COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2017) 1–39
Copyright © 2017 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12496

Naïve and Robust: Class-Conditional Independence in Human Classification Learning*

Jana B. Jarecki,[a,b] Björn Meder,[b] Jonathan D. Nelson[b,c]

[a]Department of Psychology, University of Basel
[b]Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development
[c]School of Psychology, University of Surrey

Received 6 March 2016; received in revised form 19 September 2016; accepted 18 November 2017

Abstract

Humans excel in categorization. Yet from a computational standpoint, learning a novel probabilistic classification task involves severe computational challenges. The present paper investigates one way to address these challenges: assuming class-conditional independence of features. This feature independence assumption simplifies the inference problem, allows for informed inferences about novel feature combinations, and performs robustly across different statistical environments. We designed a new Bayesian classification learning model (the dependence-independence structure and category learning model, DISC-LM) that incorporates varying degrees of prior belief in class-conditional independence, learns whether or not independence holds, and adapts its behavior accordingly. Theoretical results from two simulation studies demonstrate that classification behavior can appear to start simple, yet adapt effectively to unexpected task structures. Two experiments—designed using optimal experimental design principles—were conducted with human learners. Classification decisions of the majority of participants were best accounted for by a version of the model with very high initial prior belief in class-conditional independence, before adapting to the true environmental structure. Class-conditional independence may be a strong and useful default assumption in category learning tasks.

Keywords: Classification; Class-conditional independence; Learning; Naïve Bayes; Markov property; Bayesian model; Probabilistic inference; Heuristics

1. Introduction

Categorization—grouping objects into classes and identifying class membership—is a fundamental cognitive ability. From a computational perspective, category learning poses formidable challenges, yet humans excel at it. Cognitive science has investigated how humans

Correspondence should be sent to Jana B Jarecki, Department of Psychology, University of Basel, Missionsstrasse 62A, CH-4055 Basel, Switzerland. E-mails: jj@janajarecki.com; jana.jarecki@unibas.ch

*Parts of this research were presented at the 2013 Annual Meeting of the Cognitive Science Society.

induce category structures, which representations they acquire, and which models account for learning and generalization. A number of studies have asked whether people transition between different strategies during categorization learning (Bourne, Healy, Kole, & Graham, 2006; Briscoe & Feldman, 2011; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Johansen & Palmeri, 2002; Medin & Smith, 1981; Smith & Minda, 1998). Although there is controversy about which model best describes early learning (e.g., Bourne et al., 2006; Johansen & Palmeri, 2002; Smith & Minda, 1998), one insight from this work is that people start with *simple strategies* before progressing to more computationally intense strategies. Johansen and Palmeri (2002) found that people initially apply unidimensional categorization rules (i.e., make classification decisions based on a single feature) and adopt more complex rules (i.e., a similarity-based strategy using multiple features) only if necessary. Smith and Minda (1998) observed that early in learning people tended to use a simple prototype-based strategy and only later shifted to a computationally more demanding exemplar-based strategy (but see Nosofsky & Zaki, 2002, for an alternative interpretation). These findings suggest that people transition from computationally simple to more complex strategies.

One benefit of starting simple is computational efficiency. As long as the initial simple rule gives good enough performance, a computationally more intense strategy need not be invoked. The performance of strategies also depends on the environmental structure (e.g., linearly vs. nonlinearly separable environments; Blair & Homa, 2001; Medin & Schwanenflugel, 1981). The environmental structure, however, is unknown at the beginning of learning. One important question for a simple strategy is whether it performs robustly across different environments. How can a strategy appear simple, and yet also be robust, and perform well in situations with unexpected environmental structures? Robust performance is especially important if the potential task structures are numerous and complex.

1.1. A computational perspective on classification

At the computational level, classification implies several challenges. One such challenge consists in the *curse of dimensionality* (Bellmann, 1961): As the number of features and categories in the environment grows, the number of feature–category combinations increases exponentially. This holds even in the simple case of just two categories and binary features: the number of possible feature-category combinations grows from $2^3 = 8$ with two features to $2^5 = 32$ with four features to $2^9 = 512$ with eight features. The curse of dimensionality is particularly important in real-world situations where the number of features that could be considered is large. Yet human performance, for example, in classification of images consisting of many continuously valued features, has outperformed computer algorithms (Russakovsky et al., 2015) until recently (He, Zhang, Ren, & Sun, 2015).

Another challenge, especially early in learning, is to make inferences from limited data. In many real-world situations, people will not have observed all possible feature–category combinations, necessitating inferences about novel instances from limited information. Yet humans seem able to readily classify even previously unseen objects, such as galaxies (Lintott et al., 2008).

1.2. Scope and goals

The present study investigates one way to address the computational challenge of probabilistic classification: the statistical principle of *class-conditional independence*. Machine learning research shows that this principle is simple and robust (Domingos & Pazzani, 1997). We introduce a probabilistic category learning model, the *dependence/independence structure and category-learning model* (DISC-LM), that can incorporate any level of prior belief in class-conditional independence. We designed two experiments in which the presumption of class-conditional independence would lead to specific error patterns. Results show that a model with high prior beliefs in class-conditional independence describes human data best.

Our work complements prior research in a number of ways. We build on recent demonstrations of interindividual differences in category learning (Bartlema, Lee, Wetzels, & Vanpaemel, 2014; McDaniel, Cahill, Robbins, & Wiener, 2014), by using a learning paradigm that terminates based on individual performance (like e.g., Homa, Dunbar, & Nohre, 1991; Medin & Smith, 1981), and model individual subjects' learning and beliefs. One limitation of previous studies that argued for shifts from simple to complex classification strategies is a focus on large, discrete bins of learning trials (56 in Smith & Minda, 1998; 36 in Johansen & Palmeri, 2002), or on specific test trials interspersed during learning (Erickson & Kruschke, 1998; Nosofsky, Kruschke, & Mckinley, 1992; Nosofsky, Palmeri, & McKinley, 1994; Smith & Minda, 2002). By contrast, our design enabled us to model each trial, and to present all exemplars, throughout the full learning duration. Using optimal experimental design principles (Myung & Pitt, 2009; Nelson, 2005) enabled us to find a task such that learners who presume class-conditional independence will make strongly different classification decisions from learners who do not presume class-conditional independence.

The purpose of this paper was to investigate the descriptive validity of the statistical principle of class-conditional independence in human classification learning. Our goal was to investigate whether human category learning is guided by general default assumptions about the structure of probabilistic environments, rather than testing a particular classification model.

1.3. Class-conditional independence

Classification, from a probabilistic modeling perspective, requires estimating the probability that a stimulus s belongs to class c, which can be computed using Bayes' rule:

$$P(c|s) = \frac{P(s|c)\,P(c)}{P(s)} \tag{1}$$

where $P(s|c)$ is known as the *stimulus likelihood*, the likelihood of a feature configuration s given the class c; and $P(c)$ denotes the *class base rate*, the probability of the class. The denominator $P(s)$ is a normalizing constant given by $\sum_i P(s|c_i) \times P(c_i)$. Intuitively, the probability that a stimulus belongs to class c is directly proportional to the class frequency and how often the stimulus has co-occurred with this class.[1]

Class-conditional feature independence means statistical independence of the features given the true class (e.g., Domingos & Pazzani, 1997; Flach & Lachiche, 2004; Rish, Hellerstein, & Thathachar, 2001). In general, statistical independence entails that joint probabilities can be computed as the product of marginal probabilities. Class-conditional independence means that if the class is known, knowing one feature does not give additional ability to predict another feature. In other words, conditioning on the true class renders features independent, whereas they are unconditionally dependent. If class-conditional feature independence holds, the stimulus likelihoods $P(s|c)$ in Eq. 1 can be computed as

$$P(s|c) = \prod_{d=1}^{D} P(f_d|c) \tag{2}$$

where s denotes the stimulus (the feature configuration), c the class, f_d the d^{th} feature of this stimulus, and D the total number of features. Class-conditional independence also relates to the idea of channel separability in sensory perception (Movellan & McClelland, 2001).

1.4. Benefits of assuming class-conditional independence

One key advantage of assuming that features are independent given the true class consists in addressing the curse of dimensionality (the exponential growth of feature-category combinations with the number of features). Class-conditional independence strongly reduces the number of parameters a probabilistic model requires, compared to a model that allows for arbitrary feature dependencies (see Fig. 1). Probabilistic models need to estimate the stimulus likelihoods $P(s|c)$ and the class base rates $P(c)$ for Eq. 1. While the number of parameters for the class base rate is unaffected by the number of features, the number of stimulus likelihoods grows exponentially in the number of features.[2] The total number of parameters a probabilistic model requires for a binary classification with D binary features is $2^{D+1} - 1$, if the model allows for arbitrary feature dependencies.[3] Assuming class-conditional independence reduces the total number of necessary parameters to only $2D + 1$.

A second computational benefit of class-conditional independence is that it allows inferences about new feature configurations, beyond inferences based on the category base rates alone. Even if a feature configuration has not been observed yet, the individual features may have occurred. If two exemplars, "*ab*" and "*cd*," and the class of each exemplar, have been observed, class-conditional independence enables inferences about the unseen feature configuration "*ad*" by using the marginal feature likelihoods of feature "*a*" and "*d*" to compute the configural likelihood of "*ad*" via Eq. 2.

Another advantage of class-conditional independence is its robustness. Although class-conditional independence may not hold exactly in many real-world environments (Rish et al., 2001; Titterington et al., 1981), presuming class-conditional independence need not

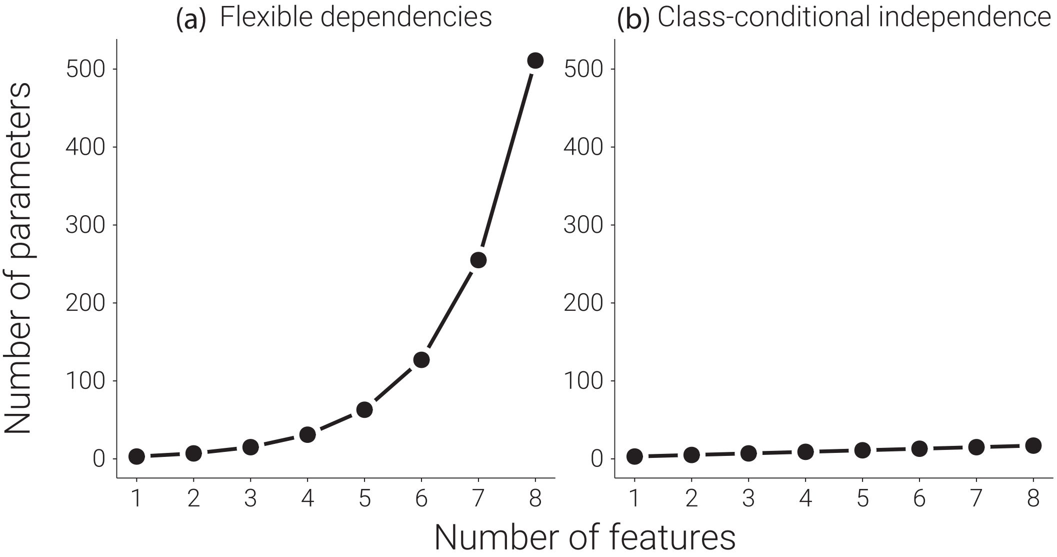

Fig. 1. Growth of the number of parameters in a binary categorization task as the number of features increases: (a) probabilistic model allowing for flexible feature dependencies; (b) probabilistic model relying on class-conditional feature independence.

impair classification performance. Both simulation studies and analytic results demonstrate the robustness of the naïve Bayes classifier, which treats features as class-conditionally independent. It often classifies accurately even if features are correlated given the true class (Domingos & Pazzani, 1997; Rish et al., 2001).

In sum, assuming that features are independent given the class has several computational benefits. From a psychological perspective, treating the environment as class-conditionally independent is a bet that the task structure complies with it. Acting accordingly can be useful due to its computational efficiency and robustness. However, dogmatically adhering to class-conditional independence in the face of substantial contradictory evidence would be a bad idea, because some category structures would be unlearnable. For instance, the naïve Bayes algorithm cannot learn the exclusive-OR problems that have been extensively studied in machine learning (Minsky & Papert, 1969), human categorization (Little & Lewandowsky, 2009; Love, Medin, & Gureckis, 2004), and causal learning (Waldmann & Martignon, 1998; Walsh & Sloman, 2008).[4] Humans can learn exclusive-OR problems (e.g., Little & Lewandowsky, 2009) and with sufficient experience can learn inferences based on configural stimuli rather than marginal features (e.g., Johansen & Palmeri, 2002; Little & Lewandowsky, 2009; Nosofsky & Bergert, 2007). Therefore, our hypothesis is not that learners always assume class-conditional independence, no matter what they have experienced. Rather, the idea is that learners use this principle as a default assumption in novel classification tasks. We introduce a new Bayesian model, the DISC-LM, that formalizes the idea of placing a particular level of prior belief in class-conditional independence, and illustrates what kind of evidence is needed to learn if that belief is incorrect in a particular task environment.

