Cooperative Inference: Features, objects, and collections

Sophia Ray Searcy, Patrick Shafto
Rutgers University–Newark

Abstract

Cooperation plays a central role in theories of development, learning, cultural evolution, and education. We argue that existing models of learning from cooperative informants have fundamental limitations that prevent them from explaining how cooperation benefits learning. First, existing models are shown to be computationally intractable, suggesting that they cannot apply to realistic learning problems. Second, existing models assume *a priori* agreement about which concepts are favored in learning, which leads to a conundrum: learning fails without precise agreement on bias yet there is no single rational choice. We introduce Cooperative Inference, a novel framework for cooperation in concept learning, which resolves these limitations. Cooperative Inference generalizes the notion of cooperation used in previous models from omission of labeled objects to the omission values of features, labels for objects, and labels for collections of objects. The result is an approach that is computationally tractable, does not require *a priori* agreement about biases, applies to both Boolean and first-order concepts, and begins to approximate the richness of real-world concept learning problems. We conclude by discussing relations to and implications for existing theories of cognition, cognitive development, and cultural evolution.

Keywords: Concept learning; Pedagogy; Cultural Accumulation; First-order Logic; Algorithmic Learning Theory; Computational Complexity

Introduction

Until recently, researchers investigating learning have been primarily concerned with the inferences learners make when the source of information is disinterested in the outcome (e.g. corresponding to the natural world or a random process). However, people routinely learn valuable information from cooperative others (e.g. family members, teachers, mentors, friends). In these situations, both the informant and the learner have a shared interest in the outcome and this shared interest can have a dramatic effect on the learning process.

The authors would like to thank Liz Bonawitz, Lee Mosher, Asheley Landrum, Kelley Durkin, and Baxter Eaves as well as the CoDaS Lab for their helpful feedback. This work was supported in part by a grant from the National Science Foundation (DRL 1149116) to P.S.

Recent empirical research has shown that learning from a cooperative, knowledgeable informant leads to qualitative changes in learning. In this literature, cooperation is construed to be any situation in which a knowledgeable and helpful informant intentionally engages a learner to provide examples of a concept and conveys semantically generalizable information (Csibra & Gergely, 2006; Shafto & Goodman, 2008; Shafto, Goodman, & Frank, 2012; Shafto, Goodman, & Griffiths, 2014). For example, consider an experiment in which Bonawitz, Shafto, et al. (2011) found that when an informant is helpful, the absence of information can itself be informative. The researchers presented children with a novel toy. In one condition, the informant, known to be both knowledgeable and helpful, *intentionally* demonstrated a single function of the toy for the children. In a second condition, the informant *accidentally* triggered the same function in the toy. The researchers then observed the children playing with the toy. Despite having seen the same evidence, children in the intentional condition explored the toy less than children in the accidental condition. This indicates that when a cooperative informant intentionally omits information, that information is interpreted to be unnecessary, allowing the learner to draw stronger inferences about the absence of other functions. This is one instance of a growing literature showing that cooperation affords stronger, and often qualitatively different inferences than learning from the very same data in non-cooperative situations (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2012; Gergely, Bekkering, & Király, 2002; Gergely & Csibra, 2003; Gweon, Tenenbaum, & Schulz, 2010; Topál, Gergely, Miklósi, ErdH ohegyi, & Csibra, 2008; Xu & Tenenbaum, 2007a).

Indeed, cooperation is central to a number of recent theories of cognitive development and human culture. Csibra and Gergely (2009) argue that at a very early age, humans are prepared to be learners in a "natural pedagogy." Learners engage a specialized learning mechanism in response to ostensive cues, which interprets information received in such situations as more generalizable than the same information in other contexts. Similarly, researchers have proposed that humans have evolved a cultural niche for social learning (Boyd, Richerson, & Henrich, 2011), that the tendency to cooperate explains our ability to accumulate information over generations in ways that other animals do not (Tomasello, 1999), and that this kind of cooperation is a uniquely human phenomenon (Csibra, 2007). Though these proposals hinge on cooperation as a mechanism of driving learning, development, and human culture, they stop short of a precise explanation of why or how cooperation may help.

Computational accounts of learning from knowledgeable, cooperative informants proposed in the cognitive science (Shafto & Goodman, 2008; Shafto et al., 2014; Xu & Tenenbaum, 2007b) and machine learning literatures (Balbach, 2008; Zilles, Lange, Holte, & Zinkevich, 2008) could provide the link between cooperative informants and cultural accumulation of knowledge. The models in cognitive science differ in the details, but in general, they all explain cooperative informants as probabilistically choosing examples based on the true concept, as opposed to sampled at random and then correctly labeled. For example, Shafto and Goodman (2008) propose a model whereby the informant chooses examples that tend to maximize the learner's probability of inferring the correct hypothesis. This sort of model explains how learning from data chosen by a helpful informant can lead to stronger inferences than the same data in non-cooperative context—the data chosen by the informant has been chosen purposefully to disambiguate the correct hypothesis from nearby,

incorrect hypotheses (cf. Tenenbaum & Griffiths, 2001b).

Research on cooperation in the machine learning literature mainly falls under the umbrella of Algorithmic Teaching (e.g. Balbach, 2008; Zilles et al., 2008), an offshoot of the Algorithmic Learning literature (Gold, 1967; Valiant, 1984). In contrast with the psychological literature, Algorithmic Teaching has focused on deterministic models, which more easily lend themselves to proofs regarding learnability. Nevertheless, the machine learning and cognitive modeling literatures have converged to similar proposals (cf. Zilles et al., 2008; Shafto & Goodman, 2008). Indeed, all of the proposals in the cognitive modeling and the machine learning literature share two key assumptions: 1) informants know the learner's bias, and 2) cooperation works by omission of selected examples.

By bias, we are drawing on the notion of an *inductive bias* in the machine learning literature (Mitchell, 1980). It refers to an *a priori* assumption that some concepts are favored, all else being equal. In Bayesian probabilistic models, this most often takes the form of a prior probability distribution over the hypothesis space. A bias can take other forms, as well. A bias can simply be an ordering on the concepts, where lower ordered concepts are learned before higher ordered concepts, which is a slightly more general version of the standard Bayesian prior. The choice of which concepts to admit into the hypothesis space in the first place is another example of bias; excluding a concept is equivalent to assuming they are not learnable under any possible data (e.g. have a prior probability of zero). Because bias is a necessary precondition for demonstrating meaningful learning and generalization (Mitchell, 1980; Watanabe, 1969), biases are a universal feature of computational models of learning, as well as all models of cooperative learning that build on such models.

In this paper, we argue that existing computational accounts fail to link cooperation in learning to cultural accumulation of knowledge, and we present an alternative framework called Cooperative Inference that does so. We begin by introducing the notion of bias-first models of cooperative learning—models that possess an *a priori* bias that prioritizes learning some concepts over others—and argue that this class includes all previous computational accounts of cooperative learning (and quite possibly every model of learning in the cognitive science literature). We show that these bias-first-cooperation models are untenable models for learning from cooperative teachers for two reasons. First, we show that existing models are computationally intractable for all but trivially simple domains. Second, through a series of results, we show that the assumption that informants know the learner's precise bias *a priori* is both necessary for these models and implies unreasonable assumptions about computation. Just as learning from randomly sampled data requires the use of a bias, so does learning from a cooperatively selected data. However, there is no single optimal bias but rather a class of optimal biases, rendering the choice of bias underspecified. Finally, if the informant and learner do not have the same or extremely similar biases—highly implausible in any realistically large space—then cooperation does not help a learner successfully acquire a concept. Together, these results cast significant doubt on the current explanations for the benefits of cooperation for learning, including explicit models in cognitive science and computer science as well as less explicit theoretical accounts: there is no explanation that shows how cooperation can lead to faster learning or cultural accumulation of knowledge while avoiding complexity problems and without assuming a shared prior.

We then offer a novel framework for explaining the effects of cooperation for learning called Cooperative Inference. Cooperative Inference consists of two major proposals

absent from existing models of cooperation. First, it proposes that cooperation *precedes* assumptions about biases. Second, it generalizes the notion of cooperation previously applied to inferences about omitted examples to omitting novel aspects such as features and other aspects of concepts. The result is a computationally efficient approach to cooperation that allows *a priori* derivation of default biases for learning in novel domains. The approach does not require unrealistic computational resources and works without the need for *a priori* agreement between the informant and learner. Moreover, because computational limitations are relaxed and the informant and learner need not agree on a bias beforehand, learning can proceed in domains with a vast—even infinite—array of candidate concepts.

The omission of features and other aspects of concepts plays an important role in Cooperative Inference, allowing an informant to communicate examples as well as information about which parts of examples are important and which are not. Additionally, the relaxation of the assumption that all features are specified in each example allows for *new* features to be added to later examples, a process called composition. Composition allows informants to teach complex concepts by starting with simple concepts and, once learned, using those concepts as building blocks for later ones. We expound in two domains of successive complexity: Boolean concepts, and first-order logic concepts. This demonstrates the framework on concept learning problems accessible to most models of cooperation and learning, and a richer, more realistic domain. We provide examples of Cooperative Inference for Peano arithmetic and natural number that illustrate the potential of our framework to offer predictions in realistically complex domains where none were previously possible.

The paper follows in 5 sections: we review models of learning from cooperative informants; we prove limitations of these previous approaches; and we introduce our Cooperative Inference framework. We conclude by grounding Cooperative Inference in the psychological literature on learning and directly linking Cooperative Inference to the cultural accumulation of knowledge and discuss the implications for cognition, cognitive development, cumulative culture, and education.

Cooperation and learning: Bias-first models

Two main literatures have attempted to explain how cooperation might enable the type of efficient learning that could lead to cultural accumulation of knowledge: cognitive science and algorithmic learning. Each intends to capture a similar notion—learning should be more efficient when evidence comes from a cooperative source. The models that appear in each literature are also similar in that they assume a learner with a bias and propose that a cooperative informant selects evidence such that the combination of evidence and bias will lead the learner to acquire the target concept. In the following, we first look at the individual literatures separately before providing a general analysis of the constraints that apply to both groups of models.

Models in cognitive science

Tenenbaum (1999b) proposed a distinction between *weak sampling*, where observed examples are generated independently of the concept, and *strong sampling*, where observed examples are positive examples drawn from the target concept. Under weak sampling the distribution of observed examples provides no information about the target concept; after

the learner rules out concepts that are inconsistent with the examples, the inference is based only on the *a priori* probability of candidate concepts. Under strong sampling, however, examples are randomly drawn from the concept's positive extension (i.e. the set of true examples), so the observed examples provide the learner with information about the true concept beyond merely ruling out inconsistent concepts. Primarily, this leads learners to assign higher probability to concepts that generalize more conservatively, and the strength of this inference increases with the number of observed examples.

More formally, in the Bayesian framework the learner's inference corresponds to the posterior probability of a concept c given a sample of examples $s = \{x_0, x_1, \dots\}$, $P(c|s)$. The posterior probability is proportional to the product of the prior $P(c)$, representing the learner's expectations independent of any examples, and the likelihood $P(s|c)$, representing the learner's expectations about the generation of examples

$$P(c|s) \propto P(s|c)P(c) \ . \tag{1}$$

Under weak sampling, the evidence s rules out a concept c if it includes an example that is inconsistent with c, otherwise all remaining concepts are equally probable,

$$P_{weak}(s|c) = \begin{cases} 1 & \forall(x \in s) : x \in c \\ 0 & \exists(x \in s) : x \notin c \ . \end{cases} \tag{2}$$

For concepts where membership is not graded, e.g. rule-based concepts, strong sampling suggests that examples are generated uniformly from the concept. Thus, the likelihood depends both on consistency, as in weak sampling, *and* the size of the concept,

$$P_{strong}(s|c) = \begin{cases} \left(\frac{1}{|c|}\right)^{|s|} & \forall(x \in s) : x \in c \\ 0 & \exists(x \in s) : x \notin c \ . \end{cases} \tag{3}$$

Given the same data, a learner will make stronger inferences with strong sampling than weak sampling. Tenenbaum and Griffiths (2001a) showed that this model is a consistent extension of earlier models of generalization from a single example (Shepard, 1987) and set theoretic similarity (Tversky, 1977). Researchers have found that human learning fits with strong sampling in a variety of situations where one can expect only positive examples, including rule-based concepts (Tenenbaum, 1999b), word learning (Xu & Tenenbaum, 2007a), and sets of numbers (Tenenbaum, 1999a). Strong sampling even explains how information from a helpful source might lead to stronger inferences (Xu & Tenenbaum, 2007b). Strong sampling has two limitations: first, it does not allow the possibility of selecting negative examples. Second, and more importantly, it does not capture the *purposeful* selection of examples; intuitively, a cooperative informant does not seem well-explained by random selection of examples.

Shafto and Goodman (2008) proposed a model of pedagogical sampling, which captures the idea that cooperative informants choose data purposefully with the goal of increasing the learner's probability of inferring the correct hypothesis (Shafto & Goodman, 2008; Shafto et al., 2012; Shafto et al., 2014; cf. Tenenbaum & Griffiths, 2001b). This model is based on cooperation between the informant and learner: a rational informant selects a sample that tends to maximize the probability of the learner inferring the target concept,

and a rational learner assumes that the informant has chosen examples in this way. Unlike strong sampling, pedagogical sampling explicitly considers what it means to cooperate and formalizes this notion for both rational informants and learners. The model defines a system of two equations that capture the recursive relationship between the informant's selection of a sample and the learner's inference of a concept:

$$P_{learner}(c|s) = \frac{P_{informant}(s|c)P(c)}{\sum_{c'} P_{informant}(s|c')P(c')} \tag{4}$$

$$P_{informant}(s|c) = \frac{P_{learner}(s|c)P(c)}{\sum_{c'} P_{learner}(s|c')P(c')} \ . \tag{5}$$

Shafto and Goodman (2008) found that in simple rule-based domains, learners make stronger inferences only when examples come from a helpful informant, as predicted by pedagogical sampling. Shafto et al. (2014) found that pedagogical sampling also predicts the inferences of learners for prototype and causally-structured concepts. Bonawitz et al. (2011) showed that even preschool-aged children are sensitive to pedagogical sampling of evidence in their exploratory play. Buchsbaum et al. (2011) showed that preschoolers used pedagogical cues to infer the necessary sequence of actions to elicit effects. Critically, however, the model assumes that the informant knows the learner's precise prior, $P(c)$ (see Equation 5). By formalizing rational cooperation, and assuming the informant knows the learner's prior bias, pedagogical sampling is able to account for the implications of purposeful selection of evidence for learning.

Models in algorithmic learning

Algorithmic learning traditionally has focused on proving statements regarding learnability, where the source of information assumed to be adversarial or indifferent (e.g. inductive inference, Gold, 1967; PAC learning, Valiant, 1984; and query learning, Angluin, 1988). More recently, researchers have turned their attention to cases where the informant chooses examples to facilitate learning by a learner with a known bias. Angluin and Kric kis (1997) defined a algorithmic learning model that included an explicitly helpful informant for partial recursive function concepts. Shinohara and Miyano (1991), Anthony, Brightwell, Cohen, and Shawe-Taylor (1992), and Goldman and Mathias (1993) independently developed what is now called the *teaching set* model for Boolean concepts. In the teaching set model, a helpful informant teaches the target concept to all valid learners by selecting the smallest set of examples that will rule out all but the target concept. In this form, the only method for ruling out other concepts is by including examples that are inconsistent with those concepts (but still consistent with the target concept).

