomenal similarity space. In contrast, the first understanding would predict a huge difference in learning behavior since the extensions of *fair dice* (or precision dice as used in casinos) is much smaller than the extension of *dice*. Another relevant comparison would be between the concepts INSECT and MAMMAL. The extension of all insects is without doubt much greater than the extension of all mammals. But concerning the perceptual in-between-similarity comparison between insects and mammals this question cannot be answered so clearly. Thus, again we would expect roughly similar learning speeds for the concepts INSECT and MAMMAL if we presupposed the second similarity-based understanding of the size principle, and we would expect vastly different learning speeds if we presupposed the cardinality-based reading of the size principle. The formal details and the philosophical advantages as well as the empirical predictions and the existing empirical support for our preferred reading of the size principle is discussed in Brössel and Poth (2019). Here we focus on solving the complex-first paradox, and for this purpose we discuss only examples for which the difference between both understandings of the size principle can be neglected.

3.2. Empirical Support for Xu and Tenenbaum's Account. To test their Bayesian model, Xu and Tenenbaum (2007) conducted a series of experiments with adults and 3–4-year-old children. Before discussing their results, let us focus on what the theory predicts and return to our worked example.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

Intuitively, we would say that the set of *Dalmatians* is narrower than the set of *dogs*. In addition, Dalmatians are typically much more similar to one another than dogs are to one another. Thus, whatever understanding of the size principle one presupposes here, for the two alternative hypotheses H_{Dal} and H_D the corresponding likelihoods should favor H_{Dal} over H_D (if we suppose that we observe one object labeled 'fuxur' that is an instance of both concepts). The likelihoods after one observation and after four observations look like this.

$$\Pr(E_1^{\text{fuxur}}|H_{\text{Dal}}) = \left[\frac{1}{\text{size(Dalmatians)}}\right]^1 > \Pr(E_1^{\text{fuxur}}|H_D) = \left[\frac{1}{\text{size(dogs)}}\right]^1 \quad (3)$$

$$\Pr(E_4^{\text{fuxur}}|H_{\text{Dal}}) = \left[\frac{1}{\text{size(Dalmatians)}}\right]^4 > \Pr(E_4^{\text{fuxur}}|H_D) = \left[\frac{1}{\text{size(dogs)}}\right]^4. \quad (4)$$

Given these likelihoods we can ask what impact the difference in likelihoods has on the difference in conditional probabilities $\Pr(H_{\text{Dal}}|E_n^{\text{fuxur}})$ and $\Pr(H_D|E_n^{\text{fuxur}})$ if measured by the posterior ratio $\Pr(H_{\text{Dal}}|E_n^{\text{fuxur}})/\Pr(H_D|E_n^{\text{fuxur}})$:

$$\frac{\Pr(H_{\text{Dal}}|E_n^{\text{fuxur}})}{\Pr(H_D|E_n^{\text{fuxur}})} = \left[\frac{\Pr(E_1^{\text{fuxur}}|H_{\text{Dal}})}{\Pr(E_1^{\text{fuxur}}|H_D)}\right]^n \times \frac{\Pr(H_{\text{Dal}})}{\Pr(H_D)}. \quad (5)$$

We know that $\Pr(E_1^{\text{fuxur}}|H_{\text{Dal}})/\Pr(E_1^{\text{fuxur}}|H_D) > 1$, due to the size principle and the difference in size of the concepts DALMATIAN and DOG. Clearly, one instance of the label 'fuxur' might not be enough to lead to an important difference in the conditional probabilities $\Pr(H_{\text{Dal}}|E_1^{\text{fuxur}})$ and $\Pr(H_D|E_1^{\text{fuxur}})$. However, obtaining more evidence leads to important differences—the difference between the left and the right side in equation (3) is smaller than the difference between the left and the right side in equation (4). Indeed, the weight of the likelihood ratio $\Pr(E_n^{\text{fuxur}}|H_{\text{Dal}})/\Pr(E_n^{\text{fuxur}}|H_D)$ for determining the posterior ratio $\Pr(H_{\text{Dal}}|E_n^{\text{fuxur}})/\Pr(H_D|E_n^{\text{fuxur}})$ grows exponentially, the more instances of the label 'fuxur' we encounter. Thus, the size of the concept—ignoring for the moment the question whether the size of the concept is understood in terms of the size of the set of objects falling under the concept or in terms of the phenomenal similarity of the (observed) instances—determines, according to Xu and Tenenbaum's (2007) account, how fast we learn one concept as compared to another concept.

Correspondingly, Xu and Tenenbaum's account of word and concept learning predicts two general effects that we should observe when we study the word-meaning generalization behavior of children and adults. First, equation (3) predicts that children and adults will prefer smaller concepts like DALMATIAN over broader ones like DOG given an equal number of examples. Second,

https://doi.org/10.1086/706077 Published online by Cambridge University Press

equations (2) and (5) predict that the more compatible examples learners have observed for the alternative concepts, the more strongly they will express their preference for learning smaller concepts like DALMATIAN over larger concepts like DOG. In other words, with additional compatible examples, learners should restrict their generalizations even more toward smaller concepts.

To test these predictions, Xu and Tenenbaum (2007) did a word-learning experiment with a training and a test phase. In the training phase, they presented adults or children with one or three labeled pictures of objects. For instance, they showed one or three pictures of a Dalmatian together with the label 'fuxur'. In the test phase, Xu and Tenenbaum asked subjects to generalize the label (e.g., 'fuxur') to other objects from an array with a variety of different kinds of objects (e.g., an array with other kinds of animals, vegetables, or vehicles). The subjects' extent of generalization was then used as a measure of their understanding of what the label means.

To test the second prediction, Xu and Tenenbaum manipulated the number of the examples (e.g., one or three) in the training phase. To test the first prediction, they manipulated the range spanned by the objects within the three-example trials. In particular, they presented subjects with three labeled objects from the same subordinate category (e.g., three pictures of a Dalmatian), three labeled objects from the same basic-level category (e.g., one picture apiece of a Dalmatian, a terrier, and a mutt), or three labeled objects from the same superordinate category (e.g., one picture of each a Dalmatian, a toucan, and a pig).

Xu and Tenenbaum's (2007) results confirm both predictions. First, learners restricted their generalizations of a word toward the smallest consistent category level in the one-example trials. For instance, given one Dalmatian labeled 'fuxur', subjects were more likely to label other Dalmatians rather than other dogs from the test array with 'fuxur'. Second, learners generalized more gradually in one-example trials and restricted their generalization behavior more strongly and distinctively when they had been given three compatible examples. For example, when learners saw one Dalmatian called 'fuxur', they were less likely to pick out any other dogs or any other nondog animals in the array. In that case, they were more likely to pick out the other Dalmatians. But given additional pieces of evidence in the training phase (e.g., two other Dalmatians called 'fuxur'), learners would typically pick out all and only Dalmatians in the test array, thereby indicating that they believed more strongly that 'fuxur' means DALMATIAN and neither DOG nor ANIMAL.

From the perspective of the size principle, it is to be expected that learners would restrict their generalization behavior to the smallest consistent category level in the test array. Likewise, we would expect that subjects' generalization behavior would become more distinct with multiple examples because, following equation (5), the evidence in the one-example cases should obtain a lower weight than in the three-example cases, and therefore it should be easier for

https://doi.org/10.1086/706077 Published online by Cambridge University Press

a Bayesian learner to rule out competing alternative hypotheses with three examples than with one example. Thus, taken together, Xu and Tenenbaum's results are consistent with the predictions from the Bayesian model and thereby illustrate that it is useful to think of word and concept learning as a rational-inductive inference task. The next section establishes that this characterization of word and concept learning is useful for making sense of the complex-first paradox.

4. The Explanation of the Complex-First Paradox and Possible Objections. Let us assume, with Xu and Tenenbaum (2007), that on the computational level word and concept learning can be understood as Bayesian inference and that something like the size principle is in place. Building on these assumptions we can now return to the main goal of this article: to explain the complex-first paradox.

Recall that the complex-first paradox consists in the fact that children learn complex substance concepts like DOG or MILK (typically associated with concrete noun words) earlier than simple, attributive concepts like BROWN and WARM (typically expressed by adjectives or abstract noun words). We have to explain why children learn the former type of concept faster than the latter type of concept. From the Bayesian perspective that we are adopting here, since we have not made any assumptions about the priors, we have to explain this with the help of likelihoods. Following Xu and Tenenbaum (2007), we assume that likelihoods obey the size principle. The less comprehensive a concept is, the higher the likelihood. Thus, to explain the complex-first paradox, we need only establish one additional claim, that is, that often complex substance concepts like DOG and MILK are smaller in size than simple, attributive concepts like RED and WARM.

Let us compare the concepts DOG and BROWN as an example. The set of all dogs is much narrower than the set of all brown things. The number of brown insects (which is much smaller than the number of brown things) is already much larger than the number of dogs. Thus, even if we understand size only in terms of the concept's extension, DOG is already much smaller than BROWN INSECT. And we come to the same conclusion when we use the similarity-based interpretation of the size of these concepts. Typically, objects that fall under an attribute concept like BROWN are similar to one another in only one respect. They are similar in color but might be extremely dissimilar to one another in all other respects: shape, height, weight, movement, sound, and so on. In contrast, objects that fall under substance concepts are often similar to one another in many respects. In the case of basic-level concepts such as DOG, for example, various instances of that concept are similar in shape, color, sound, movement, and behavior. Often the objects falling under a substance concept are not extremely similar to one another in one specific respect but considerably similar regarding all respects in which we can perceptually discriminate. We should

https://doi.org/10.1086/706077 Published online by Cambridge University Press

therefore expect instances of BROWN to be overall less similar to one another than instances of DOG. Thus, also under a similarity-based interpretation, DOG has a smaller size than BROWN.

We can thus assume, ceteris paribus, that (subordinate and basic-level) substance concepts are smaller than attributive concepts, and these concepts can be learned faster because of the size principle. (It is important to note that one probably cannot establish this for all such concept comparisons. The substance concept METAL is arguably similar in size when compared with the attribute concept METALLIC, and the various instances of metal and metallic are probably equally similar. Similarly for superordinate concepts like MAMMAL, we can expect that they display only a very small overall similarity and that they are not much more similar to one another than instances of the attributive concept HAIRY.) With this last assumption at hand, we are now in the position to explain the complex-first paradox.

P1. Ceteris paribus, (subordinate and basic-level) substance concepts are narrower than attributive concepts.
P2. Learners will learn narrower concepts earlier than less narrow concepts given comparable numbers of observed instances of the given concepts.
C. Learners will learn complex concrete noun concepts earlier than simple attributive concepts, given comparable numbers of observed instances of the given concepts.

The conclusion of this argument shows what we want to show: one can expect that children will learn substance concepts earlier than attributive concepts. The reason is that the former concepts are typically narrower than the latter, which makes them easier to learn for children because of the size principle. This resolves Werning's puzzle.

There are two empirical findings that are seemingly in conflict with our explanation of Werning's complex-first paradox and the size principle: the basic-level bias and the shape bias. The basic-level bias is an inductive preference for basic-level concepts over superordinate and subordinate concepts. The shape bias is an inductive preference for concepts that have high variability with respect to attributes like, for example, color but not with respect to shape.

Let us discuss the basic-level bias first. In particular, Rosch and Mervis (e.g., Rosch 1975; Rosch and Mervis 1975) have argued that concepts are structured in terms of the family resemblances—correlations among shared attributes—of their instances. Instances of narrower concepts are more similar because they resemble one another in terms of more attributes than instances of broader concepts, which are less specific. Rosch and Mervis use their structural account to justify the taxonomy of concepts with different sizes on which we rely here. Narrower concepts are positioned at the subordinate

https://doi.org/10.1086/706077 Published online by Cambridge University Press

level, which is the least inclusive level. Broader concepts combining the least similar instances are positioned at the superordinate level, which is the most inclusive level. Basic-level concepts are in between. The problem for our explanation of the complex-first paradox is that a number of studies show that preschoolers and adults prefer to generalize concepts on the basic level over the subordinate level (Horton and Markman 1980; Mervis and Crisafi 1982). However, basic-level concepts are relatively broad and typically span over a range of subordinate concepts. The evidence for the basic-level bias presents a challenge to our account because the size principle would predict evidence for a subordinate-level bias instead. Rosch and Mervis believe that basic-level concepts are preferred because they are the inductively most useful type of concept of the three, since they maximize the similarity among instances of a single concept while minimizing their similarity to instances of other concepts.

As already said, the shape bias is a preference for concepts that have high variability with respect to attributes like color but not with respect to shape. Indeed, among the earliest concepts that children learn, we find many concepts for artifacts like chair and table, for which color is in fact often irrelevant (Smith et al. 2002). The size principle stands in conflict with the shape bias since it predicts a preference for narrower concepts with respect to both shape and color. If children immediately generalize to concepts for which color is not as relevant as shape, they would obey the size principle only with respect to shape but not with respect to color. This bias has been found in children as young as around 24 months (e.g., Landau, Smith, and Jones 1988; Jones, Smith, and Landau 1991).

What these findings suggest is that children bring a lot of prior knowledge (if we want to call these biases 'knowledge') to the word-learning task. Where do these inductive biases come from? One option is that both of these very different biases are innate. Against this proposal, Xu and Tenenbaum (2005, 2384) provide evidence that 4-year-olds, in contrast to adults, do not show a basic-level bias in the Bayesian word-learning paradigm that we have outlined above. If a basic-level bias were innate, one would need to explain why children unlearn it at age 4 and relearn it when they become adults. Using these results and further evidential data, Xu and Tenenbaum (2007, 262) suggest that the basic-level bias is learned: with more experience, children's empirical prior for "medium-sized" basic-level concepts gets higher and higher because most concepts that parents use and teach in everyday life are medium-sized, basic-level concepts.

Similarly, Smith et al. (2002) provide evidence for the claim that the shape bias is also acquired. They designed a word-learning task to train 17-month-old children with labels for objects that had the same shapes but varied in size, texture, and color. They then tested whether children would generalize the label to novel objects that matched in shape but varied in another attribute.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

Initially children did not display the bias, but at the end of the training they did. Given that typically children do not show the bias before they are 24 months old, Smith et al.'s results indicate that the shape bias can be developed during experimental training. Using further evidence, Smith and Samuelson (2006) conclude that the shape bias is acquired and that it develops over time, depending on the perceptual properties of artifacts. Their own attentional learning account explains this development on the basis of "statistical regularities and the higher order generalizations that emerge from them" (1339), but they do not provide a model that describes how the shape bias is acquired by young children. Kemp, Perfors, and Tenenbaum (2007) provide such a model. They argue that the shape bias is the result of a process of learning to learn concepts. After learning the first couple of concepts, children learn that many concepts allow for a greater variability of colors. Their approach assumes the framework of hierarchical Bayesianism, and they show that it is compatible with a wide range of data. If children learn the shape bias, they learn to learn concepts that allow for a variability with respect to color. From this perspective, the shape bias takes the role of an empirically acquired prior and is distinct from the size principle.

This section showed that one of the most developed and successful (computational-level) theories of concept learning—Xu and Tenenbaum's (2007) Bayesian account—can also explain the complex-first paradox. But there are various empirical phenomena that a successful theory of concept learning needs to be able to explain apart from the complex-first paradox. The shape bias and the basic-level bias belong to these phenomena. This section also showed that these biases are compatible with our explanation of the complex-first paradox because they are learned. The complex-first paradox, however, is more basic and asks why the first concepts that children learn, even before they acquire any of the biases, are concrete noun concepts. In this sense, Werning's complex-first paradox is the first of many challenges that a theory of concept learning needs to explain.

5. Conclusion and Future Perspectives. Our current best neuroscientific theories of concept learning are at odds with the empirical evidence on word learning, according to Werning (2010). According to our best neuroscientific theories, children should learn concrete noun concepts like DOG later than abstract attributive concepts like BLUE, but children in fact learn such concrete nouns earlier than abstract attributive ones. In this article, we have presented a novel solution to Werning's complex-first paradox.

Our solution to the paradox starts from Marr's (1982) level of computational theory and specifies concept learning as a computational problem. Xu and Tenenbaum's (2007) Bayesian model of word learning is a worked example for this idea. In Bayesian word learning, the size principle that is characteristic of the concept-learning mechanism is the input-output function.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

The size principle predicts that, ceteris paribus, children favor smaller over larger concepts because they are associated with greater likelihoods in the word-meaning inferences. In this article, we have suggested that this rationale also explains why children learn the meanings of concrete nouns earlier than they learn the meanings of abstract attributes. Roughly, concrete noun concepts like DOG are smaller than abstract attributive concepts like BROWN. Thus, concrete noun concepts are easier to infer because the probabilities that they are associated with are greater than the probabilities associated with their abstract attributive alternatives. We have argued that the Bayesian model of word learning can thus be expanded toward a solution to the complex-first paradox. We conclude that instead of focusing all attention on the neural implementation of concepts, it is important to specify what kind of computational problem concept learning is and to find appropriate constraints on the function with which it can be solved.

While our solution to the paradox has focused on the computational problem of concept learning, it is worth noting that this cannot be sufficient for a full theory of concept learning. Thus, the next step for a theory of concept learning, aside from the specification of the computational processes in the learning mechanism, is to investigate the representational problem of concept learning and to specify the properties of the mental representations that are used to compute the size of a concept. In addition, one must investigate how specifically to combine such a theory of concept learning with the various acquired biases of concept learning, specifically the shape and the basic-level biases. Future work in this area needs to build on this new foundation of mental representations for concept learning. This new foundation should allow not only for explaining concept learning but also for explaining learning to learn concepts.

REFERENCES

Brössel, Peter. 2017. "Rational Relations between Perception and Belief: The Case of Color." *Review of Philosophy and Psychology* 8 (4): 721–44.

Brössel, Peter, and Nina Poth. 2019. "Learning Words and Acquiring Concepts." Unpublished manuscript, University of Edinburgh.

Colombo, Matteo. 2018. "Bayesian Cognitive Science, Predictive Brains, and the Nativism Debate." *Synthese* 195 (11): 4817–38.

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought*. 1st ed. Cambridge, MA: MIT Press.

———. 2019. "From Sensations to Concepts: A Proposal for Two Learning Processes." *Review of Philosophy and Psychology* 10 (3): 441–64.

Gasser, Michael, and Linda B. Smith. 1998. "Learning Nouns and Adjectives: A Connectionist Account." *Language and Cognitive Processes* 13 (2–3): 269–306.

Horton, Marjorie S., and Ellen M. Markman. 1980. "Developmental Differences in the Acquisition of Basic and Superordinate Categories." *Child Development* 51 (3): 708–19.

Jackson-Maldonado, Donna, Donna Thal, Virginia Marchman, Elizabeth Bates, and Vera Gutirrez-Clellen. 1993. "Early Lexical Development in Spanish-Speaking Infants and Toddlers." *Journal of Child Language* 20 (3): 523–49.

https://doi.org/10.1086/706077 Published online by Cambridge University Press

Jones, Susan S., Linda B. Smith, and Barbara Landau. 1991. "Object Properties and Knowledge in Early Lexical Learning." *Child Development* 62 (3): 499–516.
Kemp, Charles, Amy Perfors, and Joshua B. Tenenbaum. 2007. "Learning Overhypotheses with Hierarchical Bayesian Models." *Developmental Science* 10 (3): 307–21.
Landau, Barbara, Linda B. Smith, and Susan S. Jones. 1988. "The Importance of Shape in Early Lexical Learning." *Cognitive Development* 3 (3): 299–321.
Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.
Mervis, Carolyn B., and Maria A. Crisafi. 1982. "Order of Acquisition of Subordinate-, Basic-, and Superordinate-Level Categories." *Child Development* 53 (1): 258–66.
Millikan, Ruth G. 1998. "A Common Structure for Concepts of Individuals, Stuffs, and Real Kinds: More Mama, More Milk, and More Mouse." *Behavioral and Brain Sciences* 21 (1): 55–65.
Nelson, Katherine. 1973. "Structure and Strategy in Learning to Talk." *Monographs of the Society for Research in Child Development* 1–2 (149): 1–135.
O'Callaghan, Casey. 2016. "Auditory Perception." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University.
Poth, Nina. 2016. "A Bayesian Approach towards Concept Learning and an Answer to the Complex First Paradox." MA thesis, Ruhr-University Bochum.
Rosch, Eleanor. 1975. "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology: General* 104 (3): 192–233.
Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7 (4): 573–605.
Sandhofer, Catherine, and Linda B. Smith. 2007. "Learning Adjectives in the Real World: How Learning Nouns Impedes Learning Adjectives." *Language Learning and Development* 3 (3): 233–67.
Shepard, Roger N. 1987. "Toward a Universal Law of Generalization for Psychological Science." *Science* 237 (4820): 1317–23.
Smith, Linda B., Susan S. Jones, Barbara Landau, Lisa Gershkoff-Stowe, and Larissa Samuelson. 2002. "Object Name Learning Provides On-the-Job Training for Attention." *Psychological Science* 13 (1): 13–19.
Smith, Linda B., and Larissa Samuelson. 2006. "An Attentional Learning Account of the Shape Bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005)." *Developmental Psychology* 42 (6): 1339–43.
Werning, Markus. 2010. "Complex First? On the Evolutionary and Developmental Priority of Semantically Thick Words." *Philosophy of Science* 77 (5): 1096–108.
Xu, Fei, and Joshua B. Tenenbaum. 2005. "Word Learning as Bayesian Inference: Evidence from Preschoolers." *Proceedings of the Annual Meeting of the Cognitive Science Society* 23:2381–86.
———. 2007. "Word Learning as Bayesian Inference." *Psychological Review* 114 (2): 245–72.

https://doi.org/10.1086/706077 Published online by Cambridge University Press