1.5. Class-conditional independence and other classification models

We next outline how class-conditional independence relates to single feature rules, decision bounds, prototype models, and fast-and-frugal trees, all of which also address the curse of dimensionality. Several of these psychological models are also hypothesized to describe human behavior early in category learning (Johansen & Palmeri, 2002; Smith & Minda, 1998), but none explicitly presumes class-conditional feature independence.

Single feature rules, which classify using one feature only, can sometimes describe human categorization behavior (e.g., Pothos & Close, 2008). Accordingly, people attend to just one feature. In terms of computational complexity, single feature rules are another way to greatly simplify the classification task, ignoring the curse of dimensionality altogether. While single feature rules need to attend to one feature and can ignore all remaining features, class-conditional independence attends to *all* features given the class but ignores any feature interactions given the class. The latter imposes an assumption about feature relations rather than attention restrictions.

Decision bound models are closely related to class-conditional independence. Decision bound models estimate a parametric boundary between categories. They often implement a linear bound without feature interactions, which is another way to address the curse of dimensionality (but other functional forms are possible). Linear feature separability is related to class-conditional feature independence, because in log space the naïve Bayes classifier is an interaction-free additive model (Manning, Raghavan, & Schutze, 2009; Zhang & Ling, 2001); that is, it induces a log-linear decision bound. For binary features, class-conditional independence implies linear separability of features, although this does not in general hold for features with more than two values (Zhang & Ling, 2001).

Additive prototype models compare the current stimulus to the most typical previous stimulus from each category (e.g., Posner & Keele, 1968; Reed, 1972) and select the class whose prototype is most similar to the current exemplar. As there is only one prototype per category, the number of comparisons between a novel stimulus and the prototypes increases linearly with categories instead of exponentially with features, thereby also addressing the curse of dimensionality. The classification of prototype models, however, differs from classification based on class-conditional independence, because the latter involves no similarity-based comparison to previous exemplars.

Fast-and-frugal trees classify according to a pruned decision tree, considering features sequentially one by one (Luan, Schooler, & Gigerenzer, 2011; Martignon, Katsikopoulos, & Woike, 2008). Popular fast-and-frugal tree construction methods (Martignon et al., 2008) are based solely on the marginal relationship of the individual features to the classes, and thus do not consider feature interactions. Depending on the tree structure and the stimulus being classified, fast-and-frugal trees may be able to make classification decisions based only on a subset of features, without considering all features of the stimulus.

1.6. Class-conditional independence in psychological models

Some probabilistic models explicitly presume class-conditional independence (e.g., Anderson, 1991; Barrington, Marks, Hsiao, & Cottrell, 2008; Friedman, Massaro, Kitzis, & Cohen, 1995; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011), but the idea that people begin learning with a particular feature-dependency assumption is at most indirectly addressed in these models.

Conditional independence assumptions have been investigated more directly in research on causal reasoning. This literature has focused on inference patterns entailed by the causal Markov condition (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), which is closely related to class-conditional independence.[5] Results indicate that reasoners often tend to violate the Markov condition (e.g., Rehder, 2014; Rehder & Hoffman, 2005; Rottman & Hastie, 2016), and also that they are sensitive to relevant contextual information (Mayrhofer & Waldmann, 2014) but considerations about the underlying causal mechanisms (Park & Sloman, 2013; Walsh & Sloman, 2008).

Although there has been considerable research on conditional independence assumptions in causal reasoning, little research has specifically investigated this issue in probabilistic categorization. To directly test whether class-conditional independence is a default presumption in probabilistic classification learning, we pit a model that initially explicitly assumes class-conditional independence against one without this prior assumption, and allow the model to learn the dependency structure from experience. Empirically, we use an experience-based learning paradigm (rather than conveying information numerically or verbally) to track how people's experience with a new environment shapes their classification behavior over the course of learning. Our methodology also enables us to study people's internal beliefs via their behavior, rather than via people's responses to specific verbal or numeric questions.

2. The dependence/independence structure and category-learning model (DISC-LM)

To formalize the assumption of class-conditional independence, we developed a probabilistic model that incorporates uncertainty about whether features are independent given the class, and uncertainty about the feature likelihoods as well as uncertainty about the class base rate. We designed this model to formalize the assumption of class-conditional independence in learning, rather than as a competitor to the existing more sophisticated classification models.

The DISC-LM is a hierarchical Bayesian model. It computes the probability that a stimulus belongs to class 1 according to class-conditional feature independence, and according to flexible conditional feature dependencies. The DISC-LM weights the obtained posterior probabilities according to the match between the data and the structural assumption about feature independence using Bayesian model averaging (Chickering & Heckerman, 1997). The resulting classification decision reflects both the uncertainty about whether features are class-conditionally independent and the uncertainty within each structural model about the true values of the class base rate and the stimulus likelihoods. We next describe the DISC-LM conceptually. The formal details are outlined in the supplementary material A.

2.1. Model parameters

The DISC-LM has two parameters. The first parameter is the *structural belief* parameter π with $0 \leq \pi \leq 1$. It governs the model's prior belief that the environment complies with class-conditional independence. Higher values of π lead to inferences more similar to inferring the class assuming class-conditional feature independence. Different values of this parameter can be used to model individual differences in believing that features are class-conditionally independent. The second parameter is the *conservatism* parameter δ with $\delta \geq 1$. It governs how much experience the model requires to learn the probabilities needed for computing the class probabilities. Higher values of δ lead to more conservative learning. The conservatism parameter enables the DISC-LM to account for individual differences in learning speed.

2.2. The model

The DISC-LM is a hierarchical Bayesian model with two levels (for an introduction to Bayesian modeling see, e.g., Griffiths, Kemp, & Tenenbaum, 2008). At the lower level, it infers the class base rate and stimulus likelihoods.[6] The class base rate $P(c_1)$ is inferred by updating a Beta distribution with a symmetric prior with hyper parameters equal to δ, resulting in a uniform prior for $\delta = 1$ and a symmetric prior around 0.50 for $\delta > 1$. The stimulus likelihoods $P(s|c_1)$ and $P(s|c_2)$ are inferred twice. The first inference, which formalizes the ability to learn arbitrary feature dependencies, uses two Dirichlet distributions over the two times eight possible likelihoods, each with a symmetric prior with hyper parameters equal to δ. The second inference, which formalizes the assumption of class-conditional feature independence, estimates the marginal feature likelihoods $P(f_d|c_1)$ and $P(f_d|c_2)$ based on which the configural stimulus likelihoods are computed (Eq. 2). The marginal feature likelihoods are inferred by twice updating three independent beta distributions, each of which has a symmetric prior with hyper parameters equal to δ.

Given the inferred class base rate and the stimulus likelihoods, the model computes the class of stimulus s (Eq. 1) once based on the Dirichlet stimulus likelihood—which can capture any feature dependencies—and once based on the Beta feature likelihoods, assuming class-conditional independence. We denote these estimates as $\hat{P}(c|s;flex)$, and $\hat{P}(c|s;cci)$, respectively.

At the higher level, the DISC-LM infers the degree to which features are class-conditionally independent in the environment. Given the observed data, the DISC-LM computes a posterior structural belief in class-conditional independence, $\hat{\pi}$. This posterior structural belief equals the normalized likelihood of the observed data under the assumption of class-conditional independence, weighted by a prior probability of class-conditional independence equal to π. Depending on whether the environment obeys class conditional independence or not, the posterior structural belief $\hat{\pi}$ shifts toward 1 or 0, over the course of learning.

If the prior structural belief $\pi = 0$, the model classifies only based on allowing for flexible class-conditional feature dependencies, whereas if $\pi = 1$, the model classifies only based on class-conditional feature independence. For both $\pi = 0$ and $\pi = 1$, the posterior structural belief $\hat{\pi}$ equals the prior structural belief π throughout learning. For prior beliefs of $0 < \pi < 1$, the model classifies based on a mixture of the estimates with and without class-conditional independence.

The DISC-LM then predicts the probability that the next stimulus *s* belongs to class *c* as

$$\hat{P}(c|s) = \hat{\pi}\hat{P}(c|s; cci) + (1 - \hat{\pi})\hat{P}(c|s; flex) \tag{3}$$

where *flex* and *cci* denote whether the current estimates were generated assuming flexible feature dependencies or class-conditional feature independence, respectively; and $\hat{\pi}$ is the posterior structural belief in class-conditional independence.

A key feature of the DISC-LM is that for a high prior belief in class-conditional independence it behaves as if this property holds in the environment; however, enough learning experience can override an erroneous prior belief for prior belief values below 1.

2.3. Relation to other mixture models of classification

We briefly consider how the DISC-LM relates to three other mixture models. The *prototype-exemplar mixture model* by Medin, Altom, and Murphy (1984) formalizes classification as a mixture between a multiplicative prototype model (Reed, 1972) and an exemplar model (Medin & Schaffer, 1978) with a mixing proportion *e*. It differs from the DISC-LM with respect to the classification models it combines. Furthermore, the mixing proportion in the prototype-exemplar model is constant, whereas the DISC-LM updates its mixing parameter (the level of belief in class-conditional independence) dynamically.

Gaussian mixture models, which have been used as a general framework for classification (Rosseel, 2002), represent the probability density of the stimuli by a sum of multivariate Gaussian distributions. Each distribution is defined by a mean feature value and a feature covariance matrix. The features can, but need not be, independent; they are independent only if the covariance matrix is diagonal (Monti & Cooper, 1999). Furthermore, in such mixture models, independence is conditioned on a hidden mixture component, rather than (as in the case of class conditional independence) on the true class. For class-conditional independence, the probability density of the stimulus likelihood is a product of marginal (univariate) densities. If the marginal feature likelihood densities are Gaussian, the DISC-LM density can be represented by a Gaussian mixture model with a diagonal covariance matrix.

The *hierarchical Dirichlet process (HDP)* model by Griffiths, Canini, Sanborn, and Navarro (2007) learns to group exemplars into a number of unknown clusters that need not correspond to the actual categories; this enables the model to form prototypes across categories (a similar idea is implemented by Vanpaemel & Storms, 2008). Its objective is to learn how many clusters are needed. The HDP model differs from the DISC-LM in two ways: While the HDP model infers the clusters, the DISC-LM conditions on the true categories without inferring hidden clusters. Furthermore, while the HDP model assumes that features are conditionally independent given the current clusters, the DISC-LM learns dynamically whether features are independent given the true class.

2.4. Summary

A new Bayesian learning model, the DISC-LM, learns whether the environmental structure corresponds to class-conditional independence and the necessary parameters for making classification decisions. If the prior structural belief parameter $\pi = 1$, the model incorporates the assumption of class-conditional independence. Setting the structural belief parameter to $\pi = 0$ makes feature dependencies completely flexible. The DISC-LM with a structural belief parameter $0 < \pi < 1$ weights the two posterior class probabilities by the fit between the observed data and the posterior belief $\hat{\pi}$ about whether class-conditional independence holds.

3. Design: Statistical task environment

We designed a classification task to strongly differentiate the behavior of learners who do and do not assume class-conditional independence, respectively. Our task had three binary features and a binary class.

3.1. Optimal experimental design

We searched for a statistical task structure in which class-conditional independence fails, despite its usual robust performance across different task structures. We used optimal experimental design principles (Myung & Pitt, 2009; Nelson, 2005), that is, computationally optimizing the task's parameters to discriminate between models. A genetic numeric optimization algorithm with hill-climbing was employed to search the space of possible task parameters for a solution that maximized the disagreement between classifications based on class-conditional independence and the true environmental structure. Supplementary material B describes the procedure.

3.2. Environment 1: Deterministic task

The resulting optimized task environment contains five stimuli, denoted as 000, 001, 010, 100, and 111; three of the eight possible stimuli do not occur (110, 101, 011). This was a result of the optimization, not a deliberate choice on our part. Fig. 2a illustrates the true classification task. Assuming class-conditional independence would lead to the erroneous classifications shown in Fig. 2b. This task structure allows us to test whether people assume class-conditional independence, because it implies strongly divergent

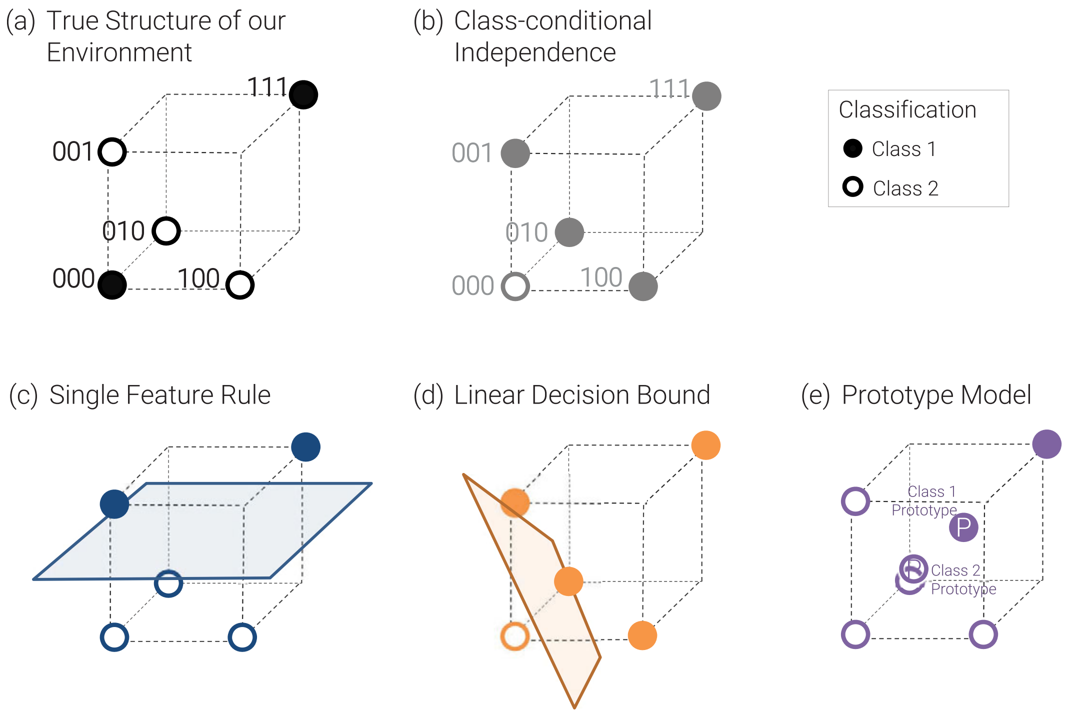

Fig. 2. True task structure compared to the classification using class-conditional independence and example classifications by various classification models.

classification decisions depending on whether class-conditional independence is assumed or not.

Table 1 summarizes the statistics of Environment 1. It shows the stimuli, their frequencies, the true probability with which they belong to class 1, and the class probability derived assuming class-conditional independence. For four of the five stimuli, the classification decision based on class-conditional independence conflicts with the actual class membership (indicated by $\neq$). We refer to those items as *critical stimuli*. Fig. 3 summarizes all parameters that describe the task.

The environment entails that a probabilistic model that assumes class-conditional independence selects a different class from that chosen by a probabilistic model that knows the true stimulus likelihoods. Both methods select the same class for only one of the five stimuli: stimulus 111 belongs to class 1 with probability 1 in the actual environment; presuming class conditional independence, it belongs to class 1 with probability .91. Thus, both models select class 1 given stimulus 111.

This equivalence in model decisions does not hold for the four critical stimuli. In the actual environment, stimulus 000 belongs to class 1 with probability 1, but under class-conditional independence it would belong to class 2 with probability .67. Similarly, the other critical stimuli (100, 010, 001) actually belong to class 2, but a learner assuming

Environment 1

	Flexible dependencies	Class-conditional independence
Class base rates	***Configural stimulus likelihoods***	***Marginal feature likelihoods***
$P(c_1) = .67$	$P(000\|c_1) = .42$ $P(001\|c_1) = 0$ $P(010\|c_1) = 0$ $P(011\|c_1) = 0$ $P(100\|c_1) = 0$ $P(101\|c_1) = 0$ $P(110\|c_1) = 0$ $P(111\|c_1) = .58$	$P(1\_\_\|c_1) = .58$ $P(\_1\_\|c_1) = .58$ $P(\_\_1\|c_1) = .58$
$P(c_2)$ $= 1 - P(c_1)$ $= .33$	$P(000\|c_2) = 0$ $P(001\|c_2) = .33$ $P(010\|c_2) = .33$ $P(011\|c_2) = 0$ $P(100\|c_2) = .33$ $P(101\|c_2) = 0$ $P(110\|c_2) = 0$ $P(111\|c_2) = 0$	$P(1\_\_\|c_2) = .33$ $P(\_1\_\|c_2) = .33$ $P(\_\_1\|c_2) = .33$

Environment 2

	Flexible dependencies	Class-conditional independence
Class base rates	***Configural stimulus likelihoods***	***Marginal feature likelihoods***
$P(c_1) = .71$	$P(000\|c_1) = .38$ $P(001\|c_1) = .04$ $P(010\|c_1) = .04$ $P(011\|c_1) = 0$ $P(100\|c_1) = .04$ $P(101\|c_1) = 0$ $P(110\|c_1) = 0$ $P(111\|c_1) = .50$	$P(1\_\_\|c_1) = .54$ $P(\_1\_\|c_1) = .54$ $P(\_\_1\|c_1) = .54$
$P(c_2)$ $= 1 - P(c_1)$ $= .29$	$P(000\|c_2) = .06$ $P(001\|c_2) = .29$ $P(010\|c_2) = .29$ $P(011\|c_2) = 0$ $P(100\|c_2) = .29$ $P(101\|c_2) = 0$ $P(110\|c_2) = 0$ $P(111\|c_2) = .07$	$P(1\_\_\|c_2) = .36$ $P(\_1\_\|c_2) = .36$ $P(\_\_1\|c_2) = .36$

Fig. 3. Assuming class-conditional feature independence reduces the number of class-conditional stimulus probabilities that are needed to describe the environment. The figure shows two ways to describe the classification task, with and without the assumption of class-conditional independence. The *class base rate* is required in either case. The flexible dependencies column shows that eight likelihoods (or class-conditional stimulus probabilities) describe the environments. The class-conditional independence column shows that only three marginal likelihoods are required under the assumption of class-conditional independence. The environment in the left panel is deterministic, the one in the right panel probabilistic.

Note. $P(c_1)$ = class 1 base rate, $P(000|c_1)$ = probability of configural stimulus 000 given class 1, $P(1\_\_|c_1)$ = probability of first marginal stimulus dimension given class 1.

class-conditional independence would assign them to class 1, because $P(c_1|s; cci) = .58$. The model disagreement is strongest for stimulus 000, with $P(c_1|000; trueenv) = 1.00$ but $P(c_1|000; cci) = .33$. Note that the stimuli are not equally frequent. The uncritical stimulus 111 is most frequent.

A model treating features as class-conditionally independent will perform poorly in this environment, irrespective of the amount of learning data. The poor performance follows from falsely assuming that features are class-conditionally independent, a structural assumption embedded in the inference mechanism. A learner who fully believes in class-conditional independence at the outset of learning cannot learn the true structure of this environment and would keep this incorrect structural belief, even after infinite experience.

3.2.1. Performance of single feature, decision bound, prototype, and fast and frugal tree models in our task

Let us return to the four simple psychological classification models outlined in the introduction and compare their class predictions in our task environment to the predictions made by assuming class-conditional independence.

Table 1
Environment 1 (deterministic task)

Stimulus s	True Frequencies	$P(c_1\|s)$		
		Flexible Dependencies		Class-Conditional Independence
1 1 1	.39	1	$\approx$	.91
1 0 0	.11	0	$\neq$	.58
0 1 0	.11	0	$\neq$	.58
0 0 1	.11	0	$\neq$	.58
0 0 0	.28	1	$\neq$	.33

Notes. $P(s)$: occurrence probability of stimulus s; *Flexible dependencies*: true class probability assuming class-conditional feature dependencies; *Class-conditional independence*: class probabilities derived assuming class-conditional independence; $\neq$: class-conditional independence yields different class decisions than the true environment.

A classifier that uses only a single feature dimension cannot learn the task. One such classifier is shown in Fig. 2c. It is easy to see that no two-dimensional plane which is parallel to the axis separates the feature combinations into the true classes. A linear classification rule without interaction terms (Ashby & Townsend, 1986) fails as well. One such example is shown in Fig. 2d; it is easy to see that no single two-dimensional plane sorts the feature combinations into the true classes. Nor can an additive prototype model with mean feature values used for prototypes (e.g., Reed, 1972) learn the task (Fig. 2e). Additive prototype models compute a representative (average) member of each class and compare how far away each feature combination is from this prototype. Fast and Frugal trees reach between 39% and 61% classification accuracy if constructed by the max(val+,val−) and zigzag(val+,val−) tree construction methods (Martignon, Vitouch, Takezawa, & Forster, 2003), respectively, while in our task 100% accuracy is achievable.

Considering single feature, decision bound, and prototype classifiers, we see that only the linear decision bound arrives at the same classifications as the class-conditional independence assumption (Fig. 2d). Single feature rules (Fig. 2c) and the additive prototype model (Fig. 2e) yield classifications that differ from the ones induced by assuming class-conditional independence. Importantly, none of these classifiers can learn the task environment.

3.3. Environment 2: Probabilistic task

An interesting property of the optimized task environment is the deterministic class membership: All stimuli belong to a class with $P = 1$ or $P = 0$. Our optimization did not explicitly aim for this, but—from a mathematical perspective—deterministic class membership best differentiates whether a classifier presumes class-conditional independence given three binary features and a binary class. However, we do not want to limit our analyses and empirical results to deterministic environments, which we suspect are fairly

Table 2
Environment 2 (probabilistic task)

Stimulus s	True Frequencies	$P(c_1 \mid s)$		
		Flexible Dependencies		Class-Conditional Independence
1 1 1	.38	.95	$\approx$	.89
1 0 0	.11	.25	$\neq$	.65
0 1 0	.11	.25	$\neq$	.65
0 0 1	.11	.25	$\neq$	.65
0 0 0	.29	.94	$\neq$	.48

Notes. $P(s)$: occurrence probability of stimulus s; *Flexible dependencies*: true class probability assuming class-conditional feature dependencies; *Class-conditional independence*: class probabilities derived assuming class-conditional independence; $\neq$: class-conditional independence yields different class decisions than the true environment.

rare, especially in situations with limited knowledge. Furthermore, research comparing learning in deterministic and probabilistic tasks (e.g., Little & Lewandowsky, 2009; Mehta & Williams, 2002) found that participants needed longer to learn probabilistic category structures (but see Seger & Cincotta, 2005). We therefore designed a second, probabilistic environment.

We manually changed the parameters of Environment 1 to design a probabilistic analog of it (Table 2). Environment 2 includes the same five stimuli, occurring with frequencies almost identical to Environment 1. The class probabilities, however, are no longer certain. For instance, in Environment 2, stimulus 010 belongs to class 2 with probability .75, rather than probability 1 in Environment 1. Importantly, Environment 2 preserves the same critical and uncritical stimuli.

Fig. 3 displays a comparison of the parameters required to specify our environments depending on whether class-conditional independence is assumed or not. The comparison shows that the number of parameters is smaller when assuming class-conditional independence: Only three marginal feature likelihoods for each class are required; in total, seven probabilities need to be estimated from data. Without the assumption of class-conditional independence, seven configural stimulus likelihoods for each class need to be estimated (the eighth stimulus likelihood is implied because the likelihoods given one class sum to 1); in total, 15 probability estimates are required.

In our experiments and simulations, we embedded the statistical task environments in a trial-by-trial category learning task. Models and human learners were presented with one stimulus, randomly drawn from the task distribution, and they received feedback about the true class after their classification decision.

4. Simulation studies

Our simulation studies investigated how the DICL-LM's prior structural belief in class-conditional independence, π, influences model behavior in our two learning environments, over the time course of learning. We simulated the DISC-LM for $N = 200$ learners with 50 learning trials each for both the deterministic and the probabilistic task.

4.1. Method and parameters

We simulated the posterior probability of the class for each trial and the posterior structural belief according to the DISC-LM, using Monte Carlo simulation. We binarized the posterior point estimates by an arg max response rule:

$$\text{class choice} = \begin{cases} \text{class 1} & \text{if } \hat{P}(c_1|s) > .5 \\ \text{class 2} & \text{if } \hat{P}(c_1|s) < .5 \\ \text{random} & \text{otherwise} \end{cases} \tag{4}$$

where c_1 is class 1 and s the stimulus.[7]

The DISC-LM was simulated with the conservatism parameter δ fixed at 1, and with different values of the prior belief in class-conditional independence $\pi \in \{0, .90, .99, .999999, 1\}$. Appendix A shows simulations for different values of δ. The uneven grid for π resulted from the fact that DISC-LM learners with values of $\pi < .7$ converged to behavior indistinguishable from learners with $\pi = 0$ very quickly.

4.2. Results

We investigated how the prior belief in class-conditional independence influences learning behavior (how fast can the model with different prior beliefs in class-conditional independence learn?). We also investigated the development of the structural beliefs (how quickly does the model learn that class-conditional independence is violated?). Fig. 4a shows the results for Environment 1, and Fig. 4b shows the results for Environment 2.

4.2.1. Learning curves

The $\pi = 0$ DISC-LM without structural belief in class-conditional independence quickly learns to correctly classify all stimuli, in both environments. The maximum performance the model can achieve is lower in the probabilistic (Fig. 4b) than in the deterministic environment (Fig. 4a). The variations in learning speed of the $\pi = 0$ model learner across the different stimuli reflect the unequal frequencies of the stimuli (see Tables 1 and 2). The model learns the more frequent stimuli 111 and 000 fastest, compared to 001, 010, and 100.

The $\pi = 1$ DISC-LM with persistent structural belief in class-conditional feature independence learns to correctly classify the uncritical stimulus 111 quickly in both environments, but it fails for the critical stimuli 100, 010, 001, and 000. Even with infinite experience this model will—in these particular environments—never learn, because the model's persisting structural belief prevents learning the true feature dependencies.

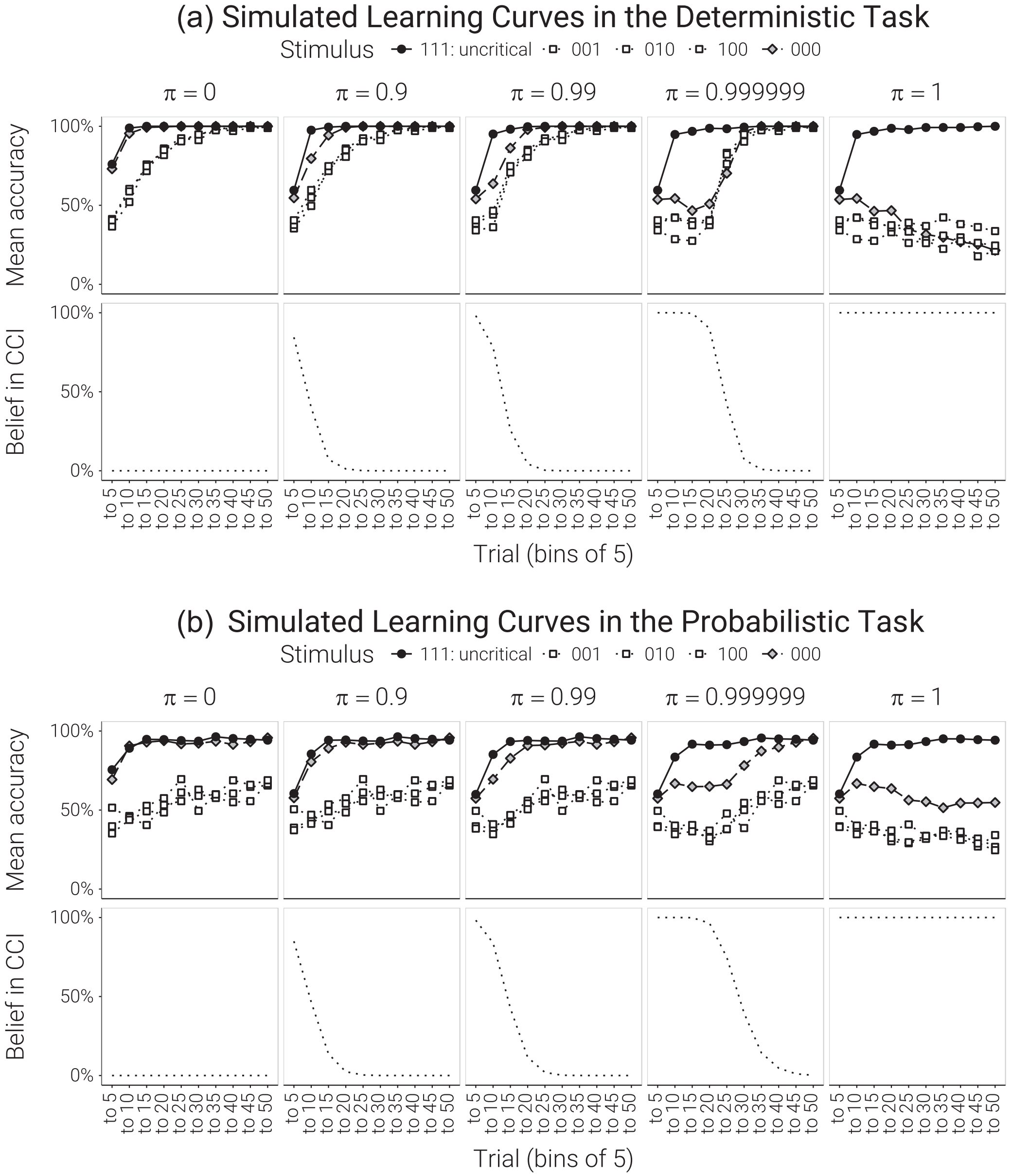

Fig. 4. (a) Deterministic environment; (b) Probabilistic environment. *Mean accuracy:* How quickly the DISC-LM learns to classify the stimuli in Environment 1 depends on the prior beliefs in class-conditional independence (CCI), π. Stronger prior beliefs in class-conditional independence result in slower learning, but only for the critical stimuli. Also note that the leftmost model (π = 0) performs above chance for stimuli 000 and 111 in the first bin. This is because the DISC-LM with π = 0 infers the class of stimuli 000 and 111 correctly from the class base rate within five trials. *Belief in CCI:* The belief in CCI decreases with experience in the environment, for prior belief values of 0 < π < 1 (higher values represent stronger beliefs).
Note. The *x*-axis shows the trials in bins of five while keeping the presentation order of stimuli.

Table 3
Predictions for the stimuli in our task, based on the simulation study

Simulation Result	Description
Superiority of 111	Stimulus 111 is learned most quickly, largely independent of the prior belief in class-conditional independence π
Initial slowing of 000	Learning of stimulus 000 is slower in the first trials the stronger the prior belief in class-conditional independence
Slowing of 001, 010, 100	Learning to classify stimuli is slowed down uniformly with stronger prior beliefs in class-conditional independence
Similarity of 000 and 111	The model with $\pi = 0$ predicts that stimuli 000 and 111 are learned almost equally fast

The DISC-LM with high, but nondeterministic, prior structural belief values ($\pi = .9, .99, .999999$) learns the uncritical stimulus as quickly as the $\pi = 0$ DISC-LM. However, high values of π result in *slower* learning of the critical stimuli. The stronger the prior belief in class-conditional independence, the slower the learning.

4.2.2. *Learning the feature dependency structure*

The higher the prior structural belief, the slower the DISC-LM learns that features are not class-conditionally independent.

4.3. *Summary*

With strong prior structural belief in class-conditional independence, learning of the DISC-LM is influenced asymmetrically across feature combinations, in both environments. With stronger prior structural beliefs, the critical stimuli 000, 100, 010, and 001 are learned more slowly, but learning of stimulus 111 is not impaired. This is true in the deterministic and probabilistic environments alike. These simulation results are the basis of our predictions for our experiment with human subjects, listed in Table 3.

5. Experiment 1—Deterministic Task

Experiment 1 was designed to investigate the extent to which humans treat features as class-conditionally independent early in learning. Our experiments used a supervised trial-by-trial learning paradigm (e.g., Ashby & Maddox, 1992) adapted from previous studies (e.g., Meder & Nelson, 2012; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Experiment 1 was based on the deterministic task environment shown in Table 1.

5.1. *Participants*

Thirty people (M_{age} 23.8 years, range 19–33 years, 67% female) participated; remuneration was 12 euros. We recruited via the Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development in Berlin, Germany. Data were collected from September to December 2012 at the center; the experiment was conducted in accordance with the ethical and data protection guidelines there.

5.2. Materials and procedure

Participants classified “plankton” stimuli differing in eye, claw, and tail appearance (binary features) into species A and species B (class). Fig. 5 illustrates the material. Each plankton specimen corresponded to one feature configuration in Table 1, but the assignment of physical features and class labels was randomized across participants.

In the beginning, participants were familiarized with the feature locations. In each trial they classified a plankton specimen drawn from the probability distribution in Table 1 and received feedback about the true class (letters “A” or “B”) and a smile emoticon after a correct decision or a frown emoticon otherwise. Learning was self-paced. Participants were instructed to always choose the most likely class. The presentation of stimuli ended when participants reached a learning criterion defined as both (a) having made at most four classification errors over the last 200 trials (98% of 200 correct), and (b) having chosen the most likely category for the last five times that each individual stimulus appeared within the random sequence of stimuli.

After 15 learning trials, participants saw “frequently asked questions,” which, among other things, reminded them to always pick the most likely class and informed them that

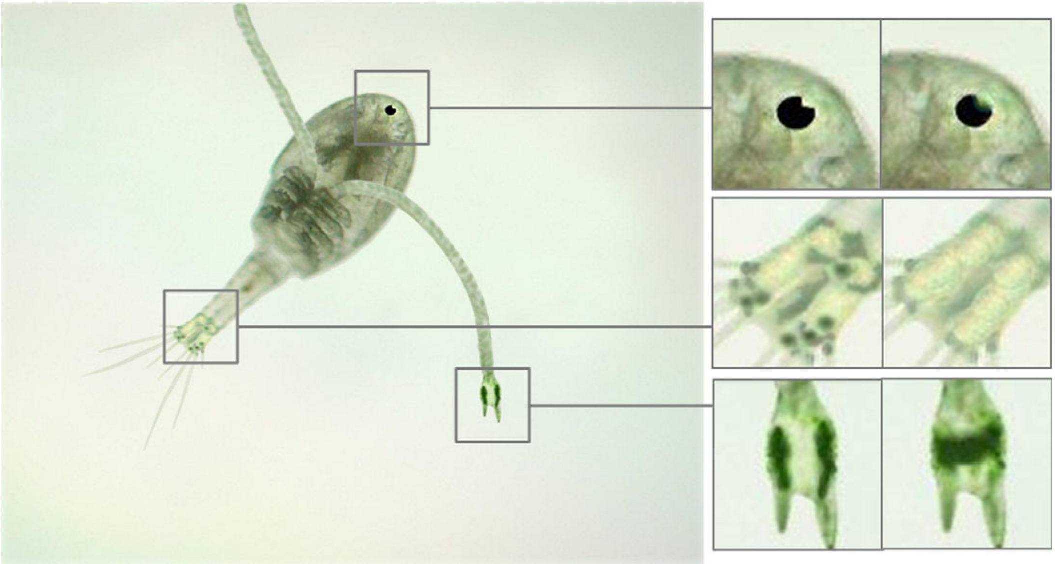

Fig. 5. Sample stimulus used in Experiments 1 and 2 (from Nelson, 2010). In each trial, participants saw and classified one plankton specimen (left) on the basis of three binary features. The gray boxes and magnification of the three features (right) are for illustrative purposes only.

it usually takes 300–400 trials to reach criterion performance. At regular intervals (every 100 trials, from trial 200 onward), participants were informed about their current performance and the maximum possible accuracy in the task. They were reminded to attend to all features rather than to focus on only one feature, but no information regarding how to integrate features was given (see Appendix B).

5.3. Behavioral results

All participants reached the learning criterion, in 200 to 798 trials (median = 338, *M* = 381, *SD* = 155).

5.3.1. Classification errors

The environmental probabilities were designed so that the presumption of class-conditional independence would lead to incorrect classification of the critical stimuli (000, 001, 010, 100). Accordingly, we expected more classification errors for the critical stimuli than for the uncritical stimulus, overall. We derived error rates separately for each stimulus, because stimuli were not equally frequent (see Table 1). As Fig. 6a shows, participants misclassified the critical stimuli more frequently than the uncritical stimulus, over the whole course of learning.[8] This finding is consistent with the idea that people treated features as class-conditionally independent and classified stimuli accordingly.

Aggregating errors over time and individuals, as in the above analysis, ignores interpersonal variability and temporal dynamics (e.g., Estes & Maddox, 2005). The temporal development of errors is key to our hypotheses. That all participants achieved criterion performance indicates that they did not treat features as class-conditionally independent throughout learning (otherwise they would have failed to reach criterion performance). Our hypothesis, formalized in the DISC-LM, is an initial assumption of class-conditional independence. This should slow down early learning, in particular. We next analyze individual learning dynamics.

5.3.2. Learning curves

The dependence-independence structure and category learning model simulations showed that a high prior belief in class-conditional independence impairs learning asymmetrically for the critical stimuli, but not for stimulus 111 (Fig. 4). All versions of the DISC-LM predicted a superiority of stimulus 111, which the human data in Fig. 6b confirm. Importantly, the human data show slower learning of 000 compared to 111. This was only predicted by the DISC-LM with high prior beliefs in class-conditional independence; it contradicts behavior of the DISC-LM without structural priors ($\pi = 0$, according to which 000 and 111 should be learned equally fast). Moreover, the DISC-LM with prior beliefs in class-conditional independence > .90 predicted that stimuli 001, 010, and 100 would suffer an initial phase of stagnation, before being learned. Participants' learning curves also show this pattern, supporting the idea that learners initially treat features as class-conditionally independent.

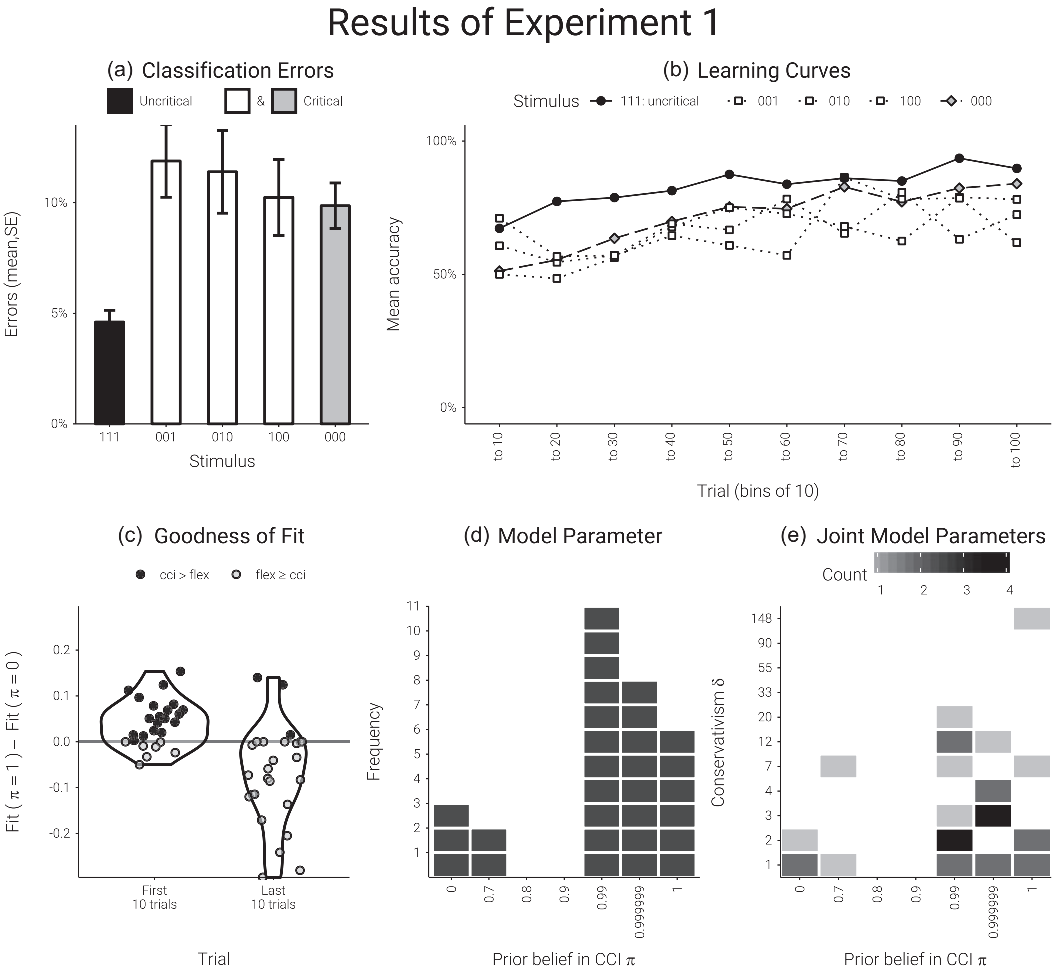

Fig. 6. Results of Experiment 1. (a) Mean error rates across all trials for each stimulus. Error bars = standard errors, bootstrapped with 100 replications. Error rates were lowest for the uncritical stimulus (111) and higher for the critical stimuli (000, 001, 010, 100). (b) Proportion of correct choices in the first 100 trials. Dots represent correct (i.e., most likely) classifications per stimulus, averaged within bins of width 10. The curves show that the uncritical stimulus (111) was learned faster than the critical stimuli. Among the latter, learning stimulus 000 was easiest, but still harder than 111. (c) In the first 10 trials, a model assuming class-conditional independence ($\pi = 1$) fits the data better than a model without the assumption ($\pi = 0$) for most participants; this reverses for the last 10 trials. (d) Distribution of the prior structural belief parameter π when predicting individual choices of participants ($N = 30$). A value of $\pi = 0$ means flexible conditional feature dependencies; a value of 1 means a strong fixed prior belief in class-conditional independence. Models with a high prior on class-conditional independence best account for the majority of participants in Experiment 1. *Notes*. Model fit was assessed by mean squared error (*MSE*, see main text). (e) Joint distribution of the obtained parameter values. Squares show which parameter combinations occurred; darker colors denote higher occurrence frequencies.

5.4. Modeling results

5.4.1. Data preprocessing

Because the DISC-LM assumes symmetric prior distributions, it predicts a mean class base rate in trial one of $p(c_1) \approx p(c_2) \approx .50$, except for small Monte Carlo errors. We tested if participants' first choices deviated from randomness and found no difference (19 of 30 class *a* choices in the first trial, exact binomial test for equal proportions: $p = .21$). To ensure that the Monte Carlo error did not distort the comparison of model and human behavior, we predicted participants' decisions only starting from trial 2. Furthermore, we used the observed decisions only up to $T - 200$, where T is a participant's last trial, because in the last 200 trials our learning criterion enforced 98% correct choices. This excluded one participant who needed exactly 200 trials, for whom we assumed no prior belief in class-conditional independence, that is, $\pi = 0$. This left 29 participants for the subsequent analyses.

5.4.2. Prediction generation

To investigate at a more fine-grained level whether class-conditional feature independence is a default assumption in human category learning, we were particularly interested in the parameter π of the DISC-LM (i.e., the prior belief in class-conditional feature independence). We applied individual parameter selection through one-trial-ahead prediction, using mean squared error (*MSE*) as the criterion for the quality of the prediction.[9] *MSE* was our measure of choice because it emerged as the best measure in a parameter recovery simulation, compared to mean absolute error and likelihood-based measures (see supplementary material C).

We modeled the decision of each learner in each trial with the DISC-LM. The classification probabilities were derived from Monte Carlo simulations using a grid of a priori fixed parameter values. This grid included $\pi \in \{0, .7, .8, .9, .99, .999999, 1\}$; and conservatism values of $\delta \in \{1, 2, 3, 4, 7, 12, 20, 33, 55, 90, 148\}$.[10]

5.4.3. Goodness of fit

If participants begin with the assumption of class-conditional independence, their early behavior but not their late behavior should be described better by a model incorporating class-conditional independence, compared to a model without such assumptions. We tested this using data from people's first 10 and last 10 choices, computing the difference in fit between a class-conditional independence DISC-LM ($\pi = 0$) and the flexible-dependency DISC-LM ($\pi = 1$), given individually adjusted δ parameters. Fig. 6c shows that the class-conditional-independence DISC-LM accounts better for the early data (it fits 19 of 26 participants better, positive mean fit difference of 4 percentage points, $t(25) = 4.02$, $p = .0005$), but it performs worse than the flexible independence DISC-LM for the late data (it fits only three participants better, negative mean fit difference of -7 percentage points, $t(25) = -3.40, p = .003$).[11]

5.4.4. Model accuracy

We obtained the parameter combinations for π and δ that jointly minimized the *MSE* between observed choices and model predictions. Given the resulting parameter combinations, the model's accuracy was 81.77%, averaged across the 29 participants.[12] By contrast, a

DISC-LM without independence assumptions that enforces flexible dependencies ($\pi = 0$ and best-fitting values of δ) is less accurate, reaching only 77.42% accuracy, with a mean difference between those models of .04 ($t(28) = 3.58$, $p = .0012$). Remember, one participant was not modeled because she hit the learning criterion in the minimum of 200 trials. To not ignore her data, we added this person (without fitting the model) as one *not* using class-conditional independence, with values of $\pi = 0$ and $\delta = 1$. This was because our learning criterion enforced good performance for the last 200 trials and by design, correct classification is only possible if any initial belief in class-conditional independence is given up.

5.4.5. Initial beliefs in class-conditional independence

We hypothesized that humans start classification learning with an initial belief that features are class-conditionally independent (high values of π). The data strongly bear out this expectation. Of our 30 participants, 25 were best accounted for by a model with prior belief in class-conditional independence of at least .99 (Fig. 6e shows results for both the class-conditional independence prior and the conservatism parameter), and two participants had moderately high values of $\pi = .7$. The joint distribution of prior belief and conservatism parameters in Fig. 6e shows that only some of the participants with higher class-conditional independence priors are more conservative. This suggests that class-conditional independence is assumed by the majority of participants early in classification learning. Only 3 of 30 participants were best accounted for by a model without prior belief in class-conditional independence (i.e., with $\pi = 0$).

5.4.6. Summary

Both the classification errors for the different stimuli on the aggregate level, and individual participants' behavior, are consistent with the idea that class-conditional independence serves as a default assumption in classification learning. The classification errors and different learning curves are in line with a model that assumes a strong initial belief in conditional independence. When fitting the π parameter of the DISC-LM to the learning data, for most participants a high value of π accounted for the data best. We next investigated a probabilistic task.

6. Experiment 2—Probabilistic Task

6.1. Participants

A total of 39 people participated. Ten had to be excluded (eight who did not reach the learning criterion in 120 min, and two due to a computer crash), leaving us with 29 participants (M_{age} 24.8 years, range 18–35 years; 79% female). They were paid 12 euros. Data were gathered from April to June 2013 at the same laboratory as in Experiment 1.

6.2. Materials and procedure

The materials and procedure were almost identical to Experiment 1, with the difference that the stimuli were drawn from the probabilistic task environment (Table 2). The correct (most likely) choices given the stimuli corresponded to Experiment 1, but the maximum achievable accuracy was 88% (instead of 100% as in Experiment 1). Learning ended when participants selected the most likely class 98 % out of the last 200 trials and had selected the most likely class for the last five appearances of each stimulus.

6.3. Behavioral results

The participants who reached the learning criterion needed between 212 to 1,156 trials (median = 627, M = 620, SD = 280) to achieve criterion performance, which is more than Experiment 1 ($t(43) = 4.03$, $p = .0003$, Cohen's $d = 1.06$, with Welch–Satterthwaite correction for variance inhomogeneity). The slower learning in the probabilistic compared to the deterministic environment corresponds to previous findings (Little & Lewandowsky, 2009; Mehta & Williams, 2002; Nosofsky & Stanton, 2005); with exceptions (Juslin, Olsson, & Olsson, 2003; Seger & Cincotta, 2005).

6.3.1. Classification errors

As in Experiment 1, we computed the proportion of errors separately for each stimulus, aggregating over time and participants (errors were defined as not choosing the most likely class). Again participants made more errors when classifying the critical stimuli, for which assuming class-conditional independence results in diverging class choices than when assuming flexible dependencies, compared to the uncritical stimulus (Fig. 7a).[13] The relatively small number of errors for the critical stimulus 000 can be explained considering Table 2, which shows that, in this environment, a classifier with class-conditional independence predicts the correct class of 000 with a rather high probability, $P(c_1|000; cci) = .48$. This means that a classifier assuming class-conditional independence is expected to select the least likely class in almost half (48 of 100) trials in this environment.

6.3.2. Learning curves

Fig. 7b shows the stimulus-wise learning curves, aggregated over participants. The pattern corroborates the results of Experiment 1: The easiest item was the uncritical stimulus 111; the critical stimulus 000 was more difficult, at least in the beginning. This pattern is predicted only by the DISC-LM learners with beliefs in class-conditional independence, and not by a model that a priori assumes flexible conditional feature dependencies ($\pi = 0$). Critical stimuli 001, 010, and 001 were the most difficult, consistent with strong beliefs in class-conditional independence. Again, the data are at variance with the pattern predicted by the DISC-LM with $\pi = 0$, according to which stimuli 000 and 111 should be learned equally quickly. These findings are in line with the results of Experiment 1, supporting the hypothesis that human learners initially treat features as class-conditionally independent.

6.4. Modeling results

We used all trials except the first trial and the final 200 trials to examine which values of the prior belief in class-conditional feature independence π best predicted participants'

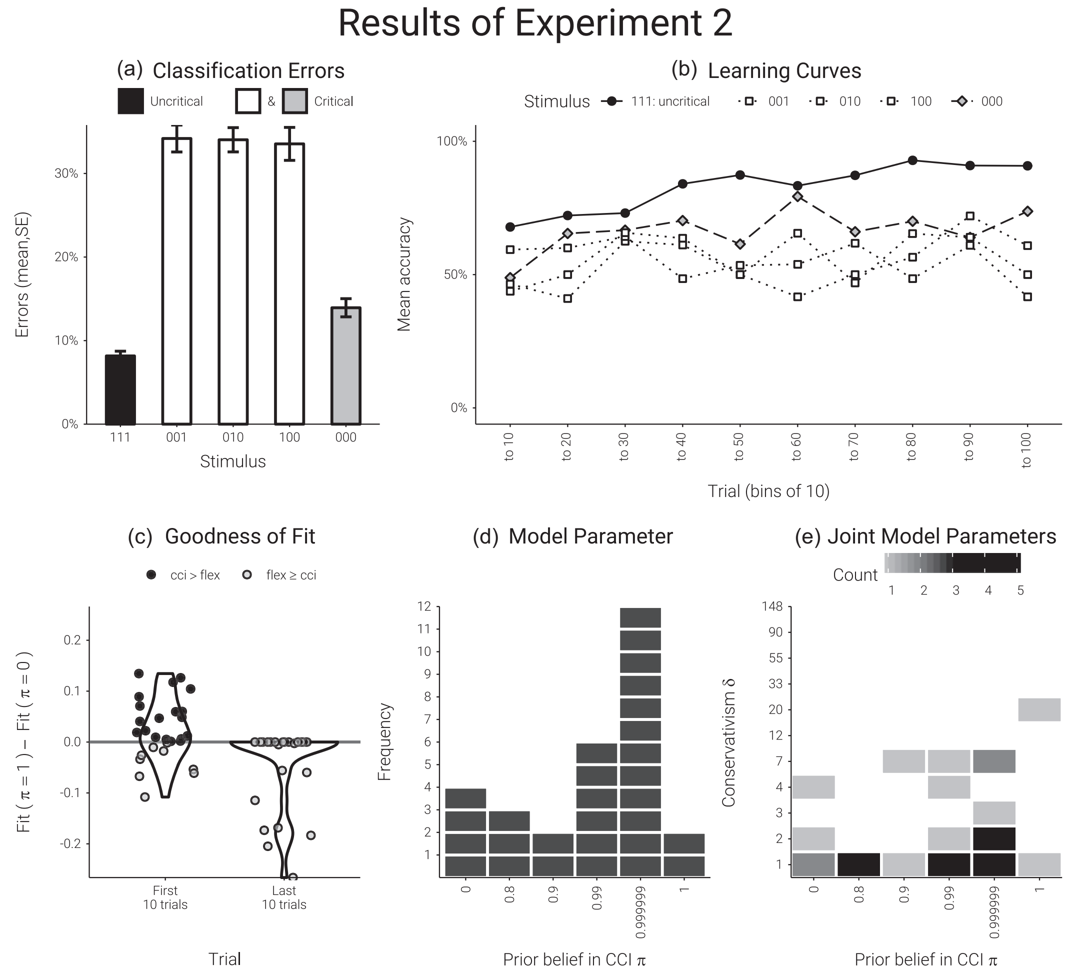

Fig. 7. Results of Experiment 2. (a) Classification errors. (b) Classification improvement in the first 100 trials. Dots represent most-likely-class choices per stimulus, averaged within bins of width 10. The curves show that participants improved fastest for the uncritical stimulus (111). For critical stimuli, improvement was slower. Among them, learning stimulus 000 was easiest, but still harder than 111. (c) In the first 10 trials, a model assuming class-conditional independence ($\pi = 1$) shows a higher fit than a model without the assumption ($\pi = 0$) for most participants; however, this reverses for the last 10 trials. (d) Distribution of best predicting values for the parameter π ($N = 29$). This parameter reflects the model's prior belief in a class-conditionally independent task structure. Values of $\pi = 0$ denote no belief; values of 1 denote the strongest structural belief. The model with high belief values predicts most participants best in Experiment 2.

Notes. Model fit was measured as mean squared error (*MSE*); we used trials $t = 2$ to $t = \max(t) - 200$ to obtain the values. (e) Joint distribution of the obtained parameter values. Squares show which parameter combinations occurred; darker colors denote higher occurrence frequencies.

decisions. We tested whether we excluded informative data by not using the first trial and found no evidence for this: 12 of 29 participants selected class 1 in the first trial, exact binomial test for equal proportions: $p = .46$. We derived predictions using

Monte Carlo simulations of the DISC-LM for different values of π and δ. The parameter grids consisted of $\pi \in \{0, .8, .9, .99, .999999, 1\}$ and $\delta \in \{1, 2, 3, 4, 7, 12, 20, 33, 55, 90, 148\}$. The grid was determined (as in Experiment 1) by selecting parameter values such that each prediction differed by more than 1 percentage point from predictions with $\pi = 0$, across trials 2–100.

6.4.1. *Goodness of fit*

Comparing how well the extreme versions of the model (the DISC-LM that has no structural priors at all and the one that always uses class-conditional independence) predict early and late learning behavior shows that the class-conditional-independence DISC-LM performs better for early choices but worse for later data. In the first 10 trials, the class-conditional independence model outperforms a flexible-dependency model for 19 of 28 participants; this holds for 0 participants in the last 10 trials. The differences in fit ($\text{fit}_{\pi=1} - \text{fit}_{\pi=0}$) are positive for the early and negative for the late trials.[14] Fig. 7c shows a trend favoring the class-conditional-independence model in the first 10 trials (positive small mean fit difference of 2 percentage points, $t(27) = 1.86$, $p = .07$); while in the last 10 trials, the class-conditional-independence DISC-LM performs worse (mean fit difference of -4 percentage points, $t(27) = -2.93$, $p = .006$).[15]

6.4.2. *Model accuracy*

We obtained the best parameter values by individual trial-by-trial predictive fitting as described for Experiment 1. There was one tie where two π values resulted in equal *MSE* scores. We conservatively selected the lower π value for a lower belief in class-conditional independence. The model's accuracy for the resulting parameter values, averaged across the 29 participants, was 84.12%. A model that does not believe in class-conditional independence ($\pi = 0$ and best-fitting δ values) has 82.83% accuracy, which yields a mean difference of 1 percentage point, ($t(29) = 1.44$, $p = .16$). Here, only the qualitative direction indicates that a $\pi = 0$ model describes behavior less well. There are two plausible reasons for the rather small accuracy difference in Experiment 2. First, in about 38 of 100 trials, when the uncritical stimulus 111 is drawn, the models make the same choice prediction irrespective of π (see Table 2). Recomputing the accuracy without stimulus 111 yields mean accuracy values of 79.34% and 77.58% (mean difference = 2 percentage points, $t(28) = 1.80$, $p = .09$) between the best-fitting-parameter and the flexible-dependency DISC-LM. The second reason is that after the structural belief parameter converges to $\hat{\pi} = 0$ (which happens quickly, after about 50 trials; see the lower panel in Fig. 4b), both models become indistinguishable. Experiment 2 involves many more late learning trials (median number of trials 627, vs. 338 in Experiment 1), increasing the difficulty of discriminating the models based on the average fit to all trials.

6.4.3. *Initial beliefs in class-conditional independence*

If learners are initially guided by high prior beliefs in class-conditional independence, this should be reflected in high parameter values of π. In line with this, we obtained values of $\pi \geq .9$ for the majority of participants (see Fig. 7d; Fig. 7e shows joint results for both parameters). As in Experiment 1, few participants (4 of 29) were best described by a fully flexible model with no beliefs in class-conditional independence. The behavior of most participants (20 of 29) was best accounted for by strong beliefs in class-conditional independence, with π values of .99 or higher. Similar to Experiment 1, the joint distribution of prior belief and conservatism parameter in Fig. 7e reveals slower learning for a subset of the participants with high structural priors (initial belief in class-conditional independence). Thus, Experiment 2's results support the hypothesis that class-conditional independence is a default assumption in human classification learning.

7. Comparison to single-dimensional classification

Initially presuming class-conditional independence is one way to cope with the combinatorial explosion in classification learning. But it is not the only way. A different rather simple model uses only one feature to classify, which is a strategy that would fail in our environment just like the class-conditional independence assumption.

To compare how well another simplifying assumption represents our data, we fitted a single feature attention model to participants' data and compared it to the fit of the class-conditional independence assumption embedded in the DISC-LM. The single feature model was formalized as restricted version of the generalized context model (GCM; Nosofsky, 1986). The model assumes that participants classify the current stimulus based on how similar the features of the current stimulus are to the previous stimuli. The current exemplar is classified into the class with most similar exemplars. The GCM has attention weight parameters allocating attention to particular feature dimensions (for formal details see Nosofsky, 1986). We used a GCM with city-block metric and exponential decay, with attention weights fixed to one dimension, and fitted the discriminability c as a free parameter by maximum likelihood. The attention weights can be restricted in three ways (attending to only the first, second, or third dimension). For each participant, we selected the single feature attention allocation that described the participant's data best.

The single feature model achieved a fit in terms of $1 - MSE$, averaged across participants, of 77.87% compared to 86.71% by the DISC-LM in Experiment 1 ($t(51) = -7.419$, $p < .001$). In Experiment 2, the single feature model achieved a fit of 79.54% vs. 88.78% by the DISC-LM ($t(49) = -10.835$, $p < .001$, both tests with Welch–Satterthwaite correction). The DISC-LM outperformed the single feature model for 28 of 29 participants in Experiment 1 and for 29 of 29 participants in Experiment 2.

8. General discussion

A variety of theoretical arguments and machine learning results suggest that the assumption of class-conditional independence of features could be very helpful in probabilistic category learning. Research to date with human subjects has not specifically focused on this question. We used computer simulations to identify statistical environments in which learners who presume class-conditional independence will make strongly different classification decisions than fully flexible learners.

With a new Bayesian learning model, the DISC-LM, we formalized varying degrees of belief in the class-conditional independence assumption in order to study the kinds of inferences that would be made over the course of learning, according to whether or how strongly one initially presumes class-conditional independence. Different versions of the DISC-LM, with different prior beliefs about environmental structure, showed very different patterns of learning trajectories, especially early in learning. Importantly, the DISC-LM learns, over time, whether or not class-conditional independence holds and adjusts its beliefs and classification decisions accordingly. Based on the model behavior, we derived a number of specific predictions for human learners' behavior, on the same tasks.

Experiment 1 consisted of a deterministic classification learning task designed to optimally discriminate between people who treat features as class-conditionally independent and those who do not. The task involved four critical stimuli, for which classifications derived assuming class-conditional feature independence disagreed with the class choices derived from the true feature dependencies, and one uncritical stimulus, for which the independence assumptions did not entail diverging classifications. Participants made more classification errors for the critical stimuli than for the uncritical stimulus. Participants also learned the critical stimuli more slowly compared to the uncritical stimulus. This stagnation in learning was predicted only by models with nonzero prior belief in class-conditional independence. We also modeled individual choices using the DISC-LM. Most participants' classification decisions were best predicted by versions of the model with very high prior beliefs in class-conditional feature independence.

Experiment 2 followed a similar rationale but used a probabilistic task to reflect that most real-world categorization environments are not deterministic (either inherently or due to incomplete knowledge). Results replicated the first experiment: Most participants' initial classification decisions were best accounted for by a DISC-LM with a high prior belief in class-conditional independence.

Results from all analyses, across both experiments, found that models that place extremely strong (but not 100%) initial belief in class-conditional independence best account for human behavior. Note that the version of the DISC-LM that does not correct its initial assumption of class-conditional independence (i.e., $\pi = 1$) did not capture participants' behavior; neither did a version of the model that allowed for completely flexible conditional feature interactions (i.e., $\pi = 0$) throughout learning. The model best capturing behavior was one that dynamically adapted to the environmental structure. Although class-conditional independence performs well across many environmental structures, people can learn from experience to overcome their structural prior beliefs when the learning input contradicts their assumptions.

We used environments with three binary features and two binary categories. The literature includes many tasks with two or three features (e.g., Meder & Nelson, 2012; Rehder & Burnett, 2005; Sanborn, Griffiths, & Navarro, 2010; Vigo, 2013; but see Nosofsky et al., 1994); thus, we had a priori reason to believe that such tasks would be learnable.

In environments with more than three features, the curse of dimensionality is stronger. Thus, making a simplifying initial assumption, such as class-conditional independence, would be even more important in more complex environments.

The late learning behavior, in which our participants had learned the category structure, can potentially be described by various models. It may be described with an exemplar-based strategy, which can learn exclusive-OR structures such as our task environments (Nosofsky, 1992). Late learning may alternatively be described with a rule-plus-exception strategy, for example, "classify as class 1 if all features = 1, otherwise classify as class 2, except if all features = 0." Early learning could be modeled by, for example, Anderson's (1991) rational model of categorization or the model by Sanborn, Griffiths, and Navarro (2006), starting with a single cluster fully implementing class-conditional independence. As noted before, our purpose was not to develop a new categorization model, but rather to test whether participants bring a specific assumption that is justified from a computational perspective (class-conditional independence) when learning novel categorization tasks.

It should be noted that the probabilistic DISC-LM is situated at Marr's computational level (Marr, 1982) and does not make claims about the underlying cognitive information-processing steps. We designed the model to test a specific hypothesis about people's behavior, rather than a cognitive process model. The insights from our studies of the DISC-LM and human learners are potentially relevant to researchers building various kinds of learning models. It should also be noted that, at the computational level, only the class-conditional independence principle itself addresses the curse of dimensionality; the computations underlying the late-learning behavior of the DISC-LM themselves are subject to the curse of dimensionality.

8.1. Issues for future empirical research and development of the DISC-LM

Our data suggest that in the kinds of tasks that we studied, people have strong prior assumptions of class-conditional independence. It is possible that people would bring different assumptions to other tasks. For instance, radially symmetric organisms (like starfish) might be presumed to have highly correlated individual arms. Suppose that the presence of a red spot on an individual arm favors class 1. If the arms are presumed to be highly correlated with each other (almost to the point of redundancy), observation of a red spot on an additional arm would provide little additional information in favor of class 1, and class-conditional independence would not apply (for similar discussions regarding the Markov condition, see Cartwright, 1993; Hausman, 1999; Park & Sloman, 2013). This intuition could be incorporated into a different, additional component of the DISC-LM.

Other steps in developing the DISC-LM include (a) to test the predictions that it provides about the development of the learners' beliefs about the structure of the task (Fig. 4), and (b) to translate the assumption of class-conditional independence into specific process model predictions, by tweaking the model such that it predicts a second, independent data dimension such as reaction times or electroencephalogram data or gaze pattern (Jarecki, Tan, & Jenny, 2016), in addition to choice predictions.

8.1.1. Implication for strategy selection

Our findings complement studies on how people adapt decision and inference strategies to the nature of a task (Gigerenzer, Todd, & the ABC Research Group, 1999; Glöckner & Betsch, 2008; Gluth, Rieskamp, & Büchel, 2014; Lieder & Griffiths, 2015; Marewski & Schooler, 2011; Mata, von Helversen, & Rieskamp, 2011; Payne, Bettman, & Johnson, 1993; Rieskamp & Otto, 2006). For instance, Gluth et al. (2014) found evidence in multiple-cue inference tasks that people's behavior and neuronal data is best described by a model formalizing dynamic switches between decision strategies over time. Our approach shows that beliefs about the nature of the task (feature dependencies) could at least implicitly be a guiding principle by which people learn to adapt inference strategies.

8.1.2. Implications for knowledge-specific learning

Our data also inform the literature on the interaction between the context of a task and the specific structural knowledge that people apply. For example, participants in an experiment by Wattenmaker, Dewey, Murphy and Medin (1986) expected to learn a linearly separable categorization structure when the cover story of a person-classification task was such that the features of one category coincided with aspects of one personality trait. However, participants did not expect linear separability when the features associated with one category belonged to different character traits. Thus, background knowledge influences structural assumptions during learning. Our findings show that even despite strong structural expectations, people can overcome their initial beliefs and fully adapt to a novel environmental structure. Thus, our findings emphasize the dynamic and adaptable nature of structural assumptions.

8.1.3. Implications for causal reasoning

In the literature on causal reasoning, conditional-independence assumptions have been investigated within the causal Markov condition in Bayes nets theory (Pearl, 2000; Spirtes et al., 1993). At least two aspects of our results have potential implications for this literature. The first implication is that whether people expect class-conditional independence to hold may depend on whether learning is through experience or from explicit verbal descriptions, and on whether choice behavior or explicit numerical judgments are measured. Causal reasoning studies (e.g., Mayrhofer & Waldmann, 2014; Rehder & Hoffman, 2005) often measure probability judgments about feature occurrences after giving participants an explicit description of a situation. Our results suggest that more implicit, behavioral, and learning-based measures may reduce violations of the causal Markov condition. This suggests an alternative methodological approach for investigating independence assumptions in causal learning and reasoning. Secondly, the dynamic adaptation of structural beliefs we found in our experiments may also hold for the degree of Markov violations in causal reasoning. Causal reasoning studies could adapt a similar paradigm by investigating causal inference in environments in which the data do or do not warrant the validity of the causal Markov condition. This approach enables systematically investigating the match between people's assumptions and inferences, the presumed causal structure of the environment, and the available learning data (e.g., von Sydow, Hagmayer, & Meder, 2016).

8.2. Beyond simplicity in early category learning

Our findings emphasize the transition between inference strategies during learning. In this sense, they are consistent with the finding that in the early stages of category learning people employ a simpler inference and categorization strategy and then gradually learn more computationally intense strategies (Love et al., 2004; Smith & Minda, 1998). Our work extends these findings by adding a notion of robustness to the notion of simplicity. The simple initial categorization strategy we proposed—assuming class-conditional feature independence—is additionally a robust strategy that often leads to accurate classification, despite its unrealistically simplistic structural assumptions (Domingos & Pazzani, 1997; Rish et al., 2001). In this sense, class-conditional independence can be viewed as a heuristic default assumption, providing an efficient means to reduce computational complexity, which works well in many situations.

Early in learning, when little information has been obtained, it is helpful to have computationally simple strategies that facilitate making inferences and decisions. But simplicity is not a virtue if it only works in very few selected statistical environments. Robustness to violation of initial assumptions is also important for cognitive systems to guard against potentially costly mistakes. Simple and robust strategies for early inferences may buy time to gather more experience, and adapt to the nuances of an environment's structure. The literature on strategy transitions in categorization is limited with respect to the question of whether the models proposed for early learning, for example prototype or linearly separable models, are robust. Our results show that people may use strategies that get the best of simplicity, robustness, and adaptability.

Acknowledgments

We thank Gregor Caregnato for data collection and Laura Martignon, Michael Waldmann, Ralf Mayrhofer, and Bob Rehder for helpful comments. This research was supported in part by grants NE 1713/1-2 to JDN, and ME 3717/2-2 to BM, from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

Notes

1. More precisely, classification requires estimating the probability that a given stimulus s_j belongs to the ith class c_i: $P(C = c_i|S = s_j) = \frac{P(S=s_j|C=c_i)P(C=c_i)}{P(S=s_j)}$ where $C \in \{c_1, \ldots, c_n\}$ denotes the class random variable, and $S \in \{s_1, \ldots, s_m\}$ the stimulus random variable. Each stimulus s_j represents one possible feature

configuration. In the text, we omit the capital letters for random variables and most subscripts to increase readability.

2. To illustrate, consider a stimulus with binary feature values. How many possible stimuli (feature configurations) exist if there are two, three, or four features? Two features yield $2^2 = 4$ stimuli, three features yield $2^3 = 8$ stimuli, four features yield $2^4 = 16$ stimuli, and so forth.
3. Generally (beyond binary classes and features), for a class vector $\boldsymbol{c}$ and D features, the number of parameters is $(\prod_{d=1}^{D} |\boldsymbol{f}_d| - 1) \cdot |\boldsymbol{c}|$, where $|\boldsymbol{c}|$ denotes the number different classes, D is the number of different features, and $|\boldsymbol{f}_d|$ the number of values the dth feature can take. By contrast, if class-conditional independence holds, the number of parameters is $\sum_{d=1}^{D}(|\boldsymbol{f}_d| - 1) \cdot |\boldsymbol{c}|$.
4. For instance, if coffee is "expensive" if it is either from Brazil or lightly roasted, but not when it is from Brazil *and* lightly roasted and also not when it is neither from Brazil nor lightly roasted, this class structure is in line with exclusive-OR. More formally, an object belongs to class $C = 1$ if it has either feature $f_1 = 1$ or feature $f_2 = 1$, but not when both or neither of the two features are present.
5. The Markov condition states that a variable in a causal network is independent of all other variables, conditional on its direct causes, except its causal descendants. The close relationship between class-conditional independence and the Markov condition is best illustrated with a common-cause network. Consider a binary cause C with three binary effects, E_1, E_2, and E_3. Applying the causal Markov condition to this causal structure entails that the three effects are independent of each other conditional on their common cause C. Now, if C represents a binary class variable and E_1, E_2, and E_3 represent three binary features, the assumption of class-conditional independence is equivalent to the causal Markov condition. Thus, class-conditional independence can be considered a special case of the Markov condition applied to a common-cause model.
6. One alternative implementation (instead of the two-part inference of base rate and stimulus likelihoods) would be a direct inference of the probabilities of the eight stimuli given the class. This direct inference, however, is not suitable for implementing class-conditional feature independence which enters only through the stimulus likelihoods.
7. In modeling, there are three primary reasons for using the deterministic arg max choice rule. First, our research focus is on comparing the model predictions with respect to the parameter π, that is, the initial belief in class-conditional independence. A probabilistic choice rule could improve the absolute fit of the model but leave the *relative* performance depending on π unaffected. A logistic transformation of the class 1 probability, such as a softmax response rule (Wills & Kruschke, 2008), shifts the posterior probabilities toward .50 but does not shift them beyond this threshold, such as from .75 to .25. Remember that our task involves four critical stimuli for which class-conditional independence predicts one class and flexible dependencies predict the opposite class. We are interested exactly in whether the response switches from below .50 to above .50. Therefore, a probabilistic response

rule would add model complexity (adding another parameter) without adding value to answer our question. Second, probabilistic choice rules require aggregating data over individuals or trials (which is common practice, e.g., Friedman et al., 1995). Aggregating over trials assumes little or no covariance of choices over time (Hannan, 1985). However, learning data are characterized by time dependencies. Therefore, time aggregation would not do justice to our data. Aggregating over individuals is also not possible because people varied greatly in their learning speed (i.e., the number of trials they needed to hit the learning criterion in our task). The third reason for the arg max rule is pragmatic. The deterministic choice makes it easiest to illustrate how the parameter π changes the DISC-LM's performance.

8. This analysis used all trials, that is, including the last 200 trials for which our learning criterion enforced 98% correct choices, because excluding the last 200 trials resulted in 19 (of 30) participants being left with fewer than 20 learning trials for one or more of the five stimuli. If we compute the median of the proportion of errors after excluding the last 200 trials, the qualitative result is unchanged, that is, fewest errors for the uncritical stimulus ($111 < 000 \approx 100 \approx 010 \approx 001$ with median error rates .09, .22, .21, .22, .20, respectively).
9. For each participant, the individual *MSE* was computed as $MSE = \frac{1}{T}\sum_t (x_t - \hat{p}_t)^2$, where t indexes trials, T is the number of trials used for parameter selection, x_t denotes the participant's choice for trial t, and $\hat{p}_t$ denotes the predicted probability for the class. This was the simulated expected value of the classification beliefs for each trial (see supplementary material A for details).
10. We used a finer grid resolution for π close to 1 because the model predictions in the lower grid regions were rather similar to each other, ceteris paribus. Each prediction by models with values of $\pi \leq .6$ differed by <1 percentage point from predictions by a model with $\pi = 0$, when comparing trials 2 to 100. By contrast, changing π from .9 to .99 resulted in a substantial difference in the predicted point estimates of the class membership. We rounded the predictions to the fourth digit.
11. The reduced number of degrees of freedom (*df* = 25 instead of 28 with 29 participants) result from the fact that some people were excluded because they had less than 20 (i.e., 10 early, 10 late) learning trials. Note that learning trials were the total trial number $T - 200$.
12. Accuracy was defined as the number of trials in which observations corresponded to the model predictions after binarizing the probabilistic predictions by an arg-max response rule (Eq. 4).
13. Again, our analysis used all trials, that is, including the last 200 trials for which our learning criterion enforced 98% correct choices, because excluding the last 200 trials resulted in 9 (of 29) participants with fewer than 20 choices for at least one stimulus type. When excluding the last 200 trials, the order of the median error rates corresponds to using all trials: The order is $111 < 000 < 100 < 010 \approx 001$ (0.09, .15, .33, .37, .39, respectively).

14. Again, we computed the fit for $\pi = 0$ and $\pi = 1$ given individually adjusted δ values.
15. Again, the number of subjects used for this analysis is lower than the total number (29) because one subject had less than 20 (10 early, 10 late) learning trials.

References

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14,* 471–485. doi:10.1017/S0140525X00070801

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 50–71. doi:10.1037/0096-1523.18.1.50

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179. doi:10.1037//0033-295X.93.2.154

Barrington, L., Marks, T. K., Hsiao, J. H.-W., & Cottrell, G. W. (2008). NIMBLE: A kernel density model of saccade-based visual memory. *Journal of Vision*, *8*, 1–14. doi:10.1167/8.14.17

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. doi:10.1016/j.jmp.2013.12.002

Bellmann, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton, NJ: Princeton University Press.

Blair, M., & Homa, D. L. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition*, *29*, 1153–1164. doi:10.3758/BF03206385

Bourne, L. E., Healy, A. F., Kole, J. A., & Graham, S. M. (2006). Strategy shifts in classification skill acquisition: Does memory retrieval dominate rule use? *Memory & Cognition, 34,* 903–913. doi:10.3758/BF03193436

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*, 2–16. doi:10.1016/j.cognition.2010.10.004

Cartwright, N. (1993). Marks and probabilities: Two ways to find causal structure. In F. Stadler (Ed.), *Scientific philosophy: Origins and development* (pp. 113–119). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Chickering, D. M., & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, *29*, 181–212. doi:10.1023/A:1007469629108

Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130. doi:10.1023/A:1007413511361

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140. doi:10.1037/0096-3445.127.2.107

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408. doi:10.3758/BF03193784

Flach, P. A., & Lachiche, N. (2004). Naïve Bayesian classification of structured data. *Machine Learning*, *57*, 233–269. doi:10.1023/B:MACH.0000039778.69032.ab

Friedman, D., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. *Journal of Mathematical Psychology*, *39*, 164–178. doi:10.1006/jmps.1995.1018

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*(3), 215–228.

Gluth, S., Rieskamp, J., & Büchel, C. (2014). Neural evidence for adaptive strategy selection in value-based decision-making. *Cerebral Cortex*, *24*, 2009–2021. doi:10.1093/cercor/bht049

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154. doi:10.1080/03640210701802071

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 323–328). Austin, TX: Cognitive Science Society.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (Chapter 3). Cambridge, UK, Cambridge University Press.

Hannan, M. T. (1985). Problems of aggregation. In H. M. Blalock (Ed.), *Causal models in the social sciences* (2nd ed., Chapter 17, pp. 403–440). Hawthorne, NY: Transaction Publishers.

Hausman, D. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, *50*, 521–583. doi:10.1093/bjps/50.4.521

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Available at http://arxiv.org/abs/1502.01852. Accessed June 01, 2015.

Homa, D. L., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 444–458. doi:10.1037/0278-7393.17.3.444

Jarecki, J. B., Tan, J. H., & Jenny, M. A. (2016). *What is a cognitive process model? A disambiguation*. Berlin: Max Planck Institute for Human Development. doi: 10.2139/ssrn.2544831

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482–553. doi:10.1016/S0010-0285(02)00505-4

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*, 133–156. doi:10.1037/0096-3445.132.1.133

Lieder, F. & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1362–1367). Austin, TX: Cognitive Science Society.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J.. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*, 1179–1189. doi:10.1111/j.1365-2966.2008.13689.x

Little, D. R., & Lewandowsky, S. (2009). Better learning with more error: Probabilistic feedback increases sensitivity to correlated cues in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1041–1061. doi:10.1037/a0015902

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332. doi:10.1037/0033-295X.111.2.309

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*, 316–338. doi:10.1037/a0022684

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An introduction to information retrieval (online)*. Cambridge, UK: Cambridge University Press. doi:10.1109/LPT.2009.2020494

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*, 393–437. doi:10.1037/a0024143

Marr, D. (1982). General introduction. In D. Marr (Ed.), *Vision: A computational investigation into the human representation and processing of visual information* (pp. 3–7). San Francisco, CA: W. H. Freeman.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361. doi:10.1016/j.jmp.2008.04.003

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (Chapter 10, pp. 189–211). Chichester, UK: John Wiley & Sons.

Mata, R., von Helversen, B., & Rieskamp, J. (2011). When easy comes hard: The development of adaptive strategy selection. *Child Development*, *82*, 687–700. doi:10.1111/j.1467-8624.2010.01535.x

Mayrhofer, R., & Waldmann, M. R. (2014). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*, 65–95. doi:10.1111/cogs.12132

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668–693. doi:10.1037/a0032963

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148. Available at: http://journal.sjdm.org/12/12314/jdm12314.html

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 333–352. doi:10.1037/0278-7393.10.3.333

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. doi:10.1037//0033-295X.85.3.207

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 355–368. doi:10.1037//0278-7393.7.5.355

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 241–253. doi:10.1037//0278-7393.7.4.241

Mehta, R., & Williams, D. A. (2002). Elemental and configural processing of novel cues in deterministic and probabilistic tasks. *Learning and Motivation*, *33*, 456–484. doi:10.1016/S0023-9690(02)00008-5

Minsky, M. & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Monti, S. & Cooper, G. F. (1999). A Bayesian network classifier that combines a finite mixture model and a naïve Bayes model. In K. B. Laskey & H. Prade (Eds.), *UAI'99 proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 447–456). San Francisco, CA: Morgan Kaufmann.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148. doi:10.1037/0033-295X.108.1.113

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518. doi:10.1037/a0016104

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999. doi:10.1037/0033-295X.112.4.979

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969. doi:10.1177/0956797610372637

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1992). Exemplar, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, Chapter 8, pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 999–1019. doi:10.1037/0278-7393.33.6.999

Nosofsky, R. M., Kruschke, J. K., & Mckinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 211–233. doi:10.1037/0278-7393.18.2.211

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79. doi:10.1037//0033-295X.101.1.53

Nosofsky, R. M., & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 608–629. doi:10.1037/0096-1523.31.3.608

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940. doi:10.1037//0278-7393.28.5.924

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*, 186–216. doi:10.1016/j.cogpsych.2013.09.002

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363. doi:10.1037/h0028558

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*, 581–602. doi:10.1016/j.cognition.2007.11.007

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407. doi:10.1016/0010-0285(72)90014-X

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. doi:10.1016/j.cogpsych.2014.02.002

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314. doi:10.1016/j.cogpsych.2004.09.002

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41. doi:10.1016/j.cogpsych.2004.11.001

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236. doi:10.1037/0096-3445.135.2.207

Rish, I., Hellerstein, J., & Thathachar, J. (2001). An analysis of data characteristics that affect naïve Bayes performance. IBM Technical Report RC21993, IBM Watson Research Center.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *210*, 178–210. doi:10.1006/jmps.2001.1379

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134. doi:10.1016/j.cogpsych.2016.05.002

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, Alexander, C., Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. doi:10.1007/s11263-015-0816-y

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 1–6.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167. doi:10.1037/a0020511

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *The Journal of Neuroscience*, *25*, 2941–2951. doi:10.1523/JNEUROSCI.3401-04.2005

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*, 1–25. doi:10.1016/j.cognition.2011.02.010

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 69–69. doi:10.1037/h0090333

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811. doi:10.1037//0278-7393.28.4.800

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Cambridge, MA: MIT Press.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., & Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society. Series A (General)*, *144*, 145. doi:10.2307/2981918

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749. doi:10.3758/PBR.15.4.732

Vigo, R. (2013). The GIST of concepts. *Cognition*, *129*, 138–162. doi:10.1016/j.cognition.2013.05.008

von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, *44*, 469–487. doi:10.3758/s13421-015-0568-5

Waldmann, M. R. & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Lawrence Erlbaum Associates.

Walsh, C. & Sloman, S. (2008). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (Chapter 21, pp. 345–357). New York: Lawrence Erlbaum Associates.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*, 158–194. doi:10.1016/0010-0285(86)90011-3

Wills, A. J. & Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (Chapter 9, 2002, pp. 984–1022). New York: Cambridge University Press. doi:10.1006/jmps.2001.1379

Zhang, H. & Ling, C. X. (2001). Geometric properties of naïve Bayes in nominal domains. In *Machine learning: {ECML} 2001, 12th European conference on machine learning* (Vol. 2167, pp. 588–599). Freiburg, Germany: Springer Berlin Heidelberg. doi:10.1007/3-540-44795-4_50

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Supplementary material A.

Appendix A: Simulation results Conservatism parameter δ

Fig. A1 shows simulation results of the DISC-LM where we vary the conservatism parameter δ and the values of the prior structural belief parameter π. Each of $N = 200$ simulated subjects experienced stimuli drawn at random from the deterministic and the probabilistic task environment environment (panel a, at top). A separate 200 simulated subjects experienced stimuli drawn at random from the probabilistic environment (panel b, at bottom).

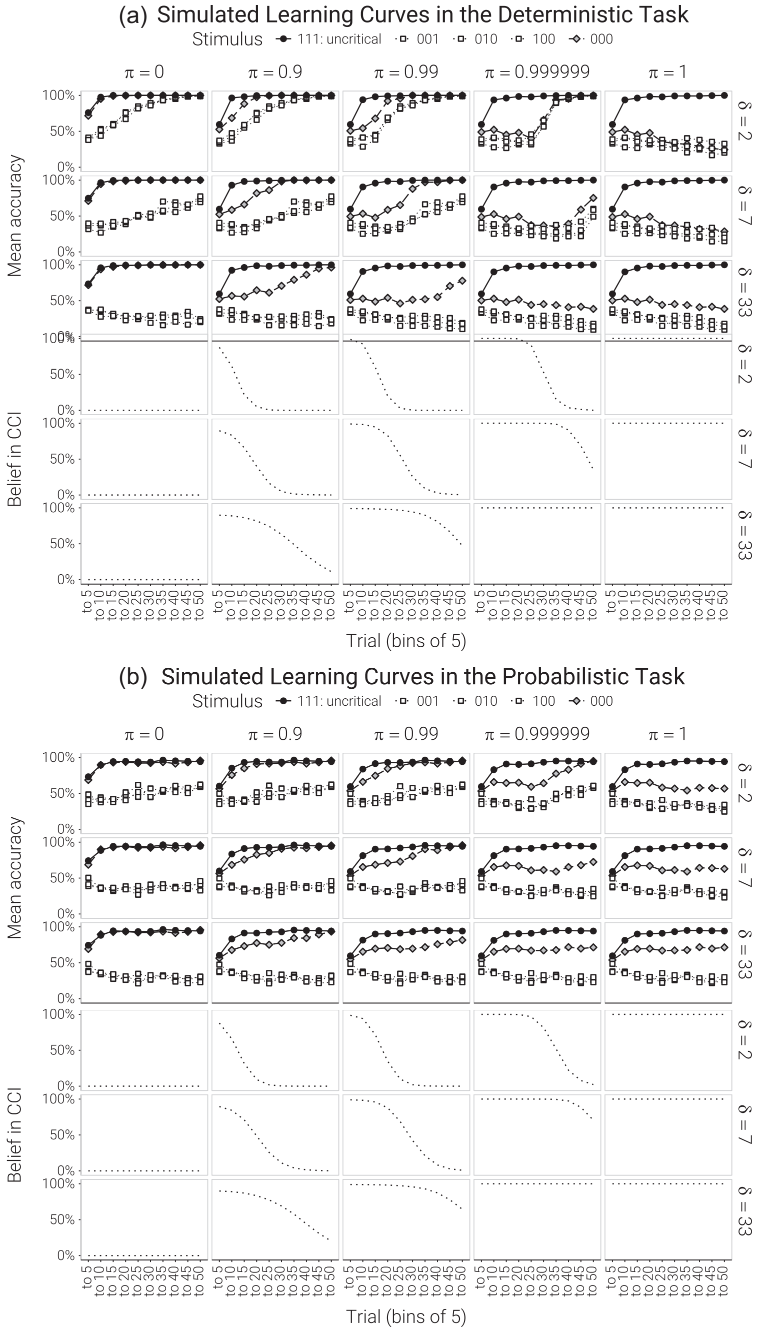

Fig. A1. Simulation of learning given different values of the conservatism parameter δ. (a) Simulated learning curves in the deterministic task. (b) Simulated learning curves in the probabilistic task. The higher the value of δ, the slower the learning for all stimuli. Higher values of π affect the critical stimuli but not the uncritical stimulus 111. CCI, class-conditional independence.

Appendix B: Experimental instruction feedback during learning

The feedback that participants received every 100 trials was as follows:

> *How are you doing? If you continue responding like in the last 200 trials, you will average about x% correct. The optimal strategy achieves about y%.*
> *Mini-FAQ: Q: I've only learned one feature. Is that okay? A: No. More than one feature matters. You must learn all the features to be able to learn to categorize the plankton specimen.*

The variable x was the accuracy that would be achieved on average if the participant would respond in the same way as in the most recent 200 trials, and the stimulus configurations would occur exactly according to their average frequencies. The variable y was the maximum achievable average accuracy, if stimuli would occur according to their average frequencies. (Each stimulus was chosen at random according to the theoretical frequencies of occurrence, in each trial in the learning task. Because of this, a participant's actual accuracy is typically not identical to the theoretical accuracy that would be achieved by their pattern of responses to the various stimuli.) Both numbers were rounded to the nearest tenth of a percent. See Tables 1 and 2 for the expected classification accuracies $P(class|stimulus)$ in Experiment 1 and Experiment 2, respectively.