How Category Selection Impacts Inference Reliability: Inheritance Inference From an Ecological Perspective

Paul D. Thorn, Gerhard Schurz

Duesseldorf Center for Logic and Philosophy of Science (DCLPS), Heinrich Heine University

Received 15 June 2020; received in revised form 15 March 2021; accepted 16 March 2021

Abstract

This article presents results from a simulation-based study of inheritance inference, that is, inference from the typicality of a property among a “base” class to its typicality among a subclass of the class. The study aims to ascertain which kinds of inheritance inferences are reliable, with attention to the dependence of their reliability upon the type of environment in which inferences are made. For example, the study addresses whether inheritance inference is reliable in the case of “exceptional subclasses” (i.e., subclasses that are known to be atypical in some respect) and attends to variations in reliability that result from variations in the entropy level of the environment. A further goal of the study is to show that the reliability of inheritance inference depends crucially on which sorts of base classes are used in making inferences. One approach to inheritance inference treats the extension of any atomic predicate as a suitable base class. A second approach identifies suitable base classes with the cells of a partition (of a preselected size $k$) of the domain of objects that satisfies the condition of maximizing the similarity of objects that are assigned to the same class. In addition to permitting more inferences, our study shows that the second approach results in inheritance inferences that are far more reliable, particularly in the case of exceptional subclasses.

Keywords: Default inheritance; Nonmonotonic reasoning; Probabilistic reasoning; Conceptual spaces; Ecological rationality
1. Introduction

In executing an inheritance inference, one reasons from a premise stating that a given property is “typical” among a class of individuals and concludes that the property is typical among a subclass of the class. Inheritance inference is exemplified by the following schema, where C is known as the “base class” for the inference:

Property ϕ is typical among the members of class C.
SC is a subclass of C.
-----------------------------
ϕ is typical among the members of SC.

Inheritance inference has been studied extensively by logicians, formal epistemologists, and researchers in artificial intelligence. For example, a formalism known as “semantic inheritance networks” was developed for the purpose of modeling normatively correct inheritance inference, with a focus on characterizing the complex ways in which inheritance inferences may interact (Horty, 2007; Horty, Thomason, & Touretzky, 1990; Touretzky et al., 1987). Questions concerning inheritance inference have also been central to the study of nonmonotonic logics. A central concern within the field of nonmonotonic logic is to decide which weakening of the monotonicity property of classical logic (e.g., cautious monotonicity, rational monotonicity, or some stronger principle) is appropriate for respective nonmonotonic logics, given a logic’s interpretation and intended application (Adams, 1975, 1986; Brewka, 2009; Gabbay, 1984; Goldszmidt & Pearl, 1991, 1996; Halpern, 2009; Hawthorne & Makinson, 2007; Kraus, Lehmann, & Magidor, 1990; McDermott & Doyle, 1980; Pearl, 1988, 1990; Poole, 1988, 1994; Reiter, 1980, 1987; Schurz, 1997, 1998; Schurz & Thorn, 2012; Thorn, Eichhorn, Kern-Isberner, Schurz, 2015; Thorn & Schurz, 2014, 2016). Inheritance inference is often at center stage in discussions of nonmonotonic logic, because inheritance inference is one of the central forms of nondeductive inference that nonmonotonic logics are meant to codify, and because claims about which weakening of classical monotonicity is appropriate translate directly into claims about which sorts of inheritance inference should be permitted.

In this article, we present results from a simulation-based study of inheritance inference. The study aims to ascertain which kinds of inheritance inferences are reliable, with attention to variations in reliability that are contingent upon the type of environment in which inferences are made. For example, our study addresses whether inheritance inference is reliable in the case of exceptional subclasses (i.e., subclasses that are known to be atypical in some respect), and it attends to variations in reliability that result from variations in the entropy level of the distribution of properties in the environment. In addition to the preceding, a principal goal of our study was to investigate the manner in which the reliability of inheritance inference depends on which sorts of base classes are used in making inferences.

In conducting our study, we considered two approaches to identifying suitable base classes. Both approaches assume that, in any given situation, there are a number of atomic predicates that serve as the basic means of characterizing objects. As a matter of convention, we call the extension of an atomic predicate an “atomic property.” The first approach treats
any atomic predicate as determining a suitable base class for a cogent inheritance inference. This approach, thereby, guarantees that the members of any suitable class will be similar in regard to a single atomic property, but it places no further limitation on the heterogeneity of the objects that are members of a given class. The second approach is motivated by the idea that a well-adapted system of categories will maximize the overall similarity of objects that are assigned to the same category and maximize the overall dissimilarity of objects that are assigned to different categories (Kaufman & Rousseeuw, 1990, p. 68). In order to implement this idea, the second approach identifies suitable base classes with the cells of a partition (of a preselected size $k$) of the domain that minimizes within class variance. Such partitions are generated by a method known as $k$-means clustering, which we describe in Section 3.

The second approach to identifying suitable base classes requires that the boundaries between classes be drawn in order to maximize the similarity of objects assigned to the same class. As a consequence, the boundaries of the class categories identified by the second approach are determined by contingent features of the distribution of atomic properties in the environment (i.e., by the relative frequencies of objects possessing respective combinations of atomic properties). Since the categories corresponding to such classes are “fitted” to the distribution of atomic properties in the environment, we call these classes “fitted classes.” Such classes correspond to convex regions of an $n$-dimensional quality space, and thereby correspond to concepts in a conceptual space, in the sense of Gärdenfors (2000).

According to Gigerenzer and others (Gigerenzer, Todd & the ABC Research Group, 1999; see also Schurz, 2005; Simon, 1955, 1956), it is often fruitful to adopt an “ecological” perspective and evaluate the rationality of an inference rule by considering its performance (e.g., reliability and frugality) within the kinds of environments in which the rule is typically used. If we adopt such an ecological perspective, then, it is appropriate to evaluate inheritance inference under conditions where class categories are “fitted” in the sense described in this article, since it is plausible to expect that the actual class categories adopted by humans are adapted to their environments as a result of biological and cultural evolution, in ways analogous to fitted classes (Douven & Gärdenfors, 2020; Jäger, 2007). This ecological perspective on inheritance inference will be supported by the results of this article: In addition to permitting more inferences, inheritance inference based on fitted classes is far more reliable than inheritance inference based on unfitted classes. The difference in the performance of the two approaches is particularly great in the case of exceptional subclasses.

While our study is focused on identifying factors that are relevant to the normative correctness of inheritance inference, the results of the study also recommend extensions of previous psychological studies in order to determine whether human subjects are sensitive to these factors (Connolly, Fodor, Gleitman, & Gleitman, 2009; Gagné & Spalding, 2014; Hampton, Passanisi, & Jönsson, 2011; Jönsson & Hampton, 2008; Jönsson & Hampton, 2012; Strößner & Schurz, 2020). Before sketching the connection of our study with psychological work in Section 3, Section 2 considers some formal reasons for thinking that inheritance inference is cogent and introduces a form of inheritance inference (i.e., inheritance in the case of exceptional subclasses) whose cogency has been a topic of some controversy in the fields of logic and artificial intelligence.
2. Cautious monotony and inheritance in the case of exceptional subclasses

As we already mentioned, inheritance inferences have the following form:

Property $\phi$ is typical among the members of class C.
SC is a subclass of C.
-----------------------------
$\phi$ is typical among the members of SC.

Within our study, the reliability of inheritance inference is identified with the tendency of inheritance inferences to deliver true conclusions, given true premises. In order to make such reliability claims, we settle on a straightforward semantics for typicality claims. We say that a property is “typical” of a class (or subclass) if and only if the relative frequency of the property among the members of the class (or subclass) meets or exceeds a given fixed threshold $r$, where $0 < r < 1$. (Different values for $r$ were considered in our study.) Although alternative criteria for typicality are possible, the present criterion permits us to address the question that is of critical interest in discussions of inheritance inference, namely, whether it is reasonable to infer a high frequency of a property among a subclass, given a high frequency of the property among the superordinate class. Given our simple typicality criterion, the general form of inheritance inference may be represented as follows:

The relative frequency of individuals with $\phi$ among C is at least $r$.
SC is a subclass of C.
-----------------------------
The relative frequency of individuals with $\phi$ among SC is at least $r$.

As a prelude to our experimental evaluation of inheritance inference, it is of interest to consider what a priori considerations bear on the reasonableness of inheritance inference. The first important observation is that inheritance inference is invalid (i.e., the truth of the premises of an inheritance inference is consistent with the falsity of its conclusion). In fact, the truth of the premises of an inheritance inference (for given $\phi$, SC, C, and $r$) places no constraint whatsoever on the relative frequency of individuals with $\phi$ among SC.

Despite its invalidity, there are conditions under which inheritance inference is reasonable. One such condition is codified by an argument form known as Cautious Monotony (CM):\(^1\)

$\phi$ is typical among the members of C.
$\psi$ is typical among the members of C.
-----------------------------
$\phi$ is typical among the members of $C \cap \psi$.