More recently in Algorithmic Teaching, researchers have converged on notion of cooperation similar to that in Cognitive Science. These models use cooperation to allow the informant to teach certain concepts using fewer examples. Balbach (2008) extended the teaching set to assume that a rational informant chooses a minimal, sufficient teaching sample; intuitively, if a student is given a sample of 3 examples, that student should be able to rule out any concept that can be taught with fewer examples. In the Optimal Teacher Teaching Dimension (OTTD) the student rules out logically inconsistent concepts (as in the teaching set) and additionally rules out all consistent concepts that could otherwise be

taught with a smaller teaching sample (Balbach, 2008). The OTTD begins by using the teaching set to define the minimal number of examples for each concept but it results in concepts being taught with even smaller samples. An iterated version (IOTTD) then repeats this process until convergence.

While IOTTD characterizes cooperation as elimination of concepts based on the size of the teaching sample, Zilles et al. (2008) defined an alternative cooperative model, Subset Teaching Set (STS), that allows learners to eliminate concepts based on the content of the sample. When a learner receives a sample, the learner considers the sample an optimal informant would choose for all concepts and eliminates any concept whose optimal sample does not contain the received sample. This sets up a recursive reasoning process, much like that independently proposed by Shafto and Goodman (2008).

Bias-first cooperation

The two classes of models share an important insight: the power of cooperation derives from reciprocal reasoning by the informant and learner about what information is (and is not) necessary to achieve learning. Moreover, in each of these models, the only information an informant may provide is a subset of the examples in the target concept (the informant may not, for instance, omit features from these examples). Each model also requires that the learner has a bias over the space of possible concepts, and that cooperation between the informant and learner is applied to that bias.

We introduce a framework that encompasses these previous models, *bias-first models of cooperation*, in order to analyze their common assumptions. As the name implies, the most important part of the framework is the use of an *a priori* learning bias. While the nature of the bias differs from model to model (e.g. in probabilistic models a bias is a probability distribution over the possible concepts and in deterministic models a bias often simply excludes all but a certain class of concepts) each model makes use of some kind of bias.[1] Our analysis will use the deterministic models from Algorithmic Learning Theory because it allows strong statements about learnability while the similarities between *all* bias-first models ensure that our points apply broadly.

In the next section, we investigate the necessary conditions for bias-first models to show benefits to learning from examples. To presage the results, we find that these models require assumptions that are both unreasonable and necessary for success, and we therefore conclude that they are inadequate explanations of implications of cooperation for learning, development, and culture.

The limitations of bias-first models of learning

Previous models of cooperative learning share a set of basic assumptions: that the learner has some bias and that the informant selects evidence such that the combination of evidence and the learner's bias leads the learner to acquire the target concept. However, the reciprocal reasoning between informants and learners intertwines possible concepts and possible examples, leading to a computationally intractable reasoning problem. Moreover,

[1] Indeed, because all models that demonstrate meaningful learning and generalization assume some bias (Mitchell, 1980; Watanabe, 1969), these analyses apply to any model of learning that is augmented with reasoning about cooperative informants.

just like learning is not possible without an *a priori* learning bias (Watanabe, 1969), learning from a cooperative informant is not possible without an *a priori* bias. Informants, then, must know what biases learners have in order to choose the examples that will lead the learner to infer the target concept. However, it is not clear how informants could know the learner's biases. In what follows we formalize this quandary, demonstrating that the reliance on biases is a fatal limitation for previous models.

To do so, we adopt the *teaching set* framework. This is a deterministic approach that is well-suited to proofs regarding the conditions under which concepts are teachable. Although the framework assumes deterministic inferences, our key results merely depend on the existence of a bias which is a key assumption of probabilistic models used in cognitive science, and the arguments hold for these as well as many other possible models of learning that apply cooperation over an *a priori* learning bias. Specifically, the first proof, regarding computational complexity[2], applies to all biases; the second proof demonstrating the necessity of a bias trivially applies to all biases; the third proof, regarding the rational choice of bias, focuses on a specific kind of deterministic bias and we discuss what this implies for a more broad class that would include all deterministic and probabilistic biases; finally, the analysis regarding bias mismatch necessarily applies to all biases.

Building on previous research in cognitive science and machine learning, we focus on the domain of Boolean concepts (Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961; Feldman, 2000; Valiant, 1984; Goodman, 1955). Boolean concept learning typically focuses on examples with binary features, and allows concepts to be formed of any subset of the possible examples. The set of Boolean concepts is the powerset of possible examples in a given domain. Many of the results we show depend on the size and complexity of the concept space, and the Boolean domain represents a useful lower bound on size and complexity.

The technical results are presented in the following subsections of this section. Before beginning, we present an abbreviated and less technical version of each result.

One point, common to many of our technical results, is that the structure of concepts leads the concept space—the set of possible concepts a teacher and learner must consider—to grow extraordinarily quickly. Our notation begins with features at the very bottom and then builds the next level of notation, examples, from combining features with labels. Simple combinatorics governs the number of possible examples that results from using a certain set of features and labels—we begin by using boolean features (so 'red' which can be labeled true or false rather than 'color' which can be labeled 'green', 'red', etc.) whose two possible labels means that the number of examples will follow $n_e = 2^{n_f}$. The set of possible concepts is built on the set of examples much like the set of possible examples is built on the set of features. Because an example can only either be in a concept or not (equivalent to two labels) the number of possible concepts will follow $n_c = 2^{n_e} = 2^{2^{n_f}}$.

This means that the number of concepts that might be considered grows considerably fast. The bar for tractability in many complexity analyses is polynomial growth—where the time it takes to complete an operation grows according to some fixed exponent, x^2 or even x^{20}, in terms of the size of the input. For concepts built in the manner we have laid

[2]Computational complexity refers to the resources used by a particular algorithm or program as a function of the size of its input. It is common to consider how the amount of time until the program finishes changes as a function of the size of the input (time complexity).

out, the growth is not just exponential in terms of the number of features (which would itself be intractable) but double-exponential. This means that any theory or model whose usefulness depends on the size of the concept space growing slowly (or simply not growing extraordinarily quickly) will fail.

The growth of the concept space is only a problem, however, if we assume that the number of features becomes large (or very large in some cases) and a skeptical reader might question this assumption. After all, the majority of experiments we consider use only a few features (Feldman (2000) tested concepts with 3 and 4 features while Goodman, Tenenbaum, Feldman, and Griffiths (2008) tested concepts with 7 features). Suffice it to say that this is one of many ways in which these experiments differ from the real world; there are vastly more features with which we can make concepts and many of these features are much more complex than Boolean features. Concept learning experiments use a few simple features for at least two reasons: it is easier to conduct a well-designed experiment (e.g. one may ensure that differences in feature salience do not affect results if there are relatively few features and they are assigned randomly or counter-balanced, but this is more demanding as the number of features grows) and one can apply computationally demanding models to these experiments (more on this later).

So far, we have established that the size of the concept space (the concepts that need to be considered) grows very quickly in terms of the number of features and that in realistic cases, the number of possible features is quite large. We have also hinted that one method for dealing with this problem is to constrain the concept space by including only certain concepts. This is an instance of a *learning bias*, known as a concept class, which excludes certain logically possible concepts. One common concept class is that of the *monomials* (see e.g. Balbach, 2008) which includes concepts made of features connected by *and* (e.g. 'is_large and not is_red', 'is_square and is_spotted') but no other concepts. This concept class grows as 3^{n_f} which is a significant improvement over the overall concept space $2^{2^{n_f}}$. Other possibilities include using only concepts with up to k conjunctions (Valiant, 1984), only including certain features as input to concepts, and only allowing disjunctions (as opposed to conjunctions).

Our first result, presented in 'Computational complexity' concerns the time-complexity of models of cooperation—the amount of time it takes to complete an operation as a function of the size of the input—and builds directly from the size of the concept space. The objective is to determine if, despite being obviously tractable where computational simulations have been completed (e.g. Bonawitz et al., 2011; Shafto & Goodman, 2008), models of cooperation are tractable for more general settings, where the number of features may be considerably greater[3]. Here we use the the set of features as the 'input' so that n corresponds to the number of features. The critical question, then, is the relationship between the amount of time models of cooperation need as a function of the size of the input.

Informally, models of cooperation operate on a two dimensional space of concepts

[3] Here, we adopt the convention (Cobham, 1965) where a system must have polynomial time complexity ($\Theta(n^k)$ for some constant k) in order to be tractable. Roughly, this means that a model is tractable as long as the amount of time it takes to run on a certain input scales with the size of the input raised to some constant power. So, for example, a model that takes $\Theta(n^2)$ time to complete would be tractable but one that takes $\Theta(2^n)$ would not be. Our analysis remains consistent with more specific arguments for a tractability standard in cognitive models (e.g. Beal & Roberts, 2009; van Rooij et al., 2011).

and teaching samples. The informant intends to communicate a target concept and searches along the teaching sample dimension for a sample that is most likely to communicate the concept to the learner. In order to do this, though, the informant must calculate the probability of learning each concept for each possible teaching sample—the informant must maximize the probability that the learner will acquire the target concept relative to the remaining concepts. As we discussed earlier, the size of the concept space in terms of the number of features is a double exponential, so simply searching along this dimension alone would be sufficient to render models of cooperation intractable, but informants must search along both concept and sample dimensions. Additionally, many models of cooperation are reciprocal in nature (Balbach, 2008; Shafto & Goodman, 2008; Zilles et al., 2008) meaning that this intractable process would be repeated multiple times.

The second result, presented in 'Ineffective without a bias' demonstrates that, ignoring complexity concerns, cooperative models are ineffective when no bias is used. In our treatment, a lack of a bias means that all concepts in the concept space are simultaneously entertained without preference. The result is that, in order for a learner to acquire a concept under this condition, all possible examples must be observed. A simple demonstration shows that this is the case. Imagine that a learner receives all but a single example. There must then be two concepts in the concept space that are consistent with the received examples: the target concept and a second concept that is identical to the target concept along each received example but is labelled opposite for the example not received.

This result indicates that if we want cooperative models of learning to be more effective than simply enumerating labels for all possible examples, then a bias is necessary. In 'Many optimal biases' we consider how this bias might be chosen. We look specifically at how a rational model of cooperation might select a bias. According to the principle of rationality, the optimal bias—the one that results in the lowest average cost—would be selected. We establish a simple method for calculating cost: the number of examples it takes to communicate a concept averaged over the concept space.

Using this, we find the optimal bias for two different classes of bias. One class, the ordered bias, simply arranges concepts according to a priority and, when multiple concepts are consistent with a sample, leads the learner to acquire the concept with the highest priority. In this class of bias, there is no single optimal bias but rather a collection of biases, all of which are equally optimal. The story is similar for the more powerful functional bias which may redefine which examples are consistent with which concepts. We show that there is a collection of optimal functional biases rather than a single one.

Finally, in 'Impossible for unknown bias', we consider how models of cooperation work when the teacher does not know the bias of the learner. We assume a simple and conservative case where the learner uses no bias and the informant's bias considers only some subset of the concept space. In such a case, the success of models of cooperation depends on the concepts shared by the learner's and the informant's bias: obviously only concepts that are shared by both can be successfully communicated, but even this is not guaranteed. One possibility is that the informant will provide examples sufficient to rule out all concepts other than the target concept according to the informant's bias but not according to the learner's bias. To determine how probable this possibility is compared the possibility of successful acquisition of the concept, we assume that the informants bias is constructed by randomly sampling concepts to exclude from the informant's bias. (This is equivalent to assuming that

the informant has no information about the learner's bias). The probability that a random concept in the informant's bias can be successfully taught depends on how many concepts in the learner's bias that the informant's bias omits. However, we show that unless the size of the informant's bias grows at a similar rate to the size of the concept space—and thus implies intractability and ineffectiveness according to our first two results—then successful teaching becomes vanishingly unlikely as the size of the concept space increases.

Together, our 4 results show that, using existing models of cooperation, a bias is necessary in order for teaching to be tractable and effective. However, the principle of rationality does not offer a method for coordinating bias, and successful teaching rapidly becomes unlikely when the learner and informant do not know each others' biases. For the remainder of this section we present the technical results. Following that, we introduce a novel model of cooperation that alleviates these problems and provides new insights into the relationship between cooperation, learning, and representation.

Notation

We begin with the notation for Boolean concepts and formally define the teaching set model. Later, the Boolean notation will be extended to first-order logic concepts.

Let $\mathcal{F} = \{f_0, f_1, \dots\}$ be the *feature space*. Let the *instance space* be the function space from $\mathcal{F}$ to the Boolean labels $\mathcal{X} = \{\mathrm{F}, \mathrm{T}\}^{\mathcal{F}}$. And let the *concept space* be the function space from $\mathcal{X}$ to the Boolean labels $\mathcal{C} = \{\mathrm{F}, \mathrm{T}\}^{\mathcal{X}}$. A concept class is some subset of the concept space $C \subseteq \mathcal{C}$.

Ordered pairs formed of features and Boolean labels such as those found in an instance $(f, b) \in x$ are referred to as *specifications* and b takes on either value $\{\mathrm{F}, \mathrm{T}\}$. Ordered pairs between instances and Boolean labels such as those found in concepts $(x, b) \in c$ are referred to as *examples*. Finally, let the *sample space* be any set of examples that can be found in a concept, $\mathcal{S} = \{s \mid s \subseteq c \wedge c \in \mathcal{C}\}$. A concept and sample are *consistent* if each example in the sample is also in the concept,

$$\mathrm{Cons}(s, C) = \{c \mid s \subseteq c \wedge c \in C\} \ . \tag{6}$$

Within the teaching set framework, a learner is both *consistent* and *class-preserving*. Consistent means that learners will only learn concepts that are consistent with the teaching sample. Class-preserving means that learners will only learn a concept that is in the concept class. Equivalently, the learner can be thought of as beginning with all concepts in the concept class and ruling out concepts that are inconsistent with the sample.

Assuming that the informant knows the concept class of the learner, a teaching set is the sample with the fewest examples that will teach the target concept for any consistent, class-preserving learner—i.e. that will rule out all other concepts in the concept class.

Definition: teaching set. The teaching set for a target concept is the minimal sample that teaches the target concept to all consistent and class-preserving learners,

$$\mathrm{TS}(c, C) = \arg\min_{s \in \mathcal{S}} \{|s| \mid \mathrm{Cons}(C, s) = \{c\}\} \ . \tag{7}$$

The *teaching dimension* for a target concept is the size of the teaching set.

Computational complexity

We begin by considering the time complexity of bias-first models of learning. A major challenge for this endeavor is that these models represent a computational level (Marr, 1982) or rational analysis (Anderson, 1991; Chater & Oaksford, 1999) explanation and thus do not specify implementation details. With this in mind, the following analysis gives special attention to the bounds within which *all* sampling models must fall. The complexity of bias-first models depends on two basic components: the complexity of the concept space considered, and the cost of applying the algorithm over that space. The Boolean concept domain (e.g. Bruner et al., 1956; Feldman, 2000; Nosofsky, Palmeri, & McKinley, 1994; Shepard et al., 1961; Valiant, 1984) represents a reasonable minimal set of concepts, and thus lower bound on concept complexity, so that any finding of intractability should hold for more complex domains. We also show that while the implementation of many bias-first models have extreme complexity cost, bias-first models remain intractable even when the costs specific to a particular implementation are ignored.

First, consider the growth of the size of the sample space in terms of the number of features. Using $n = |\mathcal{F}|$, each Boolean concept contains 2^n examples. Given a target concept c, the set of candidate teaching samples is the set of all subsets of c and has a size,

$$|\mathcal{S}| = |\mathcal{P}(c)| = 2^{|c|} = 2^{2^n} \ . \tag{8}$$

This implies that a model with polynomial complexity $\Theta(n^k)$ in the number of features n must then have $\Theta(\log \log m)$ complexity in number of teaching samples m in order to be computationally tractable.

Most models define cooperation as the selection of a sample that maximizes the likelihood that the learner acquires the concept (Shafto & Goodman, 2008; Tenenbaum & Griffiths, 2001a) or the smallest teaching sample that guarantees the learner will acquire the concept (Balbach, 2008; Zilles et al., 2008). Thus, a general template for sampling models is an algorithm that computes the learner's inference for each sample and searches for the teaching sample that maximizes some measure, such as the likelihood that the learner acquires the concept,

$$\arg\max_s P_{learner}(c|s) \ . \tag{9}$$

The complexity of calculating of the probability $P_{learner}(c|s)$ can be expanded[4] as (e.g. Shafto & Goodman, 2008; Tenenbaum & Griffiths, 2001a),

$$P_{learner}(c|s) = \frac{P(s|c)P(c)}{\sum_{c' \in \mathcal{C}} P(s|c')P(c')} \ . \tag{10}$$

The probability $P_{learner}(c|s)$ can be found by setting up a $|S| \times |C|$ matrix and computing $P(s|c)P(c)$ for each unique pair of concepts and samples. For each sample, the sum can be computed from a row in the matrix and then each $P(s|c)P(c)$ is divided by its respective sum. If each $P(s|c)P(c)$ is calculated at constant cost, doing so for all $c \in \mathcal{C}$ has

[4] This assumption additionally allows the extension of this analysis to deterministic models (e.g. Balbach, 2008; Zilles et al., 2008) which require the computation of $\mathrm{TS}(c, \mathcal{C})$ for each $c \in \mathcal{C}$. Both the probabilistic calculation of $P(s|c)P(c)$ and the deterministic calculation of $\mathrm{TS}(c, \mathcal{C})$ require *at minimum* constant time.

$\Theta(|\mathcal{C}|)$ complexity, as do the sum and division operations. This must then be calculated for all $s \in \mathcal{S}$, resulting in

$$\Theta(|\mathcal{C}||\mathcal{S}|) = \Theta\left(m^2\right) = \Theta\left(2^{2^{n^2}}\right) . \tag{11}$$

This result applies to all models where an informant intentionally selects examples that are most helpful to the learner, such as non-iterative derivatives of the teaching set model (e.g. OTTD in Balbach, 2008). The analysis also applies to the first step of iterative versions of both probabilistic (Shafto & Goodman, 2008) and teaching set (STS in Zilles et al., 2008; IOTTD in Balbach, 2008) models, and is thus a lower bound on their complexity. Each of these models, then, has double-exponential time complexity and is far from tractable.

A potential criticism of this approach is that these models intend to describe the *outcome* of the process rather than the specific *implementation* (e.g. Shafto & Goodman, 2008; Shafto et al., 2014). It is worth considering, then, the general complexity requirement that any theoretical implementation must satisfy in order to remain tractable. The size of the Boolean sample space is $\Theta\left(2^{2^n}\right)$ in the number of features. In order to remain tractable in the number of features a cooperation model must have a complexity of $\Theta(\log\log m)$ in the number of samples. This means that bias-first learning in eq. (9) cannot be made tractable by even the most generous stand-in psychological process. Consider a model that possesses an *a priori* a list of $P_{learner}(c|s)$ for all $s \in \mathcal{S}$ and thus requires no explicit calculation. Selecting the index of the greatest entry in a list of $P_{learner}(c|s)$ has a complexity of $\Theta(m)$ (Cormen, Leiserson, Rivest, & Stein, 2009) in the number of concepts and thus $\Theta\left(2^{2^n}\right)$ in the number of features. Thus even the most generous implementation-level assumptions do not allow for such models to be tractable.

As specified, existing bias-first models are not tractable accounts of cooperation. Moreover, the bias-first approach disallows any tractable account of cooperation because cooperation is by definition applied over the set of possible concepts *and* samples. The space of concepts and the space of possible teaching samples grows so quickly as the number of features increases that even the most generous possible assumption—that the informant only has to select the best examples using already computed calculations—results in an intractable model. This is, of course, assuming that all candidate concepts are considered with non-zero weight—it is not a coincidence that very few models have utilized such a concept space without some ad hoc restriction. The plausibility of sampling models, then, depends entirely on how the concept space is restricted. We turn our analysis to this for the following three sections.

Ineffective without a bias

Models of cooperation have assumed various learning biases. But is it *necessary* to adopt a bias? Results for learning without cooperation have shown that, in the absence of a bias that prioritizes some concepts, even after observing some evidence, all consistent concepts are equally probable. Thus meaningful learning requires an *a priori* bias (in the sense that it would support generalization to not previously observed objects) (see Ugly Duckling Theorem in Watanabe, 1969, p. 376). A similar result holds for bias-first models when learning from a cooperative source; even with a cooperative informant there is no meaningful learning without a bias.

Theorem: Learning without bias Given a set of n features, $\mathcal{F}_n$, and a concept class, $\mathcal{X}_n = \{\mathrm{F}, \mathrm{T}\}^{\mathcal{F}_n}$, $C_n = \{\mathrm{F}, \mathrm{T}\}^{\mathcal{X}_n}$. The teaching set for any $c \in C_n$ must include every instance in $\mathcal{X}_n$ labeled according to the concept such that $\mathrm{TS}(c, C_n) = c$.

Proof Let s, the teaching sample for c, label all but m instances in c, such that $|c \setminus s| = m$. The set of unlabeled instances can be used to form a set of concepts $X^* = \{x \mid (x, b) \in c \setminus s\}$, $C^* = \{\mathrm{F}, \mathrm{T}\}^{X^*}$. These concepts may be used to construct $\mathrm{Cons}(s, C_n)$, the set of all concepts in C_n consistent with s,

$$\mathrm{Cons}(s, C_n) = \left\{ s \cup c' \mid c' \in \{0, 1\}^{X^*} \right\} . \tag{12}$$

It follows that $\mathrm{Cons}(s, C_n) = \{c\}$ iff $X^* = \emptyset$—i.e. $m = 0$ and no instances are left unlabeled—and thus $\mathrm{TS}(c, \mathcal{C}) = c$. ■

Without a bias, previous models of cooperative learning allow a learner no generalization beyond received examples. This means that an informant must label *every* instance in order to successfully teach the target concept—learning from a cooperative informant is no better than other sampling methods, all of which trivially allow a learner to learn a concept when allowed to sample the entire example space. This result, that a bias is necessary for cooperation, along with the previous result, that a restricted concept class is necessary for tractability, together point to the paramount importance of the choice of bias for these models.

Many optimal biases

If a learner needs a bias in order for cooperation to be effective, the most sensible choice for the bias is the one that would, on average, lead to the most efficient learning. To analyze optimal biases, we formalize two types of biases that previous researchers have used. The first, *ordered bias*, adopts a total pre-order on the concepts in $\mathcal{C}$ such that lower ordered concepts are given priority. One example of this is the Occam's Razor bias (Balbach, 2008)[5] where concepts are judged by the number of terms (separated by 'or's) in the description, e.g. 'feps are red and square' which has one term would be given priority over 'feps are red or feps are blue and square' which has two. The second, *functional bias*, permits a redefinition of the set of concepts that are consistent with any sample. This bias is more complex but also opens up many more possibilities. Examples of this include pedagogical sampling (Shafto & Goodman, 2008) and the Subset Teaching Set (Zilles et al., 2008) where a learner rules out concepts based on a prediction of the teaching set rather than based on whether or not an example is consistent. For both ordered and functional biases, we show that there are many optimal biases and thus no *a priori* rational choice. Moreover, agreement about a bias through prior communication would allow arbitrarily efficient learning—effectively equivalent to telepathy.

Formally a learner with an ordered bias learns the lowest-order concept that is consistent with the examples. Thus, an ordered bias amounts to a modification of the Cons() function. Given c and $\preceq$, we use $C_{\preceq c} = \{c' \mid c' \in \mathcal{C} \wedge c' \preceq c\}$ to refer to the set of concepts of the same or lower order as c and we modify the consistent and teaching set functions as

[5] For a probabilistic version, see Goodman et al. (2008).

follows,

$$\text{Cons}(s, \mathcal{C}, \preceq) = \{c' \mid s \subseteq c' \wedge c' \in C_{\preceq c}\} \text{ and} \tag{13}$$

$$\text{TS}(c, \mathcal{C}, \preceq) = \arg\min_{s \in \mathcal{S}} \{|s| \mid \text{Cons}(s, \mathcal{C}, \preceq) = \{c\}\} \ . \tag{14}$$

Next, we develop a novel ordered bias called the *Hamming distance bias*[6] and show that it minimizes the average teaching dimension and is thus optimal. Informally, the proof is as follows. Each ordered bias includes a least-element concept that is of equal or lower order to all other concepts in the concept class. Any ordered bias would, at minimum, require the teaching set of a target concept to include examples sufficient to rule out the least-element concept (because it is necessarily in the set of concepts of lower order than the target concept). For the Hamming distance bias, we show that the teaching set of each concept is the minimal set of examples sufficient to rule out the least-element concept and therefore it is a minimal ordered bias.

Theorem: Hamming distance bias Let $h(c_1, c_2)$ be the Hamming distance between c_1 and c_2 such that $h(c_1, c_2) = |c_1 \setminus c_2|$. Given an origin concept, c^*, the Hamming distance bias is $\preceq_{h(c^*)} = \{(c_1, c_2) \mid c_1, c_2 \in \mathcal{C} \wedge h(c^*, c_1) \leq h(c^*, c_2)\}$ and is an optimal ordered bias.

Proof Let $\preceq_{c^*}$ be an ordered bias with c^* as a least element, i.e. $c^* \preceq_{c^*} c$ for all $c \in \mathcal{C}$. Then c^* is in $C_{\preceq_{c^*} c}$ for all $c \in \mathcal{C}$ and in order to rule out c^*, the teaching set for any c must include $c \setminus c^*$,

$$\text{TS}(c, \mathcal{C}, \preceq_{c^*}) \supseteq c \setminus c^* \ . \tag{15}$$

For $\preceq_{h(c^*)}$ and any pair of concepts $c_1, c_2 \in \mathcal{C}$ such that $c_1 \neq c_2$, if c_1 is of lower order than c_2, $c_1 \preceq_{h(c^*)} c_2$ (i.e. $|c_1 \setminus c^*| \leq |c_2 \setminus c^*|$), then the teaching set formed from eq. (15) will also rule out c_1, $(c_2 \setminus c^*) \not\subseteq c_1$. So $c_2 \setminus c^*$ is sufficient to rule out any concept of a lesser order, i.e.

$$\text{Cons}(c_2 \setminus c^*, \{c_1, c_2\}, \preceq_{h(c^*)}) = \{c_2\} \tag{16}$$

$$\text{TS}(c, \mathcal{C}, \preceq_{h(c^*)}) = c \setminus c^* \ . \tag{17}$$

From eq. (15), the Hamming distance bias results in the minimal size teaching set for each concept and is thus optimal. ■

For example, if the origin concept is the concept with all negative examples[7], $c^* = \text{FFF}\ldots$, the teaching set for each target concept c will include only the examples that differ from c to c^* or, in other words, the set of positive examples in c. Thus the average teaching dimension for the Hamming distance bias is the average number of positive examples in each concept, or $\frac{2^n}{2} = 2^{n-1}$ for a concept space with n features. The Hamming distance bias is optimal without regard to the origin concept, so the average teaching dimension would not change if the origin concept were changed. What does change, however, is the number of examples required to learn particular concepts—especially the origin concept which can be learned from an empty teaching sample.

[6] Named for Hamming (1950), the Hamming distance between two binary vectors is equal to the number of bit positions where the two vectors differ, so that $000, 000$ are separated by a Hamming distance of 0 while $000, 101$ are separated by a Hamming distance of 2.

[7] For abbreviation, we use strings to stand in for samples and concepts, e.g. $c = \text{TFT}*$ would be a sample that labels the the first and third examples as true, the second as false, and does not label the fourth. These strings are over $\{\text{F}, \text{T}, *\}$ such that $(x_i, A[i]) \in s$ for all $A[i] \neq *$.

Rational selection of a functional bias is similar to that of an ordered bias, with the exception that, because a functional bias may modify which concepts are consistent with which samples in any way, there is no analogous constraint to eq. (15). Thus, a functional bias may use any set of examples to teach a target concept so long as each concept is taught with a different set of examples.

The minimal functional bias would assign a concept to each unique $s \in \mathcal{S}$, beginning with the smallest. The first concept would be taught with $s = \emptyset$, the following concepts would be taught with samples where $|s| = 1$, and so on. For $|s| = i$, and the number of features, n, the number of unique samples in $\mathcal{S}$ is $\binom{2^n}{i}2^i$. It is outside of the scope of this paper to determine the average teaching dimension for the optimal functional bias, though it is clearly smaller than that of the optimal ordered bias.

Both the optimal ordered and functional biases include free parameters that allow any concept to be taught without any examples at all. Thus, for each concept that can be taught, there is at least one unique optimal choice of bias. Indeed, if *a priori* selection of the optimal bias is allowed, learning does not require any examples. However, there is no single bias that is optimal for all concepts.

Thus far, we have shown that a bias is necessary—if the informant and learner do not use a bias, then there is no way to efficiently provide examples. We have also shown that rational selection does not help with this problem—without knowledge of the target concept, there are many optimal biases.

Impossible for unknown bias

Our final look at bias-first models concerns the situation where the informant does not precisely know the bias of the learner—i.e. where there is some difference between what the informant would guess for the bias of the learner and the actual bias. For the following analysis, we will use classes $C_i, C_l \subseteq \mathcal{C}$ to stand in for more complex biases without a loss of generality (see eq. (13)). The informant's class C_i is used to determine a teaching set that is *consistent* only with the target concept, $s_i = \text{TS}(c, C_i)$, while the learner's class C_l is the class that is *preserved* when the learner uses the sample to rule out other concepts, $C' = \text{Cons}(\text{TS}(c, C_i), C_l)$. Given C_i and C_l, we say that a concept c is *teachable* iff $\{c\} = \text{Cons}(\text{TS}(c, C_i), C_l)$ and we use the following indicator function such that $\text{Teach}(c) = \text{True}$ if the concept c is teachable and False otherwise.

To begin with, concepts not in either C_i or C_l are trivially unteachable. Of the concepts in the intersection of the informant and learner's classes $c \in C_i \cap C_l$, a concept is teachable iff each example in the teaching set from the learner's perspective: $\text{TS}(c, C_l)$ is included in the teaching set from the informant's perspective $\text{TS}(c, C_i)$. Each example in $\text{TS}(c, C_l)$ represents a necessary condition for the teachability of c.

The teaching set from the informant's perspective includes an example only when it is necessary to rule out a concept that has not already been ruled out by other examples in the teaching set. We refer to such concepts as *adjacent* concepts. Given a sample, we say that a concept is adjacent along an example when the concept would be ruled out if the example is included in the teaching set but not otherwise. To illustrate, imagine that the teaching set for a target concept from the learner's perspective is $\text{TS}(c, C_l) = \text{FFT}*$. The informant's class must have included at least one concept adjacent to the third example (either FFFF or FFFT) otherwise $\text{TS}(c, C_i)$ will not include that example (e.g. $\text{TS}(c, C_i) = \text{FF}**$).

To determine the probability that a concept is teachable, we model a process where the informant's class is determined by randomly drawing without replacement from the set of concepts. Then, for each example, the hypergeometric distribution gives the probability that all adjacent concepts are removed. Let $U = |\mathcal{C}|$ be the size of the universe of concepts and R be the number of concepts removed from $\mathcal{C}$ to get the informant's class of size $T = |C_i|$ such that $U = R + T$. Given an example in the teaching set from the learner's perspective, let A_e be the number of adjacent concepts,

$$P(\text{Teach}(c) = \text{False} \mid R, C_l, e) = \frac{\binom{A_e}{A_e}\binom{U-A_e}{R-A_e}}{\binom{U}{R}} \ . \tag{18}$$

As the number of features becomes very large, $n \to \infty$, the number of concepts, U, does as well. If R remains constant, then $\lim_{n\to\infty} P(\text{Teach}(c) = \text{F}) = 0$ meaning that a target concept will be teachable in the limit. But, because $U = R + T$, a constant R would mean that the informant's concept class has a double exponential increase, i.e. $O(T) = 2^{2^n}$, and this implies an implausible lack of constraints on the size of a concept space. Consider the result from the previous complexity analysis: at best, the complexity of a bias-first model is the square of the size of the concept space. If T increases less than a double exponential such that R approaches U in the limit $R = U - T \to U$, the target concept will not be teachable.

Stirling's approximation of the factorial—appropriate because both U and R are very large—provides a simplification of the hypergeometric distribution in for $n \to \infty$,

$$\frac{\binom{A_e}{A_e}\binom{U-A_e}{R-A_e}}{\binom{U}{R}} \sim \frac{(U - A_e)^U R^U}{U^U (R - A_e)^U} \ . \tag{19}$$

Then, because $R \to U$, $\lim_{n\to\infty} P(\text{Teach}(c) = \text{F}) = 1$; concepts are, in general, unlikely to be teachable.

This result depends on a set of reasonable assumptions: that the set of possible features is large, that the set of concepts considered by the informant and learner is much smaller than the set of all concepts, and that the informant cannot predict which concepts are in the learner's concept class. Given these assumptions, the probability that a concept could successfully be taught via sampling approaches zero. The fact that more complex biases, such as ordered, effectively use concept class bias for a given target concept (see eq. (13)) means that this result applies to all such biases.

To summarize, bias-first approaches suffer from the dependence on the assumption that the informant precisely knows the bias of the learner. Because bias-first learning works through the iterative application of a bias over examples, it is computationally intractable. The learner also must use a bias because efficient learning is impossible without a bias. However, there is no single rational choice for bias. If the informant and learner choose a bias without coordination, successful teaching becomes impossible as the number of features grows. Individually, any of these results would be a significant but manageable problem for bias-first models of cooperation. Together, they suggest that bias-first models do not explain how cooperation could lead to either efficient learning or cultural accumulation of knowledge for even reasonably-sized learning problems.

Cooperative Inference

In what follows, we propose a framework that can explain how cooperation leads to efficient learning and cultural accumulation of knowledge. Previous approaches use the omission of labeled objects (or, by symmetry, the selection of labeled objects) to facilitate learning. Building on these approaches, we consider the cooperative omission (or indications of relevance) of other aspects of a concept: specified features, labeled objects, and labeled collections of objects. The result is a generalization of the notion of cooperation. This change leads to a number of important consequences: informants may successfully teach in unbounded concept spaces, the complexity of learning from a cooperative informant depends on the complexity of the target concept irrespective of the concept space, and there is a natural representation for concepts.

Moreover, these advantages enable learning in domains that go well beyond those considered previously in the literature. To illustrate, we extend Cooperative Inference to learning concepts in first-order logic. We show how the basic operations of first-order logic—quantification, composition, and predication—can be learned from a cooperative informant, and illustrate this with demonstrations of learning arithmetic and natural number concepts via examples.

One way to understand Cooperative Inference is to think of it as a form of learning bias. Whereas bias-first models critically hinge on restricting the concepts space through an a bias unrelated to cooperation and then applying a version of cooperation, Cooperative Inference proposes that cooperation *is the bias*. Thus, in contrast with previous approaches that assume learning biases are optimized to learn from the world, *Cooperative Inference proposes that biases are optimized to learn from cooperative informants.* Thus, while it should be possible to combine the biases of Cooperative Inference with other situation-specific biases, Cooperative Inference itself should remain consistent wherever the situation applies.

Cooperative inference applies when some part of a concept is omitted by an informant in order to communicate that concept helpfully to a learner. Traditionally, this only occurs when an example is omitted from the concept, but there is no reason to suppose that informants are limited to this. We consider, in turn, what cooperative inferences are made when an informant omits features, labels, and object specifiers. While the result is different at each step, the underlying principle is the same: the learner infers that the informant is omitting information in order to helpfully communicate the concept, and the informant, aware of this, does just that.

Cooperative Inference for Boolean concepts

First, we consider the domain of Boolean concepts, the most common domain for previous work in cooperation. As stated earlier, Cooperative Inference applies a notion of cooperation to the intentional omission of aspects of concepts other than examples. For Boolean concepts, we introduce two such inferences: the omission of objects according to concept label (e.g. all objects labeled ‘false’) and the omission of features within objects.

The first Cooperative Inference we discuss for Boolean concepts regards omitting label objects. Rather than omitting objects individually, an informant may omit objects according to the concept label (i.e. omit all examples labeled not in the concept or all

Table 1
Cooperative Inference for two units of analysis

Object-Level	Specification (f, v)	**Object** $\left\{(f, v), \atop \vdots\right\}$	Concept $\left\{\left\{(f, v), \atop \vdots\right\}, \atop \vdots\right\}$
Collection-Level	Specification (f, v)	Object $\left\{(f, v), \atop \vdots\right\}$	**Collection** $\left\{\left\{(f, v), \atop \vdots\right\}, \atop \vdots, (f, v), \atop \vdots\right\}$	Concept $\left\{\left\{\left\{(f, v), \atop \vdots\right\}, \atop \vdots, (f, v), \atop \vdots\right\}, \atop \vdots\right\}$

Two types of concepts we use to demonstrate Cooperative Inference are shown with the unit of analysis in bold. In object-level concepts, an example is a set of feature-value pairs (called specifications) and a concept is a set of examples. In collection-level concepts, objects are identical to the object-level examples; collections are sets of objects along with collection-level specifications; and a concept is a set of collections.

examples labeled in the concept). Once a cooperative learner is given a teaching sample with only one concept label (either positive or negative), the learner can infer that the informant has given all examples with that label and thereby determined the concept. [8]

For the second type of inference, we consider the omission of features from an example that forms what we call a *partial example*. A partial example differs from the typical example in that any feature specification may be omitted. When presented with a partial example, cooperative omission implies that it stands in for all consistent (fully-specified) examples. A single partial example can thus effectively stand in for many consistent examples. For instance, the example with only a single specification, 'red', might stand in for both 'red and circle' as well as 'red and square'.

Let a partial example be any subset of a typical example $\mathcal{X}' = \{x' \mid x \in \mathcal{X} \wedge x' \subseteq\}$. When a partial example is given to a cooperative learner, the learner infers a set of matching examples denoted by $\text{Match}(s, C)$:

$$\text{Cons}(x', \mathcal{X}) = \{x \mid x' \subseteq x\} \ , \tag{20}$$

$$\text{Match}(s, C) = \bigcup_{x' \in s} \text{Cons}(x', \mathcal{X}) \ . \tag{21}$$

When a cooperative informant omits features from examples, the learner infers that the the example stands in for all matching examples. Imagine the addition of a third feature to our intuitive example, so that we have 'red', 'square', and 'small'. For concepts such as 'feps are red or feps are small and square', a cooperative informant would use all three

[8] Cooperative omission, together with the unrealizability of some objects, can be used to justify a principle of truth (see e.g. Johnson-Laird, 2001) where the true portion of a concept is prioritized over the false portion.

features, but for others, such as 'feps are red and square', such an informant would simplify the learning process by omitting the feature 'small'. Match() can be viewed as defining the extensional semantics of such omissions.

This use of partial examples allows for a powerful improvement in the efficiency of learning. First, we show that with Cooperative Inference an informant only needs to use features known to be relevant in order to teach a concept. A feature is considered *relevant* if, for at least one pair of differently-labeled instances in the concept, the feature is the only feature to change[9]:

$$\text{Relevant}(c) = \{f \mid (x_0, \mathrm{F}), (x_1, \mathrm{T}) \in c\} \ , \tag{22}$$

$$\text{where} \quad x_0 \triangle x_1 = \{(f, \mathrm{F}), (f, \mathrm{T})\} \ . \tag{23}$$

Theorem: Relevant instance space Given a concept, c, let F_R be a subset of $\mathcal{F}$ that contains all features that are relevant with respect to c and X_R be the set of partial instances formed of F_R, $X_R = \{\mathrm{F}, \mathrm{T}\}^{\mathcal{F}_R}$. Using Cooperative Inference, an informant may successfully teach c by labeling each partial instance $x' \in X_R$ according to any consistent full instance, $x \supseteq x'$, $x \in \mathcal{X}$.

Proof The theorem follows from the definitions. Given an example in the teaching set, $(x', b) \in s$, $x' \in X_R$, consider the set of matching examples $s_{x'} = \{(x, b) \mid x \in \text{Cons}(x', \mathcal{X})\}$. Note that each example in $s_{x'}$ is a superset of x' and so must have the same label for each relevant feature. Assume for the sake of contradiction that two examples in $s_{x'}$ are differently labeled, $(x, \mathrm{F}), (x^*, \mathrm{T}) \in s_{x'}$. We may build a series of examples $x_0, x_1, \ldots x_i$ beginning with $x_0 = x$, and for each step, changing the label for one feature in x_0 to match x^* such that $i = \frac{1}{2}|x \triangle x_i|$. Because, x_0 and x_n are differently labeled, there must exist some i such that $(x_i, b), (x_{i+1}, \overline{b}) \in c$. This implies that the feature f such that $x_i \triangle x_{i+1} = \{(f, \mathrm{F})(f, \mathrm{T})\}$ is relevant and is a contradiction. ■

Therefore, an informant who knows that only 'red' and 'square' are relevant to the concept 'fep', may omit the entire set of irrelevant features and teach a concept using only those two features.

Until this point, our formal discussion of concepts used an extensional sense, where a concept is defined by the set of outputs for all inputs. The alternative, intensional sense, is used to represent concepts as a rule that generates the appropriate output label based on the content of input and is more similar to our intuitive discussion. For example, the concept with intension $c = f_0 \vee f_1$ (i.e. 'feps are red or square') has the following extension over two features

$$f_0 \vee f_1 = \left\{ \begin{array}{l} (\{(f_0, \mathrm{F}), (f_1, \mathrm{F})\}, \mathrm{F}) \\ (\{(f_0, \mathrm{F}), (f_1, \mathrm{T})\}, \mathrm{T}) \\ (\{(f_0, \mathrm{T}), (f_1, \mathrm{F})\}, \mathrm{T}) \\ (\{(f_0, \mathrm{T}), (f_1, \mathrm{T})\}, \mathrm{T}) \end{array} \right\} \ . \tag{24}$$

So, for the example $\{(f_0, \mathrm{F}), (f_1, \mathrm{F})\}$, the extension provides the label F and for the example $\{(f_0, \mathrm{F}), (f_1, \mathrm{T})\}$, the extension provides the label T. However, the extension does not provide a label for examples beyond the set of four in eq. (24).

Definition: Intensional form A *literal* is a negated or unnegated variable, e.g. $l_1 = f$ is true for (f, T) and $l_2 = \overline{f}$ is true for (f, F). A *term* is a conjunction (i.e. 'and') of

[9]Here, $\triangle$ refers to the symmetric difference such that $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

Table 2
Intension and extension in Cooperative Inference

Object-Level	$ext2int_o$ $\bigvee_{x \in c} \left(\bigwedge_{(f,v) \in x} (f = v) \right)$
	$int2ext_o$ $\bigcup_{t \in c} \left\{ \bigcup_{(f,v) \in t} \{(f,v)\} \right\}$
Collection-Level	$ext2int_c$ $\bigvee_{x \in c} \left(\bigwedge_{o \in x} (\exists o_i) ext2int_o(o) \wedge \bigwedge_{(f,v) \in o} (f = v) \right)$
	$int2ext_c$ $\bigcup_{t \in c} \left\{ \bigcup_{o \in t} \{int2ext_o(o)\} \cup \bigcup_{(f,v) \in t} \{(f,v)\} \right\}$

Cooperative Inference indicates a relationship between the intension and extension of a concept for both concept types. This relationship takes the form of a pair of conjugate functions for each type. Here, the subscript o denotes the object-level version of these functions and the subscript c denotes the collection-level version.

literals such as $t = l_0 \wedge l_1 \wedge \dots$ and is false unless *all* of the literals are true. A *clause* is a disjunction (i.e. 'or') of literals such as $cl = l_0 \vee l_1 \vee \dots$ and is true unless all of the literals are false. Each term and clause is associated with a set of literals such that $t = \bigwedge_{l \in L_t} l$ and $cl = \bigvee_{l \in L_{cl}} l$. A concept is in *Disjunctive Normal Form (DNF)* if it is a disjunction of terms such as $(l \wedge l \wedge \dots) \vee (l \wedge l \wedge \dots) \vee \dots$ and a concept is in *Conjunctive Normal Form (CNF)* if it is a conjunction of clauses $(l \vee l \vee \dots) \wedge (l \vee l \vee \dots) \wedge \dots$.

The intensional concept $f_0 \vee f_1$ will label instances outside of eq. (24), whereas the extensional concept is only defined for the included examples. The intension $f_0 \vee f_1$ gives the positive label to the instance $\{(f_0, \mathrm{F}), (f_1, \mathrm{T}), (f_2, \mathrm{F})\}$ and the negative label to the instance $\{(f_0, \mathrm{F}), (f_1, \mathrm{F}), (f_2, \mathrm{F})\}$ while both would be undefined for the extensional definition in eq. (24). Given an extension in the form of a set of examples, an intensional representation can be derived through cooperation and we briefly describe this process.

Cooperative Inference provides a method to derive an intensional concept from an extensional definition. The first method is to collect all of the positive examples and from each, form a term that is true when the specifications for that example are true and false otherwise. At this point, each positive example has an analogous term that evaluates to true for all instances in positive examples. To form an intensional concept, these terms then need to be 'or-ed' together. The resulting concept would be one that is true for all instances in positive examples and false otherwise. Through a similar process, the negative examples can be 'and-ed' together to form a concept that is negative for all of the instances in negative examples and positive otherwise.

If we use the set of examples in eq. (24), we can form an intensional concept. From the positive examples, we have $c = (\overline{f_0} \wedge f_1) \vee (f_0 \wedge \overline{f_1}) \vee (f_0 \wedge f_1)$. In the case where

the provided examples are *partial* examples, the learner can infer the concept label for instances not covered by the examples. If the intensional form of a concept is known, the output label can be predicted for any instance defined over the same features and thus intensional concepts conveniently provide a set of potentially relevant features. Thus, an informant can communicate an intensional concept through the use of partial examples via $ext2int_o$ in table 2.

Just as intensional concepts can be derived from the extensional definition, a sample compatible with Cooperative Inference can be derived from an intensional concept. This function $int2ext_o$ in table 2 implies a logical correspondence between a teaching set for a concept constructed of partial examples and the intensional form of that concept. If an informant has a concept stored intensionally, such as $c = f_1 \vee f_0$ 'feps are red or square', the informant can convert this definition to the teaching sample $\{\{(f_0, \mathrm{T})\}, \{(f_1, \mathrm{T})\}\}$.

This means that logical operations of one form can be leveraged for the other, e.g. simplifying the intensional definition results in an equivalent simplification of the extensional teaching set. For example, the rule used to combine terms and clauses in the classic Quine-McCluskey algorithm (J. C. Roth, 2013)—e.g. 'feps are either red and small or red and not small' → 'feps are red'—can be used to simplify the intension of a concept and then, through Cooperative Inference, the simplified intension can be used to generate a simplified extension.

This equivalence between intensions and extensions based on Cooperative Inference has powerful implications for learning and cultural accumulation of knowledge. For learning, it implies that knowledge obtained by one individual is effectively transferable to another individual in the form labeled partial examples. Effective transfer of knowledge between a pair of individuals is the first step toward obtaining cultural accumulation of knowledge, a point that we will return to and make more explicit in the Discussion.

Cooperative Inference has implications for the representation of knowledge as well. If an informant has a concept in DNF form, a concept can be inferred from the results of using partial examples, $\mathrm{Match}(s, \mathcal{X})$ where $s = int2ext_o(c)$. This process can be reversed with $c = ext2int_o(s)$. This correspondence between the intensional form of a concept and its teaching set suggests that DNF is a natural representation for cooperation. When a concept is stored in this way, cooperation no longer involves searching a space of teaching samples for the solution; rather, the solution is the representation itself. In this way, learning via Cooperative Inference could inform learning biases through representation. Indeed, a DNF-based representational bias is consistent with some of the most successful models and experiments in concept learning (Feldman, 2000; Goodwin & Johnson-Laird, 2011; Goodman et al., 2008).

In summary, whereas bias-first models are incapable of learning in any domain where the bias of the learner cannot be precisely known, Cooperative Inference overcomes this limitation. Cooperative Inference generalizes a notion of cooperation that allows the informant to effectively induce a situation-specific bias that is tailored to the target concept. This means that the complexity of teaching a concept with Cooperative Inference scales according to the complexity of the concept rather than the concept space. Finally, Cooperative Inference points to a fundamental connection between the teaching of concepts and intensional representation of concepts.

Limitations of bias-first models do not apply to Cooperative Inference

We previously leveled two sets criticisms against bias-first models of cooperation. Three results regard the selection of a bias: successful teaching requires a bias, there is no single rational bias, and without coordination bias mismatch becomes a problem. Additionally, because previous models require the informant (and often the learner as well) to consider the entire space of teaching samples and concepts, they are computationally intractable as the number of features increases. Now that we have introduced Cooperative Inference, it is worth discussing how our model avoids these criticisms.

The adoption of bias is key to meaningful learning. Previous models in cognitive science and machine learning adopt a bias in the form of an *a priori* ordering on the concepts, which satisfies the need for a bias, but the question of how the learner and informant choose and coordinate their biases leads to other problems. Because there are many optimal biases, the learner and informant cannot ensure that they select the same bias based on this criteria. In the absence of a method for coordinating the bias, there is no way to guarantee that independently chosen biases support teaching concepts at all.

In Cooperative Inference, rather than adopting an *a priori* ordering on the concepts, *cooperation is the bias*; Cooperative Inference thus trivially satisfies the need for a bias. This same process allows Cooperative Inference to address the other two results. It is not a problem that there is no single optimal bias because both the learner and informant use a notion of cooperation, rather than optimality, to select their biases. And so long as both the learner and informant carry similar notions of cooperation—an assumption common to all models of cooperation of which we are familiar—bias mismatch is not a problem.

Cooperative Inference thus straightforwardly accommodates the problems regarding the selection of bias by integrating the bias into the model of cooperation. Perhaps more interesting is how the resulting bias addresses the problem of computational complexity. The most important level of Cooperative Inference here is the omission of irrelevant features. When an informant includes only the relevant features in partial examples, the informant limits the space of possible concepts to those that are formed from this set of features, which puts a ceiling on the size of this space and thus on computational complexity.

Previous models assume that cooperation is secondary to learning. As a consequence, they first address the basic quandary wherein learning requires a bias. This leads directly into problems when learning from cooperative others, where learning hinges on coordination between the informant and learner. Cooperative Inference offers a powerful way out of this trap: by adopting cooperation as the bias, not only are these problems alleviated, but a much more powerful notion of cooperation emerges.

Cooperative Inference for first-order logic

Models of cooperation for learning have focused on simple domains, often Boolean concepts or even simpler cases. However, researchers have argued for the importance of quantification and composition for a representational system of human learning (e.g. Markman, 1998; Goodman et al., 2008; Kemp, 2012). Indeed, certain concept intensions that depend on these components cannot be represented—and therefore cannot be learned—within the framework of Boolean concepts we have used so far. For example, while "gray

Figure 1. A visual representation of the example in eq. (25) from a collection-level domain.

objects" can be represented as a Boolean concept, the concept "includes at least one gray object", represented by the first-order logic formula $\exists o F_{gray}(o)$, cannot.

In order to use first-order logic concepts, we introduce concepts defined over collections of objects[10]. For instance, consider fig. 1. If this collection were labeled true then that example would be consistent with concepts such as 'includes at least one circle', $\exists o F_{\text{circle}}(o)$, as well as 'includes at least two circles' and 'includes at least one white object'. The problem for Cooperative Inference is, given a particular intension, determining what can be omitted from the corresponding extension and still result in the learner's acquisition of the target concept, assuming both the informant and learner are cooperative.

In this section, we describe how Cooperative Inference applies to quantification and composition. We conclude the section with two applications of Cooperative Inference to first-order logic concepts: Peano arithmetic and natural number.

Quantification. The first step toward expanding Cooperative Inference to cover first-order logic is to incorporate quantification. In logic, quantification allows the representation of the everyday terms "some" (i.e. at least one) and "every". One major difference this necessitates is the consideration of examples as collection of objects rather than single objects. We begin with a slight modification of the notion used for single-object examples.

Let $\mathcal{F}$ be the set of features where each feature f is paired with a set of possible values, V, so that the set of features is $\mathcal{F} = \{(f_0, V_0), (f_1, V_1), \dots\}$. These form *objects* which are analogous to those in the above Boolean treatment. At a higher level, objects can be assembled to form a *collection*; a set of collections then forms a concept. Table 1 illustrates the general form of collection-level concepts.

The following demonstration uses the domain of 6 objects and 2 features, one with 3 substitutive values and another with 2 substitutive values (the entire domain includes $(3 \cdot 2)^6 = 46,656$ distinct examples). An individual example from this domain is represented in fig. 1 as well as the following equation,

$$\left\{\begin{array}{l}\{(f_{\text{shape}}, v_{\text{circle}}), (f_{\text{color}}, v_{\text{gray}})\},\\ \{(f_{\text{shape}}, v_{\text{square}}), (f_{\text{color}}, v_{\text{gray}})\},\\ \{(f_{\text{shape}}, v_{\text{circle}}), (f_{\text{color}}, v_{\text{gray}})\},\\ \{(f_{\text{shape}}, v_{\text{square}}), (f_{\text{color}}, v_{\text{gray}})\},\\ \{(f_{\text{shape}}, v_{\text{square}}), (f_{\text{color}}, v_{\text{gray}})\},\\ \{(f_{\text{shape}}, v_{\text{triangle}}), (f_{\text{color}}, v_{\text{white}})\}\end{array}\right\}. \tag{25}$$

For Boolean concepts, we defined Cooperative Inference at two levels: the level of features (the learner infers that features omitted from examples by an informant are irrelevant) and the level of concept labels (the learner infers that a sample that contains only

[10] This definition is related to others including Kemp's (2012) treatment of "high-level semantic systems".

Table 3
Quantification

	inclusive	exclusive
existential	$\exists o_i \exists o_j (F(o_i)F(o_j))$	$\exists o_i \exists o_j (F(o_i)F(o_j)),\ i \neq j$
universal	$\forall o_i \forall o_j (F(o_i)F(o_j))$	$\forall o_i \forall o_j (F(o_i)F(o_j)),\ i \neq j$

The four quantification terms are best illustrated through logical formulas. Inclusive quantification may apply to the same object while exclusive quantification must each apply to a different object. Existential quantification is satisfied by the existence of a single object. Universal quantification is only satisfied by all objects.

one concept label contains all examples with that concept label). Consider the effect of these levels of Cooperative Inference for teaching the target concept 'includes one or more circle', which includes the example in fig. 1 as well as thousands of others. By omitting unnecessary features and negatively labelled examples, Cooperative Inference would allow an informant to omit the irrelevant feature f_{color} as well as either all positive or all negative examples. This would still require the informant to use either $2^6 = 64$ negative examples or $3^6 - 2^6 = 665$ positive ones, which is both an improvement in efficiency of orders of magnitude and an extremely large number of examples for such a simple concept.

In addition to omitting unnecessary features, a cooperative informant can omit unnecessary objects from collections. The cooperative learner will infer that the incomplete collection stands in for all collections with matching examples—any collection with an object that contains $(f_{\text{shape}}, v_{\text{circle}})$—will be inferred to be in the concept. Returning to the concept 'includes one or more circle', $\exists o F_{\text{circle}}(o)$.[11] Using Cooperative Inference, it can be taught with the single example,[12]

$$\{\{\{(f_{\text{shape}}, v_{\text{circle}})\}\}\} \ . \tag{26}$$

Cooperative inference allows that objects be simplified by omitting irrelevant specifications such as all features other than shape, so that the object in eq. (26) includes only one specification, $(f_{\text{shape}}, v_{\text{circle}})$. Cooperative Inference also allows the examples to be simplified by omitting irrelevant objects, so that the example in eq. (26) includes only one object $\{(f_{\text{shape}}, v_{\text{circle}})\}$. A learner receiving this teaching sample would then infer that the concept includes all examples that match $\{\{(f_{\text{shape}}, v_{\text{circle}})\}\}$, i.e. all collections that include one object that is a circle, $(f_{\text{shape}}, v_{\text{circle}})$. A similar process would apply for any concept with a single quantifier; thus all such concepts can be taught with exactly one example.

When multiple quantifiers are used, exclusive quantification requires that the objects that match each quantifier be different (e.g. without exclusive quantification a single object would satisfy the concept 'includes two or more gray objects') (for a related use of quantification in interactive computation, see Japaridze, 2004). Under Cooperative Inference, after objects are omitted, the remaining objects are matched *exclusively* as in exclusive quantification. This allows Cooperative Inference to operate continuously with examples that do not omit any objects; in such examples each object must be individually matched, as in

[11] For convenience, let $F_{\text{circle}}(o)$ be a function that is true when $(f_{\text{shape}}, v_{\text{circle}}) \in o$ but false otherwise.

[12] Note that the outer brackets that indicate a collection-level concept is the difference between eq. (26) which teaches 'includes one or more circles' and $\{\{(f_{\text{shape}}, v_{\text{circle}})\}\}$ which teaches 'circles'.

Figure 2. An example from a collection-level domain.

exclusive quantification. The treatment of quantifiers here can be compared to the enriched parallel individuation proposal of Le Corre and Carey (2007) where children apply numerals to a set by comparing a set in the world with one held in long-term memory on the basis of a one-to-one mapping.

To illustrate, using exclusive quantification, the following sample would match examples that contained *two or more* gray objects and is equivalent to the intension $\exists o_i \exists o_j F_{\text{gray}}(o_i) F_{\text{gray}}(o_j)$, $i \neq j$

$$\left\{ \left\{ \begin{matrix} \{(f_{\text{color}}, v_{\text{gray}})\}, \\ \{(f_{\text{color}}, v_{\text{gray}})\} \end{matrix} \right\} \right\} . \tag{27}$$

Note that exclusive quantification also applies to objects with different specifications. The following sample teaches the concept 'includes at least one triangle object and at least one different white object' which excludes the example in fig. 1 but includes the example in fig. 2

$$\left\{ \left\{ \begin{matrix} \{(f_{\text{shape}}, v_{\text{triangle}})\}, \\ \{(f_{\text{color}}, v_{\text{white}})\} \end{matrix} \right\} \right\} . \tag{28}$$

The equivalent intension for the the previous concept is $\exists o_i \exists o_j F_{\text{triangle}}(o_i) F_{\text{white}}(o_j)$, $i \neq j$.

Composition and predication. Composition involves using one or more concepts to build another potentially different concept. We define composition at both the collection- and object-level (see table 1). Composition allows concepts to make use of universal quantification—'for all'—and inclusive quantification—where a single object may satisfy multiple quantifiers.

A concept is simply a function that maps any example in a given domain to a set of output labels, one example is the concept 'is square?' that maps objects to the Boolean labels and another is 'contains at least one square' which is true for collections that contain one or more objects that satisfy the feature 'is square?'. Once a concept is taught, it can be used in the future as a feature (e.g. an object-level concept like 'circles' would then become an object-level feature whereas a collection-level concept like 'includes at least one a circle' would become a collection-level feature). When the $c_{\text{new}} \mapsto$ notation is used, it indicates that the concept c_{new} may then be used as a feature.

Consider the concept 'includes no circle objects', corresponding to the formula $\forall o \overline{F_{\text{circle}}(o)}$.[13] In order to teach the concept, an informant first teaches the inverse concept 'includes one or more circle objects' with the sample

$$c_{\text{circle}} \mapsto \{\{\{(f_{\text{shape}}, v_{\text{circle}})\}\}\} . \tag{29}$$

[13] The $\overline{(\cdot)}$ notation indicates negation.

In order to negate c_{circle}, the collection-level feature is paired with the false label $(c_{\text{circle}}, \text{F})$. Such an example will match all collections which are *not* in the concept taught by c_{circle}. Then, through composition, 'includes no circle objects' is taught with

$$\{\{(c_{\text{circle}}, \text{F})\}\} \ . \tag{30}$$

Consider, as another example, the concept 'includes at least one triangle object *and* at least one white object' and note that this concept uses the inclusive quantification (i.e. the concept includes collections with one white triangle). To teach this concept an informant would first decompose the concept into two collection-level features that can be taught using only exclusive quantifications. The following examples would teach 'includes at least one triangle' and 'includes at least one white object' respectively,

$$c_{\text{white}} \mapsto \{\{\{(f_{\text{color}}, v_{\text{white}})\}\}\} \ , \tag{31}$$

$$c_{\text{triangle}} \mapsto \{\{\{(f_{\text{shape}}, v_{\text{triangle}})\}\}\} \ . \tag{32}$$

An informant may then compose c_{white} and c_{triangle} to teach the target concept with the following sample, corresponding to the intension $\exists o_i \exists o_j F_{\text{white}}(o_i) F_{\text{triangle}}(o_j)$,

$$\left\{\left\{\begin{matrix}(c_{\text{white}}, \text{T}), \\ (c_{\text{triangle}}, \text{T})\end{matrix}\right\}\right\} \ . \tag{33}$$

This concept uses inclusive quantification so it could be satisfied by a collection that included a single white triangle and could also be satisfied by a collection that included a non-white triangle and a separate white object.

The final piece required for a model that fully expresses first-order logic is *predication*. A predicate is a term that takes as an input a previously defined concept and outputs a potentially different concept. An example of a predicate is 'contains at least one object of a given shape'.

Here, predication is implemented by first teaching one or more concepts, each with a unique concept label. Then, a predicate concept is taught that may take, as input, other concepts. For instance, an informant might teach a concept for every shape

$$c_{\text{circle}} \mapsto \{\{\{(f_{\text{shape}}, v_{\text{circle}})\}\}\} \ , \tag{34}$$

$$c_{\text{square}} \mapsto \{\{\{(f_{\text{shape}}, v_{\text{square}})\}\}\} \ , \tag{35}$$

$$c_{\text{triangle}} \mapsto \{\{\{(f_{\text{shape}}, v_{\text{triangle}})\}\}\} \ . \tag{36}$$

And then teach the predicate concept

$$P_{\text{shape}}(c) \mapsto \{\{(c, \text{T})\}\}, c \in \{c_{\text{circle}}, c_{\text{square}}, c_{\text{triangle}}\} \ . \tag{37}$$

Cooperative Inference for richly structure concepts

We have noted two shortcomings of previous models of cooperation: the restriction to less expressive representations such as Boolean logic, and to unrealistically simple concepts. With Cooperative Inference, we have shown that it is possible to describe efficient

learning for concepts that require first-order logic for representation but we have so far not demonstrated complex concepts. In what follows we do just that.

The above framework describes a method for arbitrary first-order logic concepts. This is because the model includes both a process for learning the operations corresponding to the following symbols—$P(\cdot)$, $\exists$, $\forall$, $\vee$, $\wedge$, and $\overline{(\cdot)}$—as well as composition, from a cooperative informant. Together, these operations allow for the expression of any first-order logic concept. Finally, table 2 includes the functions for transforming extensional cooperative samples into intensional formulas and vice versa. This means that Cooperative Inference shows how an informant might communicate arbitrary first-order logic concepts *through examples* and how a learner might faithfully recover the intended intension from the examples.

We demonstrate with two cases. The first shows how a system of arithmetic could be learned from a cooperative informant. This demonstrates that Cooperative Inference can be used to teach concepts vastly more complex than those of previous models. The second is less abstract in nature, describing how Cooperative Inference can be applied to natural numbers. This is an example that has been well-studied in the cognitive development literature, and thus illustrates our framework on a learning problem basic, important, and that requires first-order concepts.

Peano arithmetic by examples. The Peano axioms (see Wang, 1957) are a set of axioms for the formalization of arithmetic and the natural numbers. From these axioms, any learner that uses standard rules of logical deduction would be able to derive all the necessary facts about arithmetic (Mendelson, 2009). Here, we use the first-order formulation of the Peano Axioms from Mendelson (2009):

- **S1** $x_1 = x_2 \Rightarrow (x_1 = x_3 \Rightarrow x_2 = x_3)$
- **S2** $x_1 = x_2 \Rightarrow x_1 + 1 = x_2 + 1$
- **S3** $0 \neq x_1 + 1$
- **S4** $x_1 + 1 = x_2 + 1 \Rightarrow x_1 = x_2$
- **S5** $x_1 + 0 = x_1$
- **S6** $x_1 + (x_2 + 1) = (x_1 + x_2) + 1$
- **S7** $x_1 \cdot 0 = 0$
- **S8** $x_1 \cdot (x_2 + 1) = x_1 \cdot x_1 + x_1$
- **S9** $B(0) \Rightarrow (\forall x(B(x) \Rightarrow B(x + 1)) \Rightarrow \forall x B(x))$

Each concept can be taught using composition. First, each first-order logic formula is transformed so that it uses a functional notation. For **S1** this looks like $P_{\Rightarrow}(P_{=}(x_1, x_2), P_{\Rightarrow}(P_{=}(x_1, x_3), P_{=}(x_2, x_3)))$. Each function can then be taught as an

example-level feature:

$$S1.1 \mapsto \{\{(P_{=}(x_1, x_2), \mathrm{T})\}\} \tag{38}$$
$$S1.2 \mapsto \{\{(P_{=}(x_1, x_3), \mathrm{T})\}\} \tag{39}$$
$$S1.3 \mapsto \{\{(P_{=}(x_2, x_3), \mathrm{T})\}\} \tag{40}$$
$$S1.4 \mapsto \{\{(P_{\Rightarrow}(S1.2, S1.3), \mathrm{T})\}\} \tag{41}$$
$$S1 \mapsto \{\{(P_{\Rightarrow}(S1.1, S1.4), \mathrm{T})\}\} \ . \tag{42}$$

The axioms **S1-8** can be taught in this two-step process (refer to the appendix A for detailed derivations for each axiom). The final axiom **S9** is actually an axiom schema, representing a generalized form for an unbounded set of axioms. In order to teach the axiom schema finitely, an informant would need to make use of an even higher order concept—a concept over concepts—corresponding to a higher-order logic. We have given an example of what this might look like at the end of appendix A.

Of course, this demonstration presupposes that learners can, from the basic axioms, deduce the remaining rules of basic arithmetic. Thus, although in theory this could teach a learner arithmetic, we know that human learners are unlikely to actually follow through on this possibility. A less abstract exercise would to assume a more limited learner and a practically relevant set of concepts. In the next section we describe such an approach for the concept of natural numbers.

Natural number by example. There is a rich history of research for the learning and development of numerical concepts (e.g. Gelman & Gallistel, 1978; Fuson, 1988; Wynn, 1990; Carey, 2009). Recently Piantadosi, Tenenbaum, and Goodman (2012) have applied a formal model of representational concept learning to the problem. Still, there is no explanation for how this system of concepts for arithmetic or any similarly complex system could be learned from a cooperative informant (regardless of how learning actually procedes).

In what follows we will present a formal account that closely mirrors one account from Carey (2009). Specifically, Carey argues that natural number is bootstrapped from the primitive abilities to individuate small numbers of objects and the ability to differentiate larger and smaller set sizes for large number of objects. We take the same starting point, and using Cooperative Inference illustrate the learning of a system of concepts for natural number.

Given a object-level feature, f, $C_i(f)$ will be the number function that corresponds to 'includes exactly i objects with f'. We will use the natural numbers (non-negative integers) and we assume that both the informant and learner know the sequence of natural numbers, though the learner does not necessarily know the meaning (for empirical justification of this assumption see e.g. Fuson, Pergament, Lyons, & Hall, 1985).

We begin with $C_1(f)$ which is taught by first teaching the 'greater than or equal to'

concepts $C_{\geq 1}(f)$ and $C_{\geq 2}(f)$ and composing those two concepts to form $C_1(f)$,

$$C_{\geq 1}(f) \mapsto \left\{\left\{\{(f, \mathrm{T})\}\right\}\right\} \tag{43}$$

$$C_{\geq 2}(f) \mapsto \left\{\left\{\begin{matrix}\{(f, \mathrm{T})\},\\ \{(f, \mathrm{T})\}\end{matrix}\right\}\right\} \tag{44}$$

$$C_1(f) \mapsto \left\{\left\{\begin{matrix}(C_{\geq 1}(f), \mathrm{T}),\\ (C_{\geq 2}(f), \mathrm{F})\end{matrix}\right\}\right\} . \tag{45}$$

Then, we use $C_{\geq 3}(f)$ to teach $C_2(f)$,

$$C_{\geq 3}(f) \mapsto \left\{\left\{\begin{matrix}\{(f, \mathrm{T})\},\\ \{(f, \mathrm{T})\},\\ \{(f, \mathrm{T})\}\end{matrix}\right\}\right\} \tag{46}$$

$$C_2(f) \mapsto \left\{\left\{\begin{matrix}(C_{\geq 2}(f), \mathrm{T}),\\ (C_{\geq 3}(f), \mathrm{F})\end{matrix}\right\}\right\} . \tag{47}$$

This process can be extended to any natural number i with a concept schema as in the previous section

$$C_i(f) \mapsto \left\{\left\{\begin{matrix}(C_{\geq i}(f), \mathrm{T}),\\ (C_{\geq i+1}(f), \mathrm{F})\end{matrix}\right\}\right\} . \tag{48}$$

Intuitively, the last sample describes a set of concepts, rather than a single concept. This concept schema can be used to individually teach concepts within the schema or the concept notation can be extended to teach second-order concepts (e.g. concepts over concepts) allowing this schema, corresponding to the remainder of the natural numbers, to be learned *at once* (see **S9** in appendix A).

Psychological grounding

The previous sections have explained the problems with previous models of cooperation and argued that Cooperative Inference is an effective solution to many of these problems. Our demonstrations of Cooperative Inference have taken the form of a formal mathematical description—something previous models are less amenable to—but our exposition has so far lacked grounding in psychological phenomena. We provide that here.

First, we review evidence for the more general phenomenon of cooperation in learning. Nearly all of such research makes use of bias-first models of cooperation and experiments explicitly designed to test such models. The result is that the experiments typically only consider the selection or omission of whole examples. This means that these experiments must include additional mechanisms to make bias-first teaching tractable at all, such as ensuring that both the teacher and learner are aware of the restricted concept space[14]. As such, the results of and predictions tested in these experiments can only be applied to Cooperative Inference in an incomplete way. However, we do consider two studies which provide some evidence in favor of the treatment of features in cooperative inference.

[14]It would be possible to add a mechanism to account for restricted concept spaces, but the point of Cooperative Inference is that such an account alone is incomplete.

Second, we review the evidence for the importance of feature selection and omission to learning in general and cooperative learning in particular. The notion that learning about features is an important part of concept learning is well-established and several mechanisms have been proposed to accomplish this, from feature weights which are determined by the statistical properties of observed examples to overhypotheses—essentially hypotheses about which concepts are more and less likely. Less well-established is how information about which features are important would factor into cooperation. We review proposals about joint attention and gesture, both of which offer realistic mechanisms by which cooperative informants might indicate the relevance of features and are consistent with the Cooperative Inference model.

Existing evidence for models of learning from a cooperative informant

The Cooperative Inference model begins with the assumption that cooperative informants choose to select or omit parts of concepts in order to facilitate the learner's acquisition of the concept. Additionally, the Cooperative Inference model assumes that, when the learner knows that the observed examples are generated through cooperation, the learner will make inferences that are both i) different from those that would be made in the absence of cooperation and ii) facilitate the learning process.

The models and related behavioral demonstrations of strong sampling (Tenenbaum, 1999b) and pedagogical reasoning (Shafto & Goodman, 2008) firmly establish that cooperation can dramatically alter the process of learning and making inferences. They show that when learners know that the source of information is helpful, learners make inferences that are very different from those predicted by models that assume a disinterested source of examples. Additionally, these models, together with supporting behavioral results, have established that omitting examples based on the concept label as Cooperative Inference does (e.g. including only positive or only negative examples) is a successful cooperative strategy across several concept domains.

For instance, Tenenbaum (1999b) considered how people learn and generalize a set of examples when all of the examples are positive. When people receive such examples, they infer a concept that is coextensive with the examples but does not generalize beyond. In the case Tenenbaum used, subjects were given points on a 2d plane and asked to learn an axis-aligned rectangle. Participants inferred a rectangle that, as the number of examples increased, approached smallest rectangle consistent with the received examples.

Consistent with the notion of strong sampling, Xu and Tenenbaum (2007b) reported experiments where children were given examples from hierarchical categories (i.e. ones with a 'basic' level such as 'dogs', a subordinate level such as 'Dalmatians', and a superordinate level such as 'animals'). They found that children tend to generalize to the narrowest level that includes all the observed example, rather than showing a general bias such as a basic-level bias.

Shafto and Goodman (2008) used a similar learning task to Tenenbaum (1999b) except that the examples where chosen by participants who where asked to teach the concept to other participants. When the examples where generated from a teacher as opposed to sampled randomly, the learners behaved quite differently, generalizing to the minimal consistent rectangle on the basis of just two examples. Additionally, teachers seemed to take advantage of this by selecting examples that were diverse with respect to the target

concept—in opposite corners of rectangles. Shafto and Goodman showed that the behavior of both learner and teacher is predicted by a rational model where the teacher predicts the learner's response and selects the set of examples most likely to result in learning the target concept.

Similarly, Rhodes, Brickman, and Gelman (2008) investigated children's attention to diversity of evidence when the evidence was selected by either the teacher or the learner. In a first experiment, they showed that when 5-year-olds were presented with evidence selected by a teacher, they generalized to the category the examples were diverse with respect to, and not beyond. In a second experiment, they showed that when given a choice of two possible samples to provide to a learner, 6-year-olds choose the more diverse sample.

Like previous models, Cooperative Inference predicts effects of the number of examples on the inferred complexity of the target concept. Studies have investigated how the choice, by a knowledgeable and cooperative informant, to provide a single example affects exploration (Bonawitz et al., 2011). Results suggest that, when the evidence is selected by a cooperative informant, learners explore less, consistent with the idea that cooperative selection of examples leads learners to infer simpler concepts.

Because Cooperative Inference focuses on feature omission and the associated implications for hypothesis spaces, direct applications of the model would require working though the additional assumptions most studies of cooperation include to permit whole-example (bias-first) cooperation. As noted previously, previous models that have been applied are consistent with the principles Cooperative Inference (though incomplete without some method of controlling relevant features). However in least two studies of cooperation, the information given by the cooperative informant, is perhaps better understood as about features rather than whole examples.

Buchsbaum et al. (2011) conducted an experiment that looked at the impact of pedagogy on children's evaluation of causal information. First the researchers demonstrated that children use statistical information about cause and effect to guide play. The children saw an adult perform a series of steps (e.g. 'squish, pull knob, rub') followed by an activation of the toy according to the condition. The three conditions varied the sequence that activated the toy: a complete triplet was required, only the last pair was required, and only the last action was required. Buchsbaum et al. found that without pedagogical instructions, across all three conditions, children responded according to the statistics of the input, such as producing more triplets ('ABC') in the 'ABC' condition than '**C'. In contrast, when adult's instructions were pedagogical, children ignored this statistical information and faithfully reproduced the entire triplet.

In the previously discussed study by Bonawitz, Shafto, et al. (2011), an informant either intentionally or accidentally omits all but one feature, where each feature corresponds to a function of a toy. The researchers found that children explored the toy more in the accidental condition than in the intentional condition. Both of these studies are consistent with the idea that in the absence of cooperative information, children use different strategies to infer which features are relevant from the observed examples and we discuss the different forms this takes in the following section. Additionally, these studies indicate that information about features takes on a special importance when features are chosen to be included or excluded cooperatively. When features are intentionally *omitted* by an informant, children are more likely to infer the *irrelevance* of such features, and when features are intentionally

included by an informant, children are more likely to infer the *relevance* of such features.

As noted previously, Cooperative Inference departs from previous models in taking computational complexity, and hence generalizability to realistic problem domains, seriously. Indeed, previous models require non-trivial ad hoc assumptions in order to be applicable, even in these simple experimental scenarios. For example, Shafto and Goodman (2008) use a discretization to approximate the many possible concepts in their scenario. Similarly, the designs in Bonawitz et al. (2011) and Buchsbaum et al. (2011) used a very limited number of features rather than the potentially unbounded set that might realistically apply. However, these studies do make several important points for the Cooperative Inference model: cooperation has a dramatic effect on learning, omission of whole examples by concept label is a successful cooperation strategy, and that feature selection and omission seems to imply information about feature relevance.

Overhypotheses about, and attention to, features

The idea that there are constraints on the set of possible features has a history in learning (Kemp, Perfors, & Tenenbaum, 2007; Kruschke, 1992; Love, Medin, & Gureckis, 2004) and in theories of how cooperation affects learning (Tomasello, Carpenter, Call, Behne, Moll, et al., 2005). From a pure learning perspective, the formal problem is one of restricting the set of candidate features and resembles the notion of overhypotheses (Goodman, 1955; Kemp et al., 2007). Kemp et al. (2007) argues that hierarchical models, which admit hypotheses about hypotheses, provide a formal framework which captures the basic constraining function of overhypotheses. However, hierarchical models, and indeed overhypotheses themselves, only provide traction to the extent that they constrain, and they therefore hinge critically which constraints are posed. It is thus critical to provide an *a priori* explanation for why some constraints are natural while others are not, or run the risk of simply pushing the problem back one level.[15]

One way that hierarchical models may provide traction on the learning problem is in being sensitive to the features that have meaningful variation within a concept. This is the basis of the concept of attention, as discussed in the concept learning literature. For instance, Nosofsky (1986) argued that people tend to allocate attention in such a way that optimizes performance on the task at hand, whether identification (which requires dividing attention among essentially all dimensions) or classification (which only requires dividing attention only among dimension relevant to the classification). Billman and Knutson (1996) argued for the importance of systematicity—predictive relationships between the values of features—for learning the relevance of features. These principles have been captured directly and indirectly in computational models of concept learning (Kruschke, 1992; Love et al., 2004).

Although undeniably important, attention alone cannot explain the benefits of cooperation for learning. As shown in 'Impossible for unknown bias', the benefits of cooperation derive from avoiding mismatches in the set of relevant features, and these need not be constrained to all or only features that who satisfy specific statistical regularities.

[15]This was the problem that Goodman (1955) set out to address. As Kemp et al. (2007) note, hierarchical models do not necessarily directly embody a solution to Goodman's problem.

Joint attention and gesture

One of the primary functions of *joint attention* is to facilitate learning in cases where mere attention will not suffice. These may include situations where there are too many potentially relevant features, statistical regularities are not strong, new features need to be learned, or when the concept is especially important. Indeed, Tomasello and colleagues (Tomasello, 2000; Tomasello et al., 2005; Tomasello, Kruger, & Ratner, 1993) have argued that triadic joint attention is especially important for explaining transmission of goal-directed concepts like tool use. However, these accounts stop short of either providing a mechanism by which triadic joint attention solves the learning problem or how it addresses the complexity problems circumvented by our notion of feature relevance.

Gesture offers mechanism by which cooperative informants may indicate which features are relevant (and by omission, which are not). Research shows that gestures very commonly accompany learning opportunities (Baggett, 1984; Church, Ayman-Nolley, & Mahootian, 2004; Mayer & Anderson, 1991), especially in educational situations (Alibali & Nathan, 2007; Church et al., 2004; Flevares & Perry, 2001), and gestures facilitate learning and problem solving (Goldin-Meadow, Kim, & Singer, 1999; W.-M. Roth, 2001; Singer & Goldin-Meadow, 2005). Research specifically highlights the role of gesture in indicating problem-relevant features. For example, Valenzeno, Alibali, and Klatzky (2003) showed preschool children videos about symmetry where the lesson included examples with verbal information or verbal plus gesture information. The gestures included pointing and tracing gestures that highlighted the features that are relevant to the concept symmetry. Results suggested that children in the verbal plus gesture condition were better able to identify symmetric and asymmetric shapes at post test, consistent with the idea that gestures that highlight problem relevant features lead to better learning. Other evidence suggests that explicitly training children to themselves make gestures that highlight relevant features (grouping strategies in solving math problems) learned more than children who produced partially-correct or no gestures (Goldin-Meadow, Cook, & Mitchell, 2009). Moreover, in this work, gestures are not merely reinforcing concepts expressed in language, but help most when they differ from that which is being said (Goldin-Meadow et al., 2009; Singer & Goldin-Meadow, 2005). This critically indicates that the benefit provided by gesture is unlikely to be merely emphasis on parts of verbal expression but instead an altogether distinct contribution. Indeed, theorists have gone so far as to argue that gesture grounds thought in experience (Alibali & Nathan, 2012; Goldin-Meadow & Beilock, 2010), which is broadly consistent with the possibility that gestures link concepts with examples. Regardless, this literature provides strong support for the idea that gesture functions, in part, by indicating relevant features and this facilitates learning.

Theorists such as Csibra and Gergely (2009) amd Tomasello (2000) emphasize the role of cooperative information transmission in *early* cognitive development. The gesture literature has mainly focused on experiments with older (school-aged) children. Therefore, this evidence does not support for claims about early child development or about the foundational role of reasoning about others in early cognitive development. Is there evidence that such a mechanism is available and used to indicate feature relevance to infants? Infant-directed speech (motherese, e.g. Cooper and Aslin, 1990; Kuhl et al., 1997) and action (motionese, e.g. Brand, Baldwin, and Ashburn, 2002) provide candidate possibilities.

Infant directed speech ("motherese") is characterized by changes that have been hypothesized to facilitate learning (Kuhl et al., 1997). While the precise details of this argument are hotly debated (Cristia, 2013; De Boer & Kuhl, 2003; Eaves, Feldman, Griffiths, & Shafto, Under Revision; Kirchhoff & Schimmel, 2005; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013), the idea that speakers may be increasing variability to draw attention to aspects of the stimuli generally accepted[16]. Similarly, evidence suggest that "motionese" may also show an increase in variability for problem relevant features (Brand et al., 2002).

These provide possible mechanisms by which informants may indicate the relevant features of learning problems to very young children. Regardless of whether these possibilities work out, critical role of feature relevance in Cooperative Inference generates the prediction that, for reasoning about others to play the explanatory role posited in theories of cognitive development, there must be a mechanism for indicating to very young children, which features are relevant.

Summary

Cooperative inference captures the basic observation that there are many potential features, only a subset of which are relevant at any time. It departs from more standard notions of attention by not focusing on bottom-up statistical regularities. Instead, it offers a formalization of how something akin to joint attention may facilitate concept learning—by narrowing the set of candidate features to only those that are relevant for current purposes. Given that feature omission plays a critical role in the efficiency of Cooperative Inference, one may ask how psychologically this would be implemented. For cases where features are present or absent, then omission may simply involve selecting an example where the value is set to absent. More generally, there are a variety of ways to indicate that particular features are relevant, e.g. through pointing, richer gestures, and motionese.

In addition to these connections to empirical psychological literatures, there are theoretical contributions that speak to the nature of concepts and representation. Our approach allows new concepts to be created and used as primitives (i.e. "concepts as features"). This provides a natural way to characterize construction of systems of concepts through compositionality (similar to Goodman et al., 2008; see aslo Schyns and Rodet, 1997). There are also more speculative connections that hinge on origins of representational biases. As proposed, Cooperative Inference could be interpreted as a principle for deriving situation specific biases that is only engaged when learning from cooperative others. A more provocative possibility is that humans are optimized for learning from others, and the representational bias implied by Cooperative Inference—Disjunctive Normal Form—is applied in *all* situations and the assumptions about omission of information are applied situationally. Indeed, there is a developing consensus from studies on the complexity of learning that learning complexity very much follows DNF. Feldman (2000) conducted what is likely the concept learning experiment with the widest scope to date, systematically investigating the difficulty of learning 41 distinct concepts. The *simplicity hypothesis*—the idea that when a person learns a concept, that person tends to learn the simplest sentence under the available representation—provides a complexity of learning prediction for all representations and

[16]It is important to note that there are changes in concept-irrelevant features too, e.g. pitch, but that these are argued to be cues to the pedagogical nature of the accompanying input (Csibra & Gergely, 2009).

the prediction provided by DNF outperforms all others[17]. Furthermore, DNF has become something of a consensus choice for a representation in models of concept learning (Goodman et al., 2008; Goodwin & Johnson-Laird, 2011; Kemp, 2012). This proposal is highly speculative, but demonstrates the broad potential implications of Cooperative Inference for empirical and theoretical work in psychology.

Discussion

Through a set of related results we have shown that bias-first models are unlikely to explain cooperation's effects on learning because they depend on intractable computations and the existence of a shared bias. First, bias-first models are computationally intractable because they require that both the informant and learner consider every concept in the concept space. Second, the assumption that informants know the learner's bias requires *a priori* agreement about prior biases in order to be successful, which creates a quandary: in the absence of bias, learning can be no more efficient than randomly guessing among consistent concepts; a principle of rationality cannot be used to determine the learner's bias, as there are many equivalently optimal biases; and, for large concept spaces, if the informant and learner's biases differ even slightly, it becomes impossible for bias-first models to successfully teach any concept.

We then introduced a novel framework, which we call Cooperative Inference. It is computationally efficient, functions in the absence of an *a priori* bias, and even leads to an *a priori* choice of representation language for cooperative (and potentially non-cooperative) settings. We show that cooperation provides a means by which information can be omitted, yet lead to stronger learning.

The approach builds off previous models that used selective omission of labeled objects (Shafto & Goodman, 2008; Shafto et al., 2014). By the common formalism, concepts are structured with feature-value pairs at bottom; objects are sets of feature-value pairs; collections are sets of objects; and a concept maps a set of examples (either objects or collections) onto the concept labels True and False. Previous models have only considered cooperation as omission of labeled objects. In Cooperative Inference, choosing to omit features implies to the learner that those features are irrelevant. This allows for a radical reduction in the computational complexity of learning by limiting the set of concepts that are potentially relevant. It then becomes possible to teach complex concepts, even when the concept domain is not bounded. Further, because cooperative omission allows the instantiation of intensions extensionally, the informant need not know the bias of the learner. Whereas the effects of cooperative informants on learning had previously not been explored for first-order logic concepts, we have shown that Cooperative Inference applies to this domain and leads to predictions on two interesting cases: Peano arithmetic and natural number. These demonstrations illustrate that the framework can be applied in domains that begin to approximate the richness of human conceptual knowledge.

[17] See the analysis using 'mental models' by Goodwin and Johnson-Laird (2011) and the commentary by Feldman (2012).

Limitations

These demonstrations are subject to important limitations, pointing to opportunities for future work. In order to efficiently teach a system of concepts, we assumed that the learner made use of a valid system of deduction. It is implausible that people learn much purely through deduction. Our demonstration should therefore be viewed as a demonstration of teachability rather than a prediction about human behavior, and future work should provide a more detailed treatment that makes closer contact with relevant empirical literature. Second, while first-order logic provides a substantial increase in expressiveness over the more common propositional logic, we still had to make use of a yet higher order logic in order to teach arithmetic through the Peano Axioms. Specifically, future work should look to applying Cooperative Inference to even more expressive representations, such as relational concepts and partial-recursive functions. Third, we have also focused on learning domains where the informant has no information about the learner's beliefs. Thus, our result functions as a default bias for such learning domains. In the future, models will need to account for domains in which learners are not completely naive, and informants are privy to this. Accomplishing this will lead us to also relax our deterministic account to deal with uncertainty (e.g. about prior knowledge), which would allow us to account for variability in human behavior.

Existing models of learning from cooperative, knowledgeable informants in cognitive science have mainly adopted a probabilistic approach. Though we have not focused on this setting, our results cover these models, and our proposed solution is closely related to intuitions underlying previous models. In our analysis, we adopt deterministic inference over ordered sets of concepts. Probabilistic models assume a prior bias, which is precisely an assumption about the *a priori* ordering of concepts. Moreover, the mechanics of belief updating in probabilistic frameworks obey the basic assumptions of our deterministic approach—consistency and class preservation. Thus, the limitations identified for bias-first models apply to probabilistic models broadly.

Although our framework approach is more removed from the details of human behavior than is typical of psychological modeling, we believe the results bear out the benefits of this approach. Recent research has critiqued more abstract computational-level and rational models as offering little in the way of predictive power (Bowers & Davis, 2012; Jones & Love, 2011; Marcus & Davis, 2013). Here, by abstracting away from some of the details necessary to ensure a psychologically plausible account, the results are stronger and more general. We define a class of models—bias-first models, which encompasses many, if not all of the models in the learning literature—and show that they cannot provide an adequate account of cooperative learning from knowledgeable informants without *a priori* agreement between the informant and the learner. Moreover, we show that even relatively minor variance between the informant's assumed and the learner's true bias leads to failure given a sufficiently large set of candidate concepts. This failure is an inevitability without strong constraints on the set of potentially relevant features. We show that cooperation provides a mechanism to allow strong inference based on omission of features, which in turn reduces the concept space, and allows a form of situation-specific learning bias. By putting cooperation first, we show that the problems associated with bias-first models can be avoided.

Connecting existing psychological theories

Cooperation plays an important role in influential theories of human nature. Tomasello (1999) argues that the ability to cooperatively learn explains our ability to accumulate information over generations in ways that other animals do not—the *cultural ratchet effect*. Boyd et al. (2011) proposed that humans occupy a cultural niche and that human success is owed to the unique ability to learn from others. These claims seem to necessitate a model of efficient learning through cooperation. Previous models of cooperative learning have suggested that learning can be qualitatively changed by the availability of knowledgeable and cooperative informants, and provided initial support for the broader idea of humans inhabiting a cultural niche designed for cooperative learning. However, these accounts relied on implausibly strong assumptions—such as informants with perfect knowledge of learners' biases and unlimited computational resources—rendering them ineffective for even unreasonably simple concepts.

Psychological theorists have offered detailed proposals for how cooperative learning can lead to accumulated changes in human nature. Csibra and Gergely's theory of natural pedagogy (Csibra & Gergely, 2006, 2009) proposes two mechanisms by which cooperation facilitates learning: relevance and generalizability. Relevance represents the assumption that the information that has been communicated by the informant is pertinent and communicated efficiently (Sperber & Wilson, 1986). Generalizability represents an assumption that information selected to be conveyed applies broadly. Relatedly, Tomasello et al. (2005) has proposed a more basic notion of cooperation based on triadic joint attention. On this account, the key ability facilitating cooperative learning is the ability for an informant and learner to jointly direct attention to the objects or parts of objects most useful for the purpose of learning.

Cooperative Inference is closely related to these psychological theories. Cooperative Inference provides logical basis for relevance (limited to the context of our learning problems): the mutual goal of isolating a concept implies that omitted information can be assumed to be irrelevant. The notion of feature omission used here can be viewed as the complement of triadic joint attention proposed by Tomasello. The generalizability used by Csibra and Gergely requires abstracting away from the details of the current situation. In typical accounts of learning, generalizing from a single example is nigh impossible due to the fact that any object can be a member of many categories. Cooperative inference, by allowing inferences about omission of features, results in a straightforward account of generalizability based on internalizing the relevant aspects of the situation.

From cooperative inference to cultural evolution

Indeed, computational modelers have investigated the conditions under which we may expect cultural transmission to result in accumulation of knowledge (Beppu & Griffiths, 2009; Griffiths & Kalish, 2007). In general, mere transmission of examples from one person to another, as in a game of telephone, is not enough (Griffiths & Kalish, 2007). Results show that when examples are used to transmit a concept through a chain of learners, the result is total loss of information. This occurs because every individual who draws an inference distorts the information slightly with their own bias, and therefore does not pass on the observations completely veridically. The end result converges to people's *a priori* biases.

This is only marginally improved when each person can directly observe their own evidence (Beppu & Griffiths, 2009). A sufficient condition for cultural ratchet effect is *posterior passing*, the ability to perfectly convey one person's beliefs to the next, together with each person observing their own information (Beppu & Griffiths, 2009).

Cooperative inference satisfies the conditions for achieving the cultural ratchet. The equations in table 2 describe a logical relation between the instantiation of concepts in examples and inference from examples to concepts. Whereas Beppu and Griffiths (2009) proposed natural language as a mechanism to allow posterior passing, our results suggest that Cooperative Inference—which allows for the communication of an intension directly through extensional examples—is sufficient to achieve cultural accumulation of knowledge. This provides support for existing psychological accounts, which focus on prelinguistic abilities as mechanisms for cultural accumulation (Csibra & Gergely, 2009; Tomasello, 1999; Tomasello et al., 2005).

An *a priori* analysis of representation

In addition to providing support for theories of cultural accumulation, Cooperative Inference points to a natural basis for representing concepts. Standard approaches to learning must select a bias, but it is quite difficult to find an objective basis on which to do so. Variously, models have adopted neuron-like representations (McClelland & Rumelhart, 1988; Rogers & McClelland, 2005), logical representations of various forms (Feldman, 2000; Goodman et al., 2008; Kemp, 2012), as well as more narrowly-scoped proposals such as prototypes (Rosch, 1973; Rosch & Mervis, 1975) and exemplars (Kruschke, 1992; Nosofsky & Palmeri, 1997). Each of these (combined with many more detailed assumptions) implies a bias. For most models, due to fundamental ambiguities in our knowledge of the structure of neurons and the objective features of the external world, the choice can only be loosely justified. One of the strongest approaches is *rational analysis*, which suggests that one begin by analyzing the problem in the world, and assume that internal representations are adapted to solving the problem. This method has been used to justify choice of biases across many domains (Anderson, 1990; Chater & Oaksford, 1999; Tenenbaum, Griffiths, & Kemp, 2006) However, even this method is highly controversial (Bowers & Davis, 2012; Jones & Love, 2011; Marcus & Davis, 2013).

Our approach is both similar and different. What we have done can be viewed as a form of computational-level analysis where the focus is on the social environment rather than the external environment (see also Christiansen & Chater, 2008). Rather than adapting to some assumed properties of the external environment, we have analyzed the possibility that we are adapted to cooperate with others. This approach avoids some of the problems of previous approaches. Unlike rational analysis based on the external environment, our basic formalization of cooperative learning—as the a mutual goal of inferring a concept—is common across many formalisms (Benz, Jäger, & van Rooij, 2005; Shafto & Goodman, 2008; Zilles et al., 2008), and thus we take it to be relatively uncontroversial. Moreover, unlike some other aspects of our physical or linguistic environment, the presence of cooperative others available to support learning is a nearly universal feature of human societies that has been stable for a very long time (Boyd et al., 2011).

Perhaps the most intriguing consequence of our framework is that it leads to a particular choice of representation language. Our results indicate that Disjunctive Normal Form

(DNF) is a natural choice for a representation in which to represent Boolean and first-order concepts where cooperation is important. Interestingly, empirical and computational investigations into Boolean concept learning converge on a learning bias that is closely predicted by representational complexity of concepts represented in DNF form (Goodman et al., 2008; Goodwin & Johnson-Laird, 2011; Kemp, 2012). This is consistent with the possibility that people's default learning biases are adapted to the efficient transmission of information between cooperative others. If true, this would provide strong evidence that we indeed inhabit a cultural niche (Boyd et al., 2011).

Conclusions

The idea that learning from cooperative informants is central to human learning and human culture appears across theories of cognition, cognitive development, and cultural anthropology. Previous formal accounts have fallen short of a complete explanation of how this may be. We have presented a critique of existing models and provide a framework on which a positive account of learning through cooperation may be built. Our approach generalizes the idea of cooperative omission of information, resulting in a precise account of how cooperative information transmission may lead to rapid learning in richly structured, even unbounded, conceptual domains. Our framework instantiates many of the assumptions of previous psychological theories in a precise mathematical framework, and satisfies conditions for cultural transmission. Provocatively, it also suggests the possibility that people's learning biases are adapted for learning from others. To our knowledge, this is the first formal account of how cooperation may lead to cultural accumulation of knowledge and cultural niche based on social learning. Considerable work remains but we take this to be a positive first step to providing a computationally precise link between the cooperation, culture, and the accumulation of richly structured knowledge.

References

Alibali, M. W. & Nathan, M. J. (2007). Teachers' gestures as a means of scaffolding students' understanding: evidence from an early algebra lesson. *Video research in the learning sciences*, 349–365.

Alibali, M. W. & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: evidence from learners' and teachers' gestures. *Journal of the Learning Sciences*, *21*(2), 247–286.

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(03), 471–485.

Angluin, D. (1988). Queries and concept learning. *Machine learning*, *2*(4), 319–342.

Angluin, D. & Kric kis, M. (1997). Teachers, learners and black boxes. In *Proceedings of the tenth annual conference on computational learning theory* (pp. 285–297). ACM.

Anthony, M., Brightwell, G., Cohen, D., & Shawe-Taylor, J. (1992). On exact specification by examples. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 311–318). ACM.

Baggett, P. (1984). Role of temporal overlap of visual and auditory material in forming dual media associations. *Journal of Educational Psychology*, *76*(3), 408.

Balbach, F. J. (2008). Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, *397*(1), 94–113.

Beal, J. & Roberts, J. (2009). Enhancing methodological rigor for computational cognitive science: complexity analysis. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 99–104).

Benz, A., Jäger, G., & van Rooij, R. (2005). *Game theory and pragmatics* (A. Benz, G. Jäger, & R. van Rooij, Eds.). Palgrave Macmillan.

Beppu, A. & Griffiths, T. L. (2009). Iterated learning and the cultural ratchet. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2089–2094). Citeseer.

Billman, D. & Knutson, J. (1996). Unsupervised concept learning and value systematicitiy: a complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 458.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: why social learning is essential for human adaptation. *PNAS*, *108*, 10918–10925.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, *5*(1), 72–83.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. Transaction Books.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Butler, L. P. & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child Development*, *83*(4), 1416–1428.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Chater, N. & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558.

Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: does gesture enhance learning? *International Journal of Bilingual Education and Bilingualism*, *7*(4), 303–319.

Cobham, A. (1965). The intrinsic computational difficulty of functions. In *Proceedings of the 1964 congress for logic, methodology, and the philosophy of science* (pp. 24–30).

Cooper, R. P. & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child development*, *61*(5), 1584–1595.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. The MIT Press.

Cristia, A. (2013). Input to language: the phonetics and perception of Infant-Directed speech. *Language and Linguistics Compass*, *7*(3), 157–170.

Csibra, G. (2007). Teachers in the wild. *Trends in cognitive sciences*, *11*(3), 95–96.

Csibra, G. & Gergely, G. (2006). Social learning and social cognition: the case for pedagogy. *Processes of Change in Brain and Cognitive Development. Attention and Performance*, *21*, 249–274.

Csibra, G. & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

De Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134.

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (Under Revision). Infant-Directed speech is consistent with teaching. *Psychological Review*.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2012). Symbolic representation of probabilistic worlds. *Cognition*, *123*(1), 61–83.

Flevares, L. M. & Perry, M. (2001). How many do you see? the use of nonspoken representations in first-grade mathematics lessons. *Journal of Educational Psychology*, *93*(2), 330.

Fuson, K. C. (1988). *Children's counting and concepts of number*. Springer-Verlag.

Fuson, K. C., Pergament, G. G., Lyons, B. G., & Hall, J. W. (1985). Children's conformity to the cardinality rule as a function of set size and counting accuracy. *Child Development*, 1429–1436.

Gelman, R. & Gallistel, C. (1978). *The child's understanding of number*. Harvard University Press.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755.

Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gold, E. M. (1967). Language identification in the limit. *Information Control*, *10*(5), 447–474.

Goldin-Meadow, S. & Beilock, S. L. (2010). Action's influence on thought: the case of gesture. *Perspectives on Psychological Science*, *5*(6), 664–674.

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*(3), 267–272.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology*, *91*(4), 720.

Goldman, S. A. & Mathias, H. D. (1993). Teaching a smarter learner. In *Proceedings of the sixth annual conference on computational learning theory* (pp. 67–76). ACM.

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

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of Rule-Based concept learning. *Cognitive Science*, *32*(1), 108–154.

Goodwin, G. P. & Johnson-Laird, P. N. (2011). Mental models of boolean concepts. *Cognitive psychology*, *63*(1), 34–59.

Griffiths, T. L. & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, *29*(2), 147–160.

Japaridze, G. (2004). Computability logic: a formal theory of interaction. In D. Goldin, S. A. Smolka, & P. Wegner (Eds.), *Interactive computation: the new paradigm*. Springer.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, *5*(10), 434–442.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(04), 169–188.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological review*, *119*(4), 685–722.

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

Kirchhoff, K. & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238–2246.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.

Le Corre, M. & Carey, S. (2007). One, two, three, four, nothing more: an investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, *111*(2), 309.

Marcus, G. F. & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.

Markman, A. B. (1998). *Knowledge representation*. Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.

Mayer, R. E. & Anderson, R. B. (1991). Animations need narrations: an experimental test of a dual-coding hypothesis. *Journal of educational psychology*, *83*(4), 484.

McClelland, J. L. & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: a handbook of models, programs, and exercises*. MIT press.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378.

Mendelson, E. (2009). *Introduction to mathematical logic*. CRC press.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, *104*(2), 266.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53–79.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: evidence for developmental change. *Cognition*, *108*(2), 543–556.

Rogers, T. T. & McClelland, J. L. (2005). A parallel distributed processing approach to semantic cognition: applications to conceptual development. *Building object categories in developmental time*, 335–387.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories.

Rosch, E. H. & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Roth, J. C. (2013). *Fundamentals of logic design*. Cengage Learning.

Roth, W.-M. (2001). Gestures: their role in teaching and learning. *Review of Educational Research*, *71*(3), 365–392.

Schyns, P. G. & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 681.

Shafto, P. & Goodman, N. D. (2008). Teaching games: statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society* (pp. 1632–1637).

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Shinohara, A. & Miyano, S. (1991). Teachability in computational learning. *New Generation Computing*, *8*, 337–347.

Singer, M. A. & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*(2), 85–89.

Sperber, D. & Wilson, D. (1986). *Relevance: communication and cognition*. Harvard University Press.

Tenenbaum, J. B. (1999a). *A bayesian framework for concept learning.* (Doctoral dissertation, Massachussets Institute of Technology).

Tenenbaum, J. B. (1999b). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 59–65).

Tenenbaum, J. B. & Griffiths, T. L. (2001a). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Tenenbaum, J. B. & Griffiths, T. L. (2001b). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1036–1041).

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.

Tomasello, M. (1999). The human adaptation for culture. *Annual review of anthropology*, 509–529.

Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., et al. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–690.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and brain sciences*, *16*(03), 495–511.

Topál, J., Gergely, G., Miklósi, Á., ErdH ohegyi, Á., & Csibra, G. (2008). Infants perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*(5897), 1831–1834.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: a lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187–204.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: computationally easy or difficult? *Frontiers in Human Neuroscience*, *5*(52), 1–18.

Wang, H. (1957). The axiomatization of arithmetic. *The Journal of Symbolic Logic*, *20*(2), 145–158.

Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. Wiley New York.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.

Xu, F. & Tenenbaum, J. B. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Xu, F. & Tenenbaum, J. B. (2007b). Word learning as bayesian inference. *Psychological Review*, *114*(2).

Zilles, S., Lange, S., Holte, R., & Zinkevich, M. (2008). Teaching dimensions based on cooperative learning. In *Colt 2008* (pp. 135–146).

Appendix
Peano Axioms continued

Earlier, we demonstrated how the first of the Peano axioms would be taught using the first-order logic version of Cooperative Inference. Here, we include the remainder of the axioms.

S2 $x_1 = x_2 \Rightarrow x_1 + 1 = x_2 + 1$

$$
\begin{aligned}
S2.1 &\mapsto \{\{(P_{=}(x_1, x_2), \mathrm{T})\}\} \\
S2.2 &\mapsto \{\{(P_{+}(x_1, 1), \mathrm{T})\}\} \\
S2.3 &\mapsto \{\{(P_{+}(x_2, 1), \mathrm{T})\}\} \\
S2.4 &\mapsto \{\{(P_{=}(S1.2, S1.3), \mathrm{T})\}\} \\
S2 &\mapsto \{\{(P_{\Rightarrow}(S2.1, S2.4), \mathrm{T})\}\}
\end{aligned}
$$

S3 $0 \neq x_1 + 1$

$$
\begin{aligned}
S3.1 &\mapsto \{\{(P_{+}(x_1, 1), \mathrm{T})\}\} \\
S3 &\mapsto \{\{(P_{=}(S3.1, 0), \mathrm{F})\}\}
\end{aligned}
$$

S4 $x_1 + 1 = x_2 + 1 \Rightarrow x_1 = x_2$

$$
\begin{aligned}
S4.1 &\mapsto \{\{(P_{+}(x_1, 1), \mathrm{T})\}\} \\
S4.2 &\mapsto \{\{(P_{+}(x_2, 1), \mathrm{T})\}\} \\
S4.3 &\mapsto \{\{(P_{=}(S1.2, S1.3), \mathrm{T})\}\} \\
S4.4 &\mapsto \{\{(P_{=}(x_1, x_2), \mathrm{T})\}\} \\
S4 &\mapsto \{\{(P_{\Rightarrow}(S4.3, S4.4), \mathrm{T})\}\}
\end{aligned}
$$

S5 $x_1 + 0 = x_1$

$$
\begin{aligned}
S5.1 &\mapsto \{\{(P_{+}(x_1, 0), \mathrm{T})\}\} \\
S5 &\mapsto \{\{(P_{=}(S5.1, x_1), \mathrm{T})\}\}
\end{aligned}
$$

S6 $x_1 + (x_2 + 1) = (x_1 + x_2) + 1$

$$
\begin{aligned}
S6.1 &\mapsto \{\{(P_{+}(x_2, 1), \mathrm{T})\}\} \\
S6.2 &\mapsto \{\{(P_{+}(x_1, S6.1), \mathrm{T})\}\} \\
S6.3 &\mapsto \{\{(P_{+}(x_1, x_2), \mathrm{T})\}\} \\
S6.4 &\mapsto \{\{(P_{+}(S6.3, 1), \mathrm{T})\}\} \\
S6 &\mapsto \{\{(P_{\Rightarrow}(S6.2, S6.4), \mathrm{T})\}\}
\end{aligned}
$$

S7 $x_1 \cdot 0 = 0$

$$
\begin{aligned}
S7.1 &\mapsto \{\{(P.(x_1, 0), \mathrm{T})\}\} \\
S7 &\mapsto \{\{(P_{=}(S7.1, 0), \mathrm{T})\}\}
\end{aligned}
$$

S8 $x_1 \cdot (x_2 + 1) = x_1 \cdot x_1 + x_1$

$$
\begin{aligned}
S8.1 &\mapsto \{\{(P_{+}(x_2, 1), \mathrm{T})\}\} \\
S8.2 &\mapsto \{\{(P_{\cdot}(x_1, S8.1), \mathrm{T})\}\} \\
S8.3 &\mapsto \{\{(P_{\cdot}(x_1, x_2), \mathrm{T})\}\} \\
S8.4 &\mapsto \{\{(P_{+}(S8.3, x_1), \mathrm{T})\}\} \\
S8 &\mapsto \{\{(P_{\Rightarrow}(S8.2, S8.4), \mathrm{T})\}\}
\end{aligned}
$$

S9 $B(0) \Rightarrow (\forall x(B(x) \Rightarrow B(x + 1)) \Rightarrow \forall x B(x))$

$$
\begin{aligned}
S9.1 &\mapsto \{\{(P_B(0), \mathrm{T})\}\} \\
S9.2 &\mapsto \{\{(P_B(x_1), \mathrm{T})\}\} \\
S9.3 &\mapsto \{\{(P_{+}(x_1, 1), \mathrm{T})\}\} \\
S9.4 &\mapsto \{\{(P_B(S9.3), \mathrm{T})\}\} \\
S9.5 &\mapsto \{\{(P_{\Rightarrow}(S9.2, S9.4), \mathrm{T})\}\} \\
S9.6 &\mapsto \{\{(P_B(x_2), \mathrm{T})\}\} \\
S9.7 &\mapsto \{\{(P_{\Rightarrow}(S9.5, S9.6), \mathrm{T})\}\} \\
S9 &\mapsto \{\{(P_{\Rightarrow}(S9.1, S9.7), \mathrm{T})\}\}
\end{aligned}
$$