Where SC is $C \cap \psi$, we see that CM corresponds to inheritance inference in cases where membership in SC is typical among C. For example, suppose our base class is the set of players in the National Basketball Association ($C = \text{the set of NBA players}$), and we know that having a height of greater than 2 meters ($\phi$) is typical for NBA players, and that having an annual income greater than 6 million USD ($\psi$) is also typical for NBA players. In that case, Cautious Monotony licenses an inference to the conclusion that having a height of greater
than 2 meters ($\varphi$) is typical among NBA players with an annual income greater than 6 million USD ($C \cap \psi$).

If typicality is understood as high relative frequency (as we assume here), then it is possible to justify inference in accordance with CM, and inheritance inference in cases where SC is typical among C, by appeal to the following formula, which holds for relative frequencies, and for conditional probabilities, in general. Within the formula, $\text{freq}(T|R)$ denotes the relative frequency of $T$ among $R$ (i.e., $\text{freq}(T|R) = |T \cap R|/|R|$, if $|R| > 0$, and is undefined, otherwise):

\[
\text{IP} \quad 1 - \text{freq}(\varphi|SC) \leq (1 - \text{freq}(\varphi|C)) / \text{freq}(SC|C) \quad (\text{Adams}, 1975; \text{Schurz}, 1998).
\]

The preceding formula tells us that the difference between $\text{freq}(\varphi|SC)$ and $\text{freq}(\varphi|C)$ is bounded as a function of $\text{freq}(SC|C)$ (i.e., the relative size of SC among C). If $\text{freq}(SC|C) = 1$, then $\text{freq}(\varphi|SC)$ and $\text{freq}(\varphi|C)$ are identical. Similarly, if the relative size of SC among C is large, then $\text{freq}(\varphi|SC)$ cannot differ much from $\text{freq}(\varphi|C)$. For example, if 99% of sea birds (C) have webbed feet, and 90% of sea birds are capable of flight ($\varphi$), then [IP] tells us that at least 88.4% of sea birds with webbed feet are capable of flight.

Although CM (backed by [IP]) is suggestive of one avenue of justifying inheritance inference, that avenue is rather limited, since it only applies when the relative size of SC among C is large. But inheritance inference is plausible independently of the relative size of SC among C, as exemplified by so-called “direct inference,” where one reasons from frequency information for a class to a single-case probability for a single element of that class. For example, if I know that 95% of dachshunds live at least $9 \frac{1}{2}$ years, and I know that Max is a dachshund, and no other facts specific to Max, then it looks like it is reasonable to infer (by direct inference) that there is a .95 probability that Max will live at least $9 \frac{1}{2}$ years (Bacchus, 1990; Kyburg, 1974; Pollock, 1990; Reichenbach, 1949; Thorn, 2012, 2015; Venn, 1866).²

There are other approaches to justifying inheritance inference that appeal to combinatorial facts that relate the relative frequency of a characteristic among a set (such as C) to the relative frequencies of the characteristic among the subsets of the set (such as SC). One such approach appeals to combinatorial variants of the Law of Large Numbers, which tell us that the vast majority of the subsets of C of sufficient size differ from C by no more than a small amount regarding the frequency of any given property $\varphi$ (McGrew, 2001; Thorn, 2014, 2017). In making an inheritance inference, this approach proceeds by treating SC as if it were a randomly selected subset of C, as a basis for concluding that it is very probable that the frequency of $\varphi$ among SC is very similar to the frequency of $\varphi$ among C. A significant problem with this approach is to justify the treatment of SC as if it were a randomly selected subset of C. Beyond such worries, one might regard the described approach to justifying inheritance inference as too limited, because it cannot be applied as a basis for justifying inheritance inference in the case of so-called “exceptional subclasses,” which we will now briefly describe.

In cases where the following propositions are true of a triple $\langle SC, C, \psi \rangle$, we say that SC is an exceptional subclass of C, with respect to $\psi$:

$\psi$ is typical among C.
SC is a subclass of C.
$\psi$ is not typical among SC.
For example, since the capacity for flight ($\psi$) is typical among birds (C), but not among penguins (SC), penguins are an exceptional subclass of birds, with respect to the capacity for flight. An inheritance inference concerning a property $\varphi$ among a subclass SC is described as “inheritance inference in the case of an exceptional subclass,” when the preceding conditions are known to hold for a respective triple $\langle SC, C, \psi \rangle$. So, since penguins are exceptional birds, with respect to flight ($\psi$), inference to the conclusion that a poor sense of smell ($\varphi$) is typical among penguins (SC), given that a poor sense of smell is typical among birds (C), counts as an inheritance inference in the case of an exceptional subclass.

Within early discussions of default inheritance, the fact that existing formal systems did not license inheritance inference in the case of exceptional subclasses was regarded as a problem with the adequacy of the systems (Benferhat, Cayrol, Dubois, Lang, & Prade, 1993; Goldszmidt & Pearl, 1996; Goldszmidt et al., 1990). The problem of licensing inheritance inference in the case of exceptional classes was called the Drowning Problem, because for existing systems (e.g., Lehmann, 1989; Pearl, 1990) the exceptionality of a subclass with respect to a single property was sufficient to drown all other inheritance inferences for that subclass. Systems that avoided the Drowning Problem (licensing inheritance inference in the case of exceptional subclasses) were subsequently developed (e.g., Benferhat et al., 1993; Goldszmidt & Pearl, 1996; Goldszmidt et al., 1990; Kern-Isberner, 2001; Tan & Pearl, 1995). With numerous available solutions to the technical problem, it is now controversial whether inheritance inference in the case of exceptional subclasses is desirable. In his reference article on defeasible reasoning, Koons (2017) rejects inheritance inference in the case of exceptional subclasses, and claims that “consensus is growing that the Drowning Problem should not be ‘solved’.” Other researchers rejecting inheritance inference in the case of exceptional subclasses include Elio and Pelletier (Elio & Pelletier, 1994; Pelletier & Elio, 1997), Wobcke (1995, p. 85), and Bonevac (2003, pp. 461–465). In their reference article on nonmonotonic logic, Strasser and Antonelli (2019) describe inheritance inference in the case of exceptional subclasses as an “advanced issue” and treat the question of its desirability as an open problem.

Skeptics about inheritance inference in the case of exceptional subclasses are committed to the claim that the atypicality of properties among subclasses is probabilistically dependent: Learning that a subclass (e.g., penguin) is atypical in one respect (e.g., the capacity for flight) makes it more likely that it is atypical in other respects (e.g., having a poor sense of smell). One of the objectives of the present article is to determine whether skepticism about the reliability of inheritance inference in the case of exceptional subclasses is justified. In the end, our simulation studies suggest that being atypical in one respect is positively relevant to being atypical in other respects. But there are significant differences in the reliability of inheritance inference in the case of exceptional subclasses for fitted and unfitted classes, and it may be that the risk of error in the case of fitted classes is tolerable.

3. Inheritance inference and concept composition

Within cognitive psychology, studies of inheritance inference have mostly been connected to a debate concerning the nature of concepts and concept composition. On one side of this debate, proponents of the classical view of concepts hold that the primary mechanism of
concept composition corresponds to the set theoretic operation of intersection: The extension of a composed concept is, at least in its first instance, the intersection of the extensions of the composing concepts, with any modification of that content coming only as the result of secondary cognitive processes (Connolly et al., 2009). On the other side of this debate, proponents of the prototype theory of concepts subscribe to the view that the prototype corresponding to a composed concept is derived (prior to any modification of content resulting from secondary cognitive processes) from the prototypes associated with the composing concepts (Hampton, 1987; Smith, Osherson, Rips, & Keane, 1988).

For prototype theories of concepts, it is expected that the prototypes of composed concepts tend to incorporate the prototypical features of the prototypes from which they are derived (Hampton, 1987; Hampton et al., 2011; Jönsson & Hampton, 2008, 2012; Rosch, 1975; Rosch & Mervis, 1975; Smith et al., 1988). The formal mechanism of default inheritance, as studied by logicians, provides a plausible model of the conditions under which composed concepts inherit prototypical properties. However, contrary to Connolly et al. (2009), it would be too quick to treat failures of default inheritance as disconfirming prototype theories of concepts and concept composition (Jönsson & Hampton, 2008). Nevertheless, assuming some version of the prototype theory of concepts is correct, one would expect composed concepts to inherit the prototypical properties of composing concepts in a manner roughly in line with formal theories of default inheritance.

Preliminary evidence of default inheritance in connection with concept composition is reported in Hampton (1987). In a challenge to prototype theories, Connolly et al. (2009) report results illustrating the so-called modifier effect, showing that human subjects assign less likelihood to generic statements (e.g., ducks have webbed feet) when the subject term is modified (e.g., baby ducks have webbed feet). A number of follow-up studies have been conducted with the aim of explaining the modifier effect and assessing its relevance to prototype theories of concepts (Jönsson & Hampton, 2012; Strößner & Schurz, 2020). Of particular interest, Hampton et al. (2011) found a difference in the rated likelihood of modified generic statements when the modifier concerns a mutable property (i.e., a property that can easily be imagined to be different) in comparison to a central property (i.e., a property that has causal dependencies to other properties associated with the relevant concept). Within their study, subjects were less willing to rate a modified statement as less likely than its original, when the modifier was central as opposed to mutable. This finding is consistent with the hypothesis that people think that the distribution of the central prototypical properties of a class is less likely to be variable among subclasses. A similar hypothesis was offered by Elio and Pelletier in discussing their finding that subjects are more likely to affirm the conclusion of an inheritance inference when the involved class is a natural kind category as opposed to an artifact category; that is, persons regard the subclasses of natural kind categories as less variable in their properties than the subclasses of artifact categories (Elio & Pelletier, 1994; Pelletier & Elio, 1997). The results of Hampton et al. (2011) and of Elio and Pelletier suggest that human subjects are sensitive to general factors that are relevant to the reliability of inheritance inference, inasmuch as: (1) it is reasonable to expect less variability in the distribution of central prototypical properties among subclasses, and
less variability in the distribution of properties among the subclasses of natural kinds, and (2) less variability in these respects is conducive to reliable inheritance inference.

A central finding of the study presented in this article is that inheritance inference based on so-called fitted classes is more reliable. In light of the findings of Hampton et al. (2011), and of Elio and Pelletier (Elio & Pelletier, 1994; Pelletier & Elio, 1997), which appear to show that human subjects are sensitive to general factors that are relevant to the reliability of inheritance inference, it is natural to conjecture that human subjects are (1) sensitive to the difference between fitted and unfitted classes, and (2) less credulous concerning inheritance inference based on fitted classes. The results of our study, thus, recommend extensions of previous psychological studies to determine whether or not the preceding conjectures are correct. In connection with these conjectures, it should be observed that our distinction between fitted and unfitted classes is orthogonal to the distinction between natural kind categories and artifact categories, as discussed by Elio and Pelletier (Elio & Pelletier, 1994; Pelletier & Elio, 1997). While natural kind categories will typically correspond to fitted classes, many categories of artifact will also correspond to fitted classes, inasmuch as the categorization of artifacts is driven by the goal of classifying objects according to their similarities and differences, along multiple dimensions.

In considering the relevance of the study presented here to the work of psychologists on the nature of concepts and concept composition, it should be emphasized that the implementation of our notion of fitted classes by means of k-means clustering is intended as a proxy for a more nuanced notion of fitted classes. A more nuanced elaboration of the notion of a fitted class might well correspond to the notion of a “natural concept” (Douven & Gärdenfors, 2020), where natural concepts are understood as exactly the type of concepts to which prototype theories of concepts properly apply (Hampton, 1987). If a connection can be made between fitted concepts and the type of concepts to which prototype theories of concepts properly apply, then the results of our study are consonant with prototype theories of concepts, particularly if our conjecture is correct that human subjects are more credulous concerning inheritance inference based on fitted classes.

4. Base class suitability criteria

We assume that the classes, subclasses, and properties that are the subject of inheritance inference are formulated in a first-order language, and that this language is based upon a finite set of atomic predicates. Because our interest is in inheritance inference in general (and not some particular formal system), we will not worry too much about the precise syntax of the language in which inheritance inference is formulated. In order to simplify our presentation, we equate properties with the sets of objects possessing respective properties.

In the following sections, we evaluate two approaches to identifying suitable base classes. The first approach treats a set as a suitable base class if and only if the set is the extension of an atomic predicate (i.e., an atomic property). Although the atomic predicates adopted by a linguistic community will typically be adapted to the community’s needs and environment, we describe the first approach as employing “unfitted” classes, since class membership
(according to the approach) is dictated solely by the possession of a single atomic property, with the possibility that class members are heterogeneous in all other respects. The second approach differs from the first inasmuch as it opts for classes that are fitted to groups of objects according to overall similarity.

In implementing the second approach, we employed a clustering method known as “k-means clustering” (Lloyd, 1982; MacQueen, 1967). This method returns a partition of the relevant domain of objects that minimizes the sum of squared distances between objects and the mean values of the cells to which respective objects are assigned. In the special case considered here, each object is characterized by an \( n \)-dimensional vector of binary values, where \( n \) is the relevant number of atomic properties, \( \varphi_1, \ldots, \varphi_n \). We then employ a function \( \phi \) that returns the appropriate vector, when given an object. For example, in the case where there are only two atomic properties \( \varphi_1 \) and \( \varphi_2 \), and an object \( x \) is in \( \varphi_1 \) but not in \( \varphi_2 \), we have \( \phi(x) = (1, 0) \). Let \( \Pi = \{c_1, \ldots, c_k\} \) be a partition of the domain of objects \( D \) (i.e., for all \( i, j \): \( c_i \subseteq D \), \( c_i \cap c_j = \emptyset \), \( c_1 \cup \ldots \cup c_k = D \)). Given such a partition, \( \mu_i(\Pi) \) designates the centroid of \( c_i \), namely, the vector whose \( n \) components are the mean values of the \( n \) properties among the members of \( c_i \). Finally, k-means clustering selects a partition \( \Pi' = \{c_1, \ldots, c_k\} \) of the domain of objects that minimizes the sum of squared differences of the individual property values from the mean property values in a given cell \( c_i \), summed up for all cells of the partition. In other words, the partition selected by k-means clustering minimizes the following function:

\[
f(\Pi') = \sum_{c_i \in \Pi'} \sum_{x \in c_i} (\phi(x) - \mu_i(\Pi'))^2
\]

Fig. 1 illustrates the result of applying k-means clustering to vectors of binary-values. The four illustrations represent the result of applying k-means clustering to the same data set, differing only in the number of clusters, \( k \): The top left illustration represents the case \( k = 1 \) (one cluster), the top right the case where \( k = 2 \) (two clusters), the bottom left the case where \( k = 3 \) (three clusters), and the bottom right the case where \( k = 4 \) (four clusters). Black dots (vertices) represent the position of object types (i.e., complete specifications of atomic properties), with the number of objects of each type indicated as a percentage of the total population. Blue dots represent the position of centroids, with dotted lines indicating which object types are attached to which centroids. The four illustrations exhibit an interesting pattern that is typical of k-means clustering, in the case of binary-valued data, namely: When \( k \) is small (in comparison to the number of property dimensions), centroids are positioned in the interior of the space, but as we consider progressively larger values of \( k \), centroids are “pulled” toward the surface of the space, then to the edges, and eventually to vertices (with the latter effect occurring, if we consider values of \( k \) greater than 4). Within the present framework, the notion of typicality that we employ (i.e., a property is typical of a class if and only if the relative frequency of the property among the members of the class meets or exceeds a given fixed threshold \( r \)) has a geometrical interpretation: In the case of a space representing three properties, a property is typical for a class if and only if the Euclidean distance between the centroid for the class and the face of the cube representing the property is no greater than \( 1 - r \). More generally, in the case of an \( n \)-cube representing \( n \) properties, a property is typical
for a class if and only if the Euclidean distance between the centroid for the class and the $(n−1)$-cube representing the property is no greater than $1−r$.

We chose $k$-means clustering as our method for identifying “fitted” classes, because the method is conceptually simple (i.e., easy to grasp) and widely known (Hastie, Tibshirani, & Friedman, 2009, pp. 509, 516, 545). The fact that $k$-means clustering has a relatively fast implementation was also advantageous (Inaba et al., 1994, p. 335; Hartigan & Wong, 1979, p. 102). It should be acknowledged, however, that $k$-means clustering is not always the best algorithm to use when clustering a data set (MacKay, 2003, pp. 287–289). Indeed, problems arise for some data sets due to the assumption implicit in $k$-means clustering that objects are always assigned to the nearest centroid, which implies that all clusters are convex, and that clusters are typically of similar size. As is well known, the assumption that objects are always assigned to the nearest centroid leads to intuitively incorrect cluster models in some cases, such as for the Iris and “mouse” data sets (MacKay, 2003, p. 288).³

Further worries may arise concerning the use of $k$-means clustering in clustering vectors of binary-valued variables (i.e., vectors of 0s and 1s), where individual vectors are used to represent the characteristics of individual objects. There are two sorts of worry. One sort of
worry is conceptual and concerns the appropriateness, or “meaningfulness,” of (i) taking the mean values of such vectors as representing a meaningful feature of the collective (Huang, 1998, p. 284), and/or (ii) measuring the (dis)similarity of objects, according to (squared) Euclidean distance (Finch, 2005, p. 86). For example, in a case where the value of the centroid for some binary-valued feature, say blueness, has a nonbinary-value, say .7, it would be odd to say that, among the cluster, the objects are, on average, .7 blue. However (contrary to (i)), the values for the centroid for any given cluster do represent meaningful and highly informative features of the cluster, namely, the relative frequencies of characteristics among the collective. Contrary to (ii), it is reasonable to cluster objects, represented by vectors of binary-values, by minimizing the total distance of individual objects to centroids consisting of vectors of real numbers: Such distances represent the difference between the characteristics of an individual and the frequency of the characteristic among the cluster to which the individual is assigned (Ng, Li, Huang, & He, 2007; Salem et al., 2017).

Measuring differences between vectors using squared Euclidean distance also has several advantages. First, due to the Pythagorean Theorem, the distance between any two vectors is equal to the sum of the distances between the components of the vectors. Second, $k$-means clustering, based on minimizing squared Euclidean distance, typically results in centroid placement that is nonarbitrary and intuitively satisfying. For example, if there is a single binary-valued characteristic, and the numbers of objects with and without the characteristic are identical, then for all $r$ in $[0, 1]$, a centroid at position $r$ minimizes (unsquared) Euclidean distance, while a centroid at position .5 uniquely minimizes squared Euclidean distance.

We have heard it expressed (though not in print) that $k$-means clustering frequently leads to “arbitrary” results when it is applied to vectors of binary values, due to the possibility of there being no unique partition that minimizes $f$ (i.e., the sum of the squared Euclidean distances of objects to centroids). But $k$-means clustering only leads to nonunique results when there are considerable symmetries in the distribution of objects to object types. But the claim that binary-valued quality spaces frequently exhibit symmetric distributions is unwarranted: One would only expect such distributions to arise in special situations, such as, for example, when objects are distributed among object types according to a uniform probability distribution.

The problem of finding a partition $\Pi'$ that minimizes $f(\Pi')$ for a given domain of objects is computationally difficult (NP-hard) (Aloise, Deshpande, Hansen, & Popat, 2009), with runtime $O(n^{mk+1})$, where $m$ is the number of properties (i.e., the dimensionality of the quality space) and $n$ is the number of objects (Inaba et al., 1994, p. 335). In attempting to find a partition $\Pi'$ that minimizes $f(\Pi')$, we used a computationally efficient method, known as “Lloyd’s algorithm,” with runtime $O(nmki)$, where $i$ is the number of iterations of the update step for the algorithm (Hartigan & Wong, 1979, p. 102). For the simulations reported in Sections 5 and 6, $i$ was 30, and we ran 20 applications of the algorithm with different (randomly selected) initial centroids for each environment. For the simulations described in the following section, which involve only two environments, we set $i$ to 100 and ran 100 applications of the algorithm with different (randomly selected) initial centroids.

The studies presented in the following sections show, among other things, that inheritance inferences based on classes generated by $k$-means clustering tend to be more reliable than inheritance inferences based on classes that are not comparably fitted to the underlying
domain. Our more general suggestion is that inheritance inferences based on classes that consist of the elements of a well-clustered domain will be more reliable than inheritance inferences based on classes that are not comparably fitted to the underlying domain. Given this more general suggestion, we acknowledge that there are many alternatives to $k$-means clustering, and it is reasonable to expect somewhat different results for other clustering methods. That said, we take it that inheritance inferences based on clusters generated by $k$-means tend to be reliable, because $k$-means forms clusters by minimizing the sum of object to prototype distances. So, we would expect similar results for other methods, such as $k$-median and $k$-medoid clustering (Kaufman & Rousseeuw, 1990), that also function by partitioning a domain by minimizing the sum of object to prototype distances. On the other hand, for at least some domains, we would expect less similarity in results for connectivity-based clustering methods that form clusters by connecting objects to their nearest neighbors, and not by minimizing the sum of object to prototype distances.

5. Simulations with the Zoo and Chicago High Schools Data Sets

As a preface to our broader study, we here present simulation results based on a pair of “real world” data sets, namely, the Zoo Data Set and the Chicago High Schools Data Set. The first round of simulations that we report uses the Zoo Data Set, which was created by Richard Forsyth (Dua & Taniskidou, 2017). Forsyth’s data set features a list of 101 animals that are typical inhabitants of a zoo. The data set represents five sorts of vertebrates (20 birds, 5 reptiles, 13 fish, 4 amphibians, and 41 mammals), along with 8 insects, 4 noninsect arthropods, 3 molluscs (clam, slug, and octopus), 1 cnidarian (jellyfish), 1 echinoderm (starfish), and 1 annelida (earth worm). Strictly speaking, the animals are species, but we consider them here as individuals. The animals are described according to 16 binary-valued properties. Each binary-valued property represented in the Zoo Data Set was treated as an atomic property and as an unfitted class. Fitted classes were generated by the application of $k$-means clustering to individuals represented in the data set.

Forsyth assigned each animal to one of seven classes that correspond roughly to standard biological taxonomy. While Forsyth’s partition of the data set into seven classes is reasonable by scientific standards, the most popular criteria for selecting the correct value of $k$ for the application of $k$-means (including the elbow and silhouette method) do not yield an unequivocal answer concerning the correct value of $k$. By these standards, a value somewhere between $k = 4$ and $k = 8$ appears reasonable. In fact, all of the partitions of the Zoo Data Set generated by $k$-means for values of $k$ between 4 and 8 are fairly reasonable from the perspective of biological taxonomy, despite some classifications that are incorrect from this standpoint, for all of these values. The main differences in the partitions are found in the number of classes of mammal represented (between 1 and 3) and in which other relatively dissimilar groups of animals are lumped together (e.g., whether reptiles are lumped together with amphibians). As we will see later, the reliability of inheritance inference with fitted classes is a monotonically increasing function of $k$. Since the reliability of inheritance inference with fitted classes applied to the Zoo Data Set does not differ much for values from $k = 4$ to $k = 8$, we here
present the data for the case where \( k = 4 \), since a smaller setting of \( k \) provides a more impressive illustration of the reliability of inheritance inference with fitted classes. The partition generated by \( k \)-means, when \( k = 4 \), is also easily stated: One class consists of all 41 species of mammal. Another class consists of all 20 species of bird along with one reptile, tortoise. A third class consists of all 13 species of fish along with one reptile, sea snake. The final class is a heterogeneous catchall, consisting of the remaining 25 species.

In order to compare the performance of inheritance inference for the fitted and unfitted classes, we collected data (for each simulation) concerning the relative frequency of tuples not satisfying the \textit{conclusion condition} for an inheritance inference among the tuples that satisfy the \textit{premise conditions}.

The set of tuples satisfying the \textit{premise conditions} for “regular” inheritance inference is specified as follows:

\[
P = \{ \langle \varphi, \text{SC}, C \rangle : \varphi \text{ is typical among } C \text{ and SC is a subclass of } C \}.
\]

The set of tuples \textit{not} satisfying the \textit{conclusion condition} for “regular” inheritance inference is specified as follows:

\[
\bar{K} = \{ \langle \varphi, \text{SC}, C \rangle : \varphi \text{ is not typical among SC} \}.
\]

For the first approach to base class suitability, each atomic property counts as a class (i.e., \( C = \chi \) and \( \chi \) is an atomic property). For the second approach, the set of classes consists of the elements of the partition generated by \( k \)-means clustering. For both approaches, we require that \( \text{SC} \) be identical to \( C \cap \gamma \), where \( \gamma \) is an atomic property. We also require that \( \varphi \neq \chi, \varphi \neq \gamma, \) and \( \chi \neq \gamma \), thereby ignoring trivial deductive inferences in computing error rates.

The error rate for “regular” inheritance inference is specified by the following quotient, with the value of the quotient depending on which base class suitability criterion is employed:

\[
|\bar{K} \cap P| / |P|
\]

In addition to comparing the performance of the two approaches to base class suitability for regular inheritance inference, we compared their performance for inheritance inference in the case of exceptional subclasses. In this case, the set of tuples satisfying the premise conditions for inheritance inference is specified as follows:

\[
P_E = \{ \langle \varphi, \psi, \text{SC}, C \rangle : \varphi \text{ and } \psi \text{ are typical among } C, \text{SC is a subclass of } C, \text{ and } \psi \text{ is not typical among SC} \}.
\]

The set of tuples \textit{not} satisfying the conclusion condition is:

\[
\bar{K}_E = \{ \langle \varphi, \psi, \text{SC}, C \rangle : \varphi \text{ is not typical among SC} \}.
\]
The error rate for inheritance inference in the case of exceptional subclasses is then given by the following quotient:

$$\frac{|K_E \cap P_E|}{|P_E|}$$

Once again, for the first approach to base class suitability, each atomic property counts as a subclass (i.e., C is a class just in case C = χ and χ is an atomic property). For the second approach, the set of classes consists of the elements of the “dissimilarity minimizing” partition generated by k-means clustering. For both approaches, we require SC = C γ, where γ is an atomic property, and we require that ϕ ≠ ψ, ϕ ≠ χ, ϕ ≠ γ, ψ ≠ χ, ψ ≠ γ, and χ ≠ γ (thereby ignoring trivial deductive inferences in computing error rates).

Given the preceding conventions, Fig. 2 presents the number of typical atomic properties per class, for the Zoo Data Set, for the two approaches to base class suitability, for varied values of r (i.e., varied typicality thresholds), for the case where k = 4. The number of typical properties per class (for the two approaches) is of interest, because the number of typical properties per class is a plausible measure of the number of inheritance inferences that a given approach to base class suitability will license. With this interpretation in mind, the values presented in Fig. 2 (and in further figures, below) distinguish between ϕ being typical of C, when C ≠ ϕ, versus ϕ being typical of C, when C = ϕ, with the former represented by the opaque portion of the red bars, and the latter represented by the transparent portion of the red bars. We call the former “nontrivial” typical properties and the latter “trivial” typical properties.

Fig. 2 illustrates the obvious result that the number of typical properties per class (for both approaches) varies inversely with the magnitude of the typicality threshold r. This result is
expected, since the threshold \( r \) is the standard for typicality, with higher thresholds consisting of a more restrictive standard. The interesting result of Fig. 2 concerns the difference in the number of typical properties for the two approaches to base class suitability, and, particularly, the magnitude of the difference as \( r \) increases. The approach that employs unfitted classes has fewer typical properties per class, for all typicality thresholds except \( r = .5 \) and \( r = .6 \), and very few typical properties per class when the typicality threshold \( r \) is high. For example, when \( r = .99 \), less than two atomic properties are typical of an unfitted class, on average (with one of those being the property \( \varphi \) in the case where \( C = \varphi \)), while three atomic properties are typical of a fitted class, on average.

Fig. 3 presents the error rates for the two sorts of inheritance inference, for the Zoo Data Set, for varied typicality thresholds \( r \), where \( k = 4 \). Note that we did not present error rates for the typicality threshold \( r = .99 \), since the error rate for regular inheritance at this threshold was zero, for both fitted and unfitted classes, and the error rate for inheritance inference for exceptional subclasses, at this threshold, was thus undefined, for both fitted and unfitted classes.

The data presented in Fig. 3 exhibit two patterns that we will also see in the data presented below. The first is that inheritance inference with fitted classes is typically more reliable than inheritance inference with unfitted classes, both for regular inheritance inference and inheritance inference in the case of exceptional subclasses. Note that the slight advantage of unfitted classes over fitted classes for inheritance inference in the case of exceptional subclasses in the
case where \( r = .95 \) is reversed for the Zoo Data Set for \( k = 5, 6, 7, \) and \( 8. \) The second pattern is that inheritance inference in the case of exceptional subclasses is typically less reliable than regular inheritance inference, for both fitted and unfitted classes.

There are at least two factors that explain the superior reliability of inheritance inference with fitted classes. We will return to discuss the second factor, in detail, in Section 8, but it can be observed, immediately, that one of the key reasons that inheritance inference with fitted classes performs better is due to the fact that the frequency of a typical property in a fitted class tends to greatly exceed the relevant typicality threshold \( r \) (for values of \( r \) that are sufficiently less than 1). For example, Fig. 2 shows that, in the case of the Zoo Data Set, fitted classes have, on average, 6.75 atomic properties with frequency at least .5, and 3.75 atomic properties with frequency of at least .95. This implies that, when \( r = .5 \), more than half of the typical atomic properties are “super typical” with a within class frequency of at least .95 (exceeding the typicality threshold \( r \) by .45 or more). This kind of super typicality (which is more common for fitted classes) provides a margin of safety when conducting inheritance inference. Roughly speaking, if the frequency of some characteristic \( \varphi \) among \( C \) is very high (say .95), then it is difficult to select a (moderate-sized) subset of \( C \) in which the frequency \( \varphi \) is not moderately high (say at least .5). For example, if \( C \) has 160 members and 152 of them are \( \varphi \) (i.e., \( \text{freq}(\varphi|C) = .95 \)), then more than 99.98% of the eight member subsets of \( C \) have at least five members that are \( \varphi \).

Some of the other patterns found in Fig. 3 do not generalize to other data sets. For example, Fig. 3 suggests that inheritance inference with unfitted classes performs well if \( r \) is quite high (e.g., when \( r = .95 \)). As an illustration that this is not the case, Figs. 4 and 5 present data
for the Chicago High Schools Data Set, a data set that features prominently in the analysis of cue-based prediction methods provided by Czerlinski, Gigerenzer, and Goldstein (1999). This data set characterizes 57 Chicago public high schools according to 18 properties. The properties are relevant to predicting dropout rates, and most concern the demographic profile of the students at respective schools, relating to ethnicity, parent’s income, past academic performance, and so on.⁵

When applied to the Chicago High Schools Data Set, the most popular criteria for selecting the correct value of \( k \) for the application of \( k \)-means (including the elbow and silhouette method) do not yield an unequivocal answer concerning the correct value of \( k \). Once again, somewhere between \( k = 4 \) and \( k = 8 \) appears reasonable. In this case, we present data for the case where \( k = 6 \). Assuming our goal is to achieve a good balance between minimizing \( k \) and maximizing the reliability of inheritance inference (in the case of fitted classes), then setting \( k \) to 6 is probably the best choice, since improvements in reliability are small for higher values of \( k \), and reliability is markedly impaired for lower values of \( k \), for inheritance inference in the case of exceptional subclasses (with not much difference for regular inheritance inference). The primary difference, for \( k = 5 \), is an increase in the error rate of inheritance inference in the case of exceptional subclasses to roughly 12% and 15%, for the typicality thresholds \( r = .6 \) and \( r = .7 \). For \( k = 4 \), the primary difference is an increase in the error rate of inheritance inference in the case of exceptional subclasses to roughly 22% and 32%, for the typicality thresholds \( r = .9 \) and \( r = .95 \).
Fig. 4 illustrates patterns similar to Fig. 2. Once again, the number of typical properties per class (for both approaches to base class suitability) varies inversely with the magnitude of the typicality threshold $r$. The approach that employs atomic properties as classes has fewer typical properties per class, for all typicality thresholds except $r = .5$ and .6, and very few typical properties per class when $r$ is high. For example, when $r = .95$, we have fewer than $.5$ nontrivial typical atomic properties per unfitted class (on average), and when $r = .99$, we have only $.11$ nontrivial typical atomic properties per unfitted class (on average). On the other hand, $4.67$ atomic properties are typical of fitted classes (on average), when $r = .95$ and when $r = .99$.

Fig. 5 illustrates that setting $r$ to $.95$ is insufficient to ensure reliable inheritance inference based on unfitted classes. It should be noted, however, that for both approaches to base class suitability, there will generally be some typicality threshold $r$ that is sufficient to prevent error, simply by ensuring that $r$ is large enough that a property is typical among a class only if it is typical for all subclasses of the class. This was the case for $r$ to $.99$, which is omitted from Fig. 5.

The data presented in Figs. 2 through 5 demonstrate a clear advantage in the productivity and reliability of inheritance inference with fitted base classes over unfitted base classes. For both data sets, the number of properties that are typical for fitted classes remains high as $r$ increases, which is in strong contrast to inheritance inference with unfitted classes. For both data sets, regular inheritance inference for fitted classes is also highly reliable for all values of $r$, with an error rate that never exceeds $.07$. The error rate for regular inheritance inference based upon unfitted classes is generally higher, pressing into the high teens, for some values of $r$, for both data sets. The error rate of inheritance inference in the case of exceptional subclasses is much greater in the case of unfitted as opposed to fitted classes. Still, the performance of inheritance inference with fitted classes in the case of exceptional subclasses is mixed. For the Chicago High Schools Data Set, the error rate for inheritance inference in the case of exceptional subclasses is below $.1$, for all values of $r$. However, for the Zoo Data Set, the error rate for inheritance inference in the case of exceptional subclasses pushes into the mid-teens in the case where $r = .7$ and in the case where $r = .8$. It is important to note here that increasing the number of fitted classes to 5, 6, 7, or 8 does not alleviate the problem. One possible explanation of this result is that many animals are exhibited in a zoo precisely because they are exotic and/or exceptional. So, it would be a bad idea to place too much weight on the Zoo Data Set in drawing conclusions about the reliability of inheritance inference in the case of exceptional subclasses.

In the following section, we present the results of a more extensive simulation-based study of inheritance inference. The data we present concern the performance of inheritance inference in millions of randomly generated environments.

6. Inheritance inference in randomly generated environments

In the preceding section, we presented some results regarding the reliability of inheritance inference in reasoning with real-world data. In the present section, we present data regarding
the reliability of inheritance inference in reasoning about randomly generated environments. A key advantage of simulations using randomly generated environments is that one can freely vary parameters for the environments, in order to discover which factors are relevant to the reliability of inheritance inference (and to what extent).

The primary ingredient of each of the randomly generated environments was a randomly selected population of objects. For practical purposes, populations are represented as probability distributions over exhaustive combinations of atomic properties. To simplify terminology, we refer to exhaustive combinations of atomic properties as “object types.” For example, if there are exactly two atomic properties ϕ and ψ, then there are four object types, namely: ϕ∩ψ, ϕ∩¬ψ, ¬ϕ∩ψ, and ¬ϕ∩¬ψ. Within an individual simulation, probabilities are assigned to object types as follows: Given n atomic properties, ϕ₁, ..., ϕₙ, we generate the following $\sum_{k=1}^{n} 2^{k-1}$ independently variable (conditional) probabilities via a uniform probability distribution on [0, 1]: $P(\phi_1)$, $P(\phi_2|\phi_1)$, $P(\phi_2|¬\phi_1)$, $P(\phi_3|\phi_1\cap\phi_2)$, $P(\phi_3|\phi_1\cap¬\phi_2)$, $P(\phi_3|¬\phi_1\cap\phi_2)$, $P(\phi_3|¬\phi_1\cap¬\phi_2)$, $P(\phi_1\cap¬\phi_2\cap¬\phi_3)$, $P(\phi_1\cap¬\phi_1\cap¬\phi_2\cap¬\phi_3)$, $P(\phi_4|\phi_1\cap¬\phi_2\cap¬\phi_3)$, $P(\phi_4|¬\phi_1\cap¬\phi_2\cap¬\phi_3)$, $P(\phi_4|¬\phi_1\cap¬\phi_2\cap¬\phi_3)$, and so on. Given these probabilities, the probability of each object type is equal to a product of n independently variable (conditional) probabilities. For example, $P(\phi_1\cap¬\phi_2\cap¬\phi_3\cap¬\phi_4) = P(\phi_1) \times P(¬\phi_2|\phi_1) \times P(¬\phi_3|\phi_1\cap¬\phi_2) \times P(¬\phi_4|\phi_1\cap¬\phi_2\cap¬\phi_3)$ (where $P(¬\phi_2|\phi_1) = 1 - P(\phi_2|\phi_1)$ and $P(¬\phi_4|\phi_1\cap¬\phi_2\cap¬\phi_3) = 1 - P(\phi_4|\phi_1\cap¬\phi_2\cap¬\phi_3)$).

The preceding method of generating probability distributions was chosen because it generates a greater variety of probability distributions in comparison to obvious alternatives. For example, one alternative would be to assign each object type a weight chosen by a uniform distribution on [0, 1], and then assign each object type a probability by normalizing its weight. This alternative method tends to generate distributions that differ by only a small amount from a uniform probability distribution over the set of object types.⁶

Figs. 6 and 7 present data for random environments for varied values of the typicality threshold r. Each bar of each figure represents mean values for 1,000,000 property distributions (i.e., environments), generated as described above for the case of eight atomic properties, and thus $2^8 = 256$ object types.⁷ The environments presented in Figs. 6 and 7 differ from the ones described in the preceding section in featuring a smaller number of atomic properties, but a much larger number of exemplified object types. Indeed, although the numbers of atomic properties reported in the Zoo and Chicago High School (namely, 16 and 18) permit a large number of object types (i.e., $2^{16} = 65,536$ and $2^{18} = 262,144$), the data sets feature only 101 and 57 objects.

For the purpose of applying k-means clustering, k was set to 8, for the simulations presented in Figs. 6 and 7. For randomly generated environments with eight atomic properties, the graph of mean within cluster variance as a function of k exhibits a shallow curve. The shape of the curve to the right of k = 8 is very close to linear, suggesting that setting k to 8 is not unreasonable for these environments.

Fig. 6 illustrates the unsurprising fact that the number of typical properties per class (for both approaches to base class suitability) varies inversely with the magnitude of the typicality bound r. As with Figs. 2 and 4, the interesting result presented in Fig. 6 is the large difference in the number of typical properties per class for the two approaches. For all typicality bounds,
the number of typical properties for fitted classes exceeds the number of nontrivial typical properties for unfitted classes, with large differences for higher thresholds. As a reference point for our expectations concerning how many properties will be typical per class, recall that the environments described in Fig. 6 feature eight atomic properties. Now, since the data in Fig. 6 only concern the number of atomic properties that are typical of classes, we would expect the values presented in the figure to double, if we also counted the number of negated atomic properties that are typical of classes. Notice, then, that, in the case where \( r = .5 \), the mean number of typical atomic properties per fitted class is (approximately) 4. So, were we to consider both atomic properties and their negations, we would see (in the case where \( r = .5 \)) that for every fitted class, and every atomic property \( \varphi \), either \( \varphi \) or not \( \varphi \) is typical for the class. As we raise the typicality threshold \( r \), the number of typical properties decreases. However, at the point where \( r = .99 \), we are still left with about .84 typical atomic properties per fitted class, on average. If we also counted the negations of atomic properties, we would expect this value to double to 1.68. In contrast, in the case where \( r = .99 \), we have only .014 nontrivial typical atomic properties per unfitted class, on average.

Fig. 7 describes variations in the reliability of regular inheritance inference, and inheritance inference in the case of exceptional subclasses, as a function of variations in \( r \). Here, there is not much variation in the reliability of inheritance with fitted classes: For fitted classes, inheritance inference is pretty reliable, with error rates below 10%, for all values of \( r \). Inheritance inference for exceptional subclasses is less reliable, but still reliable, with error rates that are also below 10%, for all values of \( r \). In contrast, the picture for inheritance inference with unfitted classes is not pretty: Regular inheritance is quite
Fig. 7. Error rate as a function of typicality threshold $r$.

*The error rate for a given type of inheritance inference will be undefined, if there is no tuple satisfying the premise conditions for the type. Since the error rates for one or both types of inheritance inference are undefined for some simulated environments, the error rates reported are the mean values for environments in which a respective error rate is defined.

unreliable, with error rates that are greater than 20%, for all values of $r$. In the case of exceptional subclasses, the error rate is substantially worse, increasing from an error rate of 21.5%, when $r = .5$, to reach a peak error rate near 45%—not much better than guessing—for $r = .9$. If one favors higher typicality thresholds (e.g., $r = .8$ or greater), then the results of Figs. 6 and 7 show that inheritance inference with unfitted classes is an analogue to the Catskill’s restaurant derided in Woody Allen’s *Annie Hall*: Bad food (Fig. 7) and small portions too (Fig. 6).

Figs. 8 and 9 illustrate the impact of the number of fitted classes, $k$, on the fruitfulness and reliability of inheritance inference with fitted classes. Within the figures, each column represents the mean values for 1,000,000 randomly generated environments. For all environments, the number of atomic properties was 8, and the typicality threshold $r$ was .9.

Before examining Figs. 8 and 9, note that the behavior of inheritance inference with unfitted classes is independent of the number of fitted classes. So any variation in the behavior of inheritance inference for unfitted classes is due to random variations in the randomly generated environments.
Fig. 8. Number of typical properties as a function of the number of cluster classes $k$.

Fig. 9. Error rate as a function of number of cluster classes $k$. 
Fig. 8 shows the expected result that the number of atomic properties typical for a fitted class increases as the number of fitted classes increases. As one can see, the rate of increase is (more or less) linear with exponential increases in \( k \), which means that increases in the number of typical properties as a function of \( k \) is (more or less) logarithmic (with base 2). The preceding makes sense, since each time one doubles the number of fitted classes from \( k \) to \( 2k \), one grants oneself an increase in the number of classes sufficient to divide each cell of one’s previous partition into a pair of cells that differ with respect to one property that was neither typical nor atypical among the original cell, which thereby increases the mean number of typical properties per fitted class by .5. In fact, we suspect that the number of typical properties within the classes generated by \( k \)-means clustering applied to a space of binary-values is largely independent of the dimensionality of the space. As an explanation for this suspicion, notice the following: If a space is characterized by \( n \) properties, \( 2^n \) exemplified object types, and \( k \leq 2^n \), then there is a partition of the space into \( k \) classes such that there are \( \lceil \log_2 k \rceil \) properties such that each class is completely homogeneous with respect to those properties (i.e., for each class, \( C \), and property, \( \varphi \), we have \( \text{freq}(\varphi|C) \in \{0, 1\} \)). Within the simulation results presented in Fig. 8, such a partition is generated (obviously), in the case where \( k = 256 \), where each of the \( k \) classes is completely homogeneous with respect to all eight atomic properties, resulting in exactly four typical (positive) atomic properties per class, on average.

Fig. 9 illustrates that the reliability of inheritance inference with fitted classes increases as a function of the number of fitted classes. Obviously, one can insure highly reliable inheritance inference (not to mention a large number of typical properties per class) by setting \( k \) to a very high value. Presumably, maintaining an efficient cognitive economy militates against adopting too many class categories (i.e., a value of \( k \) that is too large) (Douven & Gärdenfors, 2020, p. 318). The data presented in Fig. 9 suggest that one can maintain high degrees of reliability, without sacrificing too much cognitive efficiency. Indeed, using eight fitted classes is normally sufficient for high reliability in a domain featuring 256 types of object. In fact, even four classes are sufficient for relatively good performance, with error rates at about 10% for both regular inheritance inference and inheritance in the case of exceptional subclasses.

The simulations described in the present section feature atomic property distributions that are characterized by moderate to high degrees of entropy. The mean (normalized) entropy level for the simulations described in the present section was about .72, with a standard deviation of about .057. In the following section, we consider environments spanning the full range of entropy levels, from extremely low to extremely high. These studies permit for a more inclusive evaluation of the difference between inheritance inferences with fitted versus unfitted base classes.

7. Randomly generated environments with varied entropy levels

Entropy is a standard measure of the degree of uniformity of a probability distribution. Where \( P \) is a probability distribution over \( n \) states (namely, states \( s_1 \) through \( s_n \)), the entropy
of $P$ is defined as:

$$e(P) = -\sum_{i=1}^{n} P(s_i) \times \log P(s_i)$$

In fact, different logarithm bases yield different entropy measures. We assume that the log is base 2, since we are concerned with distributions over binary-valued properties. Because the range of possible entropy values may vary according to $n$, we present all results in terms of normalized entropy (which yields entropy values in $[0, 1]$):

$$\eta(P) = -\sum_{i=1}^{n} \frac{P(s_i) \times \log_2 P(s_i)}{\log_2 n}$$

In order to generate property distributions with specific entropy levels, we adapted the method of generating distributions of object types that was described in the preceding section. As the first step in generating a property distribution with a (normalized) entropy level $\delta \pm .01$, we generated a property distribution exactly as described in the preceding section. Thereafter, the distribution was subject to small random perturbations until a distribution with entropy in $[\delta - .01, \delta + .01]$ was reached.$^9$

The data presented in Fig. 10 show that the number of typical properties per class varies inversely as a function of entropy, that is, higher entropy correlates with lower numbers of typical properties.$^{10}$ This is expected, both in the case of fitted and unfitted classes. In the case of unfitted classes, high entropy may absolutely exclude the possibility of nontrivial typical properties. For example, if entropy is exactly one, then every object type will have exactly the
same number of members. So, for every pair of atomic properties, \( \varphi \) and \( \psi \) (with \( \varphi \neq \psi \)), the relative frequency of \( \varphi \) among \( \psi \) will be .5, and there will be no nontrivial typical properties for any unfitted class, assuming \( r > .5 \). Similarly, if entropy is close to one, then every object type will have nearly the same number of members. So, for every pair of properties, \( \varphi \) and \( \psi \) (with \( \varphi \neq \psi \)), the relative frequency \( \varphi \) among \( \psi \) will be close to .5, and there will be no nontrivial typical properties for any unfitted class, assuming \( r \) is somewhat greater than .5. On the other hand, if entropy is zero, then there will be exactly one exemplified object type, which means that all objects in the domain have exactly the same properties. For an environment characterized by eight atomic properties (such as the environments presented in Fig. 10), the number of typical atomic properties per unfitted class will depend upon which object type is exemplified. On average, a zero entropy environment and a single exemplified object type will result in 3.5 nontrivial typical atomic properties per unfitted class. We observe slightly fewer nontrivial typical atomic properties per unfitted class for the near zero entropy distributions reported in the left most column of Fig. 10.

For all entropy levels greater than zero, the number of typical properties per fitted class exceeds the number of typical properties per unfitted class. In cases where entropy is .4 or less, the expected number of typical atomic properties per fitted class is between 3.5 and 4. If we also counted instances where the negation of an atomic property is typical of a class, then this number would double to values between 7 and 8 (of the eight atomic properties). This means that in relatively low entropy environments, objects are concentrated within a small subset of the set of all object types, so that the vast majority of objects will be very near the position of some (optimally positioned) centroid. Similarly, the vast majority of objects will be very near most of the other objects that are “captured” by the same centroid. Topologically speaking, centroids in a low entropy environment will be positioned near those vertices of the property space in which objects are concentrated, or very near the edges of the space that connect vertices in which objects are concentrated. As entropy increases, objects become less concentrated within a small number of object types, and it becomes more difficult to place centroids in such a way that the vast majority of objects are very near some centroid. Yet, even when entropy is high, centroids are placed near those (hyper)planes of the quality space that connect those object types that are most frequent (Fig. 1), with the result that fitted classes are still characterized by typical properties (unlike unfitted classes). So, even in an entirely unordered environment with maximal entropy, it is possible to form categories in such a way that their instances are highly similar to each other.

Fig. 11 shows that inheritance inference with fitted classes is reliable across all entropy levels, including both regular inheritance inference and inheritance inference in the case of exceptional subclasses. The figure shows that the error rate for regular inheritance inference tends to increase as a function of entropy, beginning with no errors at very low entropy, reaching a peak value of about .074 at high entropy, with exceptional behavior near maximum entropy. The error rate for inheritance inference in the case of exceptional subclasses exhibits less variation, falling in the range .051–.087 across the range of entropy levels, excluding very low and high entropy levels, where no exceptional subclasses were observed. In general, inheritance inference in the case of exceptional subclasses was less reliable than regular inheritance inference, though the difference is generally not great.
The reliability of inheritance inference with unfitted classes was appreciably lower than for fitted classes. For unfitted classes, inheritance inference in the case of exceptional subclasses was very unreliable across all entropy levels in which exceptional subclasses were observed. On the other hand, regular inheritance was reliable when entropy was low, increasing precipitously with entropy, up to the point where we no longer have nontrivial typical properties for unfitted classes.

8. The relevance of base class uniformity to the reliability of inheritance inference

The simulation results presented above show that inheritance inference tends to be more reliable when it is based upon fitted rather than unfitted base classes. The difference in the reliability of inheritance inference for the two sorts of classes derives from several factors. We mentioned one factor earlier, namely: The frequency of a typical atomic property among a fitted class tends to exceed the typicality threshold by a greater margin, thereby providing a greater margin of safety when conducting inheritance inference. There is another important factor that accounts for the difference in the reliability of inheritance inference for the two sorts of classes. This factor concerns the different impact of disuniformity in the distribution of object types within base classes of the two sorts.

The impact of disuniformity in the distribution of object types within a base class is much greater for unfitted classes. In the case of unfitted classes, disuniformity among a base class correlates with the existence of typical atomic properties for the class, while also correlating
with the unreliability of inheritance inference based upon the class. This yields a sort of Catch-22 for inheritance inference with unfitted base classes: Such inferences are more likely to be possible only when they are less reliable.

It is straightforward to see that disuniformity is a necessary condition for typical atomic properties among unfitted classes: An unfitted class has a typical atomic property only if there is disuniformity in the distribution of the object types included in the class. Indeed, an unfitted class C (C = ϕ, for some atom ϕ) includes all and only those object types consisting of C conjoined with literals of atoms, with the consequence that for all atoms ψ: if ψ ≠ C, then freq(ψ|C) > .5 only if there is disuniformity in the distribution of objects among the object types included in C. For example, if we have three atoms, ϕ, ψ1, and ψ2, then C = ϕ consists of the object types ϕ∩ψ1∩ψ2, ϕ∩ψ1∩¬ψ2, ϕ∩¬ψ1∩ψ2, and ϕ∩¬ψ1∩¬ψ2, and freq(ψ1|C) > .5 only if freq(ψ1∩ψ2)+freq(ψ1∩¬ψ2) > freq(¬ψ1∩ψ2)+freq(¬ψ1∩¬ψ2). More generally, it is reasonable to expect a positive correlation between disuniformity in the distribution of object types within an unfitted class and the number of typical atomic properties in the class. For example, we obtain such a correlation, if the sizes of the object types included in unfitted classes consist of a sequence of exchangeable random variables.

Disuniformity among an unfitted class also positively correlates with unreliable inheritance inference. First of all, if the distribution of object types among an unfitted class, C, is completely uniform, then for any pair of atomic properties, ϕ and ψ, we have freq(ϕ|C) = freq(ϕ|C∩ψ), with an approximation of the preceding identity, if the distribution of object types included in C approximates uniformity. So, we should expect highly uniform unfitted base classes to permit reliable inheritance inferences. On the other hand, higher levels of disuniformity tend to make inheritance inference with unfitted base classes less reliable. Indeed, the typicality of a property ϕ among an unfitted class invariably derives from the fact that the relative frequency of objects among some object types that are characterized by ϕ being substantially greater than the relative frequency of objects among other object types that are not characterized by ϕ. For this reason, it is relatively “easy” for a subclass of an unfitted class to omit one or more of the (more frequent) object types that are characterized by ϕ. Omitting such types frequently results in the atypicality of the subclass with respect to ϕ.

Unlike unfitted classes, a fitted class may have typical properties even in the case of a uniform distribution among the object types included in the class. For example, in a case where there are three atoms (ϕ1, ϕ2, and ϕ3), a (nonempty) fitted class C may consist of two object types ϕ1∩ϕ2∩ϕ3 and ϕ1∩ϕ2∩¬ϕ3 (i.e., C = (ϕ1∩ϕ2∩ϕ3)∪(ϕ1∩ϕ2∩¬ϕ3)), with the consequence that freq(ϕ1|C) = freq(ϕ2|C) = 1 regardless of the relative sizes of ϕ1∩ϕ2∩ϕ3 and ϕ1∩ϕ2∩¬ϕ3. Although it is plausible to think that there is some negative correlation between the disuniformity of a fitted class and the number of typical properties the class has, we doubt that there is a strong connection, which is confirmed by the results that we will present in a moment. As with unfitted classes, we expect a correlation between the uniformity of a fitted class and the reliability of inheritance inference with the class. However, in the case of fitted classes (unlike unfitted ones), complete uniformity is not a sufficient condition for the reliability of inheritance inference. More generally, the data presented below suggest that uniformity is less important for the reliability of inheritance inference with fitted base classes, in comparison to unfitted ones.
In an attempt to illustrate the relevance of base class (dis)uniformity to the reliability of inheritance inference, we conducted a further run of simulations where we controlled for the entropy of the base class, or to be more precise we collected data on the relationship between error rate and base class entropy. The entropy of a base class stands as a quantitative measure of its degree of uniformity. We also controlled for the fact that the frequency of a typical atomic property among fitted classes tends to exceed the typicality threshold by a greater margin than for unfitted classes. In order to control for this factor, we set the typicality threshold to $r = .8$, and only considered inheritance inferences for a base class $C$ and a property $\varphi$, if $\text{freq}(\varphi|C) \in [.89, .91]$.

Fig. 12 presents data for 10,000,000 environments, generated in the manner described in Section 5 (with no perturbation of entropy, as in Section 6). The figure presents the error rates for inheritance inferences (with $r = .8$ and $\text{freq}(\varphi|C) \in [.89, .91]$), as a function of base class entropy.

The right-most columns of Fig. 12 (which present the error rates for base classes across all entropy levels) show that inheritance inference with fitted base classes is still more reliable than inheritance inference with unfitted base classes, when one controls the degree to which a typical property exceeds the typicality threshold. It is also evident that the error rate for all four types of inheritance inference is a decreasing (roughly linear) function of base class entropy. We can also see that there are big differences in the slope of this function, with the steepest slope for inheritance inference with exceptional unfitted base classes, and the shallowest slope
Fig. 13. The distribution of base classes according to entropy compared to the distribution of inferences according to base class entropy.

for regular inheritance inference with fitted base classes. These slopes reflect the relative advantage conferred by a high entropy base class, which is greater for unfitted classes. For example, while the error rate for regular inheritance inference is almost four times as great for unfitted versus fitted classes when the entropy of the base class is very low (i.e., in (.1, .2]), the error rate for unfitted classes is actually smaller than for fitted classes (.005 vs. .02) when the entropy of the base class is very high (i.e., in (.8, .9]).

The crucial ingredient for the explanation of why inheritance inference is less reliable for unfitted classes concerns the Catch-22 mentioned above: Such inferences are more likely to be possible only when they are less reliable. Fig. 13 complements the reliability information presented in Fig. 12, and presents data comparing (1) the distribution of base classes according to entropy, with (2) the distribution of inferences according to base class entropy. The yellow transparent bars represent the percentage of unfitted classes with respective entropy levels, which we compare to (nontransparent) yellow bars representing the percentage of inferences based upon unfitted classes with respective entropy levels. Similarly, the green transparent bars represent the percentage of fitted classes with respective entropy levels, which we compare to (nontransparent) green bars representing the percentage of inferences based upon fitted classes with respective entropy levels. The data for fitted classes are unremarkable: The probability that a base class has a typical property, and thereby supports inheritance inferences, is largely independent of its entropy. This is illustrated by the similarity in the size
of corresponding green and transparent green bars. On the other hand, the data for unfitted classes nicely illustrate the claim that disuniformity is conducive to typical atomic properties within unfitted classes. This is illustrated by the difference in the size of corresponding yellow and transparent yellow bars. For example, while about 4% of the unfitted base classes had an entropy level in (.8, .9], these classes accounted for less than .05% of the inheritance inferences with unfitted base classes (illustrating the above mentioned Catch-22, since these inferences had an error rate of about .5%, Fig. 12). Similarly, while more than 50% of the unfitted base classes had an entropy level in (.7, .8], these classes accounted for less than 23% of the inheritance inferences with unfitted base classes (illustrating the above mentioned Catch-22, since these inferences had an error rate of less than 5%, Fig. 12). Conversely, while about 8% of the unfitted base classes had an entropy level in (.5, .6], these classes accounted for about 22% of the inheritance inferences with unfitted base classes (illustrating the above mentioned Catch-22, since these inferences had an error rate of about 19%, Fig. 12).

Even when we control for base class entropy and the margin by which an inferred typical property exceeds the typicality threshold, inheritance inference with unfitted classes is far less reliable in the case of exceptional subclasses. This difference derives from the fact that typicality in the case of unfitted classes derives from disuniformity in the distribution of objects among the object types included in the class. Due to the dependence of typicality on disuniformity, the atypicality of a subclass, in one respect, frequently indicates that the subclass omits the frequent object types that are responsible for the typicalities that characterize the overarching class, which makes it not unlikely that the subclass will be atypical in further respects. The situation is, of course, much different in the case of fitted classes. Unlike an unfitted class, a fitted class may feature typical properties even if the distribution of objects among the object types included in the class is uniform. So, learning that a subclass is atypical in some respect does not provide a strong reason for thinking that the subclass omits the object types that are responsible for the typicalities that characterize the overarching class. However, disuniformity in the distribution of objects among the object types within fitted classes is a possible source of typicality, and so learning that a subclass of a fitted class is atypical in some respect provides modest incremental confirmation for the proposition that the subclass is atypical in further respects. But if we are in a position to know that the entropy of a fitted base class is high, the possibility that typicality is a result of disuniformity receives incremental disconfirmation. Indeed, in cases where the entropy of a fitted class is very high, the data presented in Fig. 12 show that inheritance inference is more reliable in the case of exceptional subclasses! In these cases, where the effects of disuniformity are minimal, learning that a subclass is exceptional with respect to a property provides incremental confirmation for the conclusion that it is not exceptional with respect to another property ϕ. This effect derives from the general “pressure” on the objects in fitted classes to minimize their overall dissimilarity to other objects that are assigned to the same class, and grounds a nondeductive version of disjunctive syllogism: If both ψ and ϕ are typical among a base class, and a given subclass of the base class is exceptional with respect to ψ, this makes it all the more likely that ϕ is typical among the subclass, since too much atypicality would contradict the very grounds by which the object types were assigned to the base class in the first place.
9. Conclusion

The simulation studies presented in the preceding sections show that inheritance inference with fitted classes is more reliable than inheritance inference with unfitted classes, both for regular inheritance inference and inheritance inference in the case of exceptional subclasses. Furthermore, the superiority of inheritance inference with fitted classes holds (i) independently of the value, \( r \), of the typicality threshold, (ii) independently of the number, \( k \), of fitted classes that is used to partition a domain, and (iii) independently of the entropy level of the environment. Of particular interest, the study shows that, in the case of exceptional subclasses, inheritance inference based upon atomic properties is horrendously unreliable, whereas inheritance inference based upon fitted classes is quite reliable. The results of the study, thereby, address a long-running debate in the field of nonmonotonic reasoning, concerning whether inheritance inference is reasonable in the case of exceptional subclasses: The matter depends upon the criteria used in selecting acceptable classes! If one makes the reasonable assumption that there is close relationship between fitted classes and the type of concepts to which prototype theories of concepts properly apply, the results of our study are also consonant with prototype theories of concepts. In keeping with the results presented here, we conjecture that human subjects are sensitive to the difference between fitted and unfitted classes and are more credulous concerning inheritance inference based on fitted classes. If human subjects are more credulous concerning inheritance inference based on fitted classes, then that would provide some corroboration of prototype theories of concepts inasmuch as those theories predict that composed concepts inherit the prototypical properties of composing concepts by default.

Notes

1. We here present Cautious Monotony according to our intended application. More abstractly, Cautious Monotony takes the following form: If \( C \models \varphi \) & \( C \models \psi \), then \( C, \psi \models \varphi \).
2. So reasonable instances of inheritance inference also outstrip Rational Monotonicity, as conformity to Rational Monotonicity (i.e., if \( C \models \varphi \) & not \( C \models \neg \psi \), then \( C, \psi \models \varphi \)) would require that it not be the case that the relative size of SC among C be small.
3. Excellent visualizations of \( k \)-means clustering applied to the Iris and “mouse” data sets can be found on the Wikipedia page for \( k \)-means clustering (Wikipedia contributors 2019, June 3).
4. The data set includes 15 binary-valued properties: has hair, has feathers, lays eggs, produces milk, flies, is aquatic, is a predator, has teeth, has a backbone, breathes under water, is venomous, has fins, has a tail, is domesticated, is larger than a house cat. The data set also included one nonbinary variable “number of legs.” In order to arrive at 16 binary-valued properties, we converted the variable “number of legs” to the variable “has legs.”
5. The properties are the result of classifying the schools as at least average (1) or below average (0) with respect to 18 quantitative variables, namely: enrollment, attendance rate, graduation rate, parental involvement rate, percent limited English students,
percent low-income students, class size, students per teacher, percent white students, percent black students, percent Hispanic students, percent Asian students, percent minority teachers, average ACT score, average reading score, average math score, average science score, average social science score, and average writing score.

6. For eight atomic properties, the mean normalized entropy for such distributions is about .965 with a standard deviation of about .003. In the case of eight atomic properties, the mean normalized entropy for the method we employed is about .72, with a standard deviation of about .057.

7. Our sampling of 1,000,000 property distributions is unnecessarily large for almost all of the values and parameters settings considered, with all of the mean values presented in Figs. 6 through 11 reportable with the qualification ± .001 (based on the recommendations of Bindel & Goodman, 2009). Large samples were drawn for the sake of parameter settings that tend to produce distributions that have undefined error rates for inheritance inference in the case of exceptional subclasses (e.g., for unfitted classes when entropy is high). The largest standard errors for the mean values presented in Figs. 6–11 (rounded at the first nonzero digit) were .0009, .0004, .0004, .0003, .001, and .0003.

8. We include the case where \( k = 1 \) as a point of reference, although it is inappropriate to describe the relevant class, in this case, as “fitted.”

9. More precisely: We first generated a distribution over object types exactly as described in the preceding section. It was then checked whether the entropy of the distribution was in \([\delta - .01, \delta + .01]\). If it was, then the search for an acceptable distribution halted. If the entropy of the generated distribution was not in \([\delta - .01, \delta + .01]\), then a further step, called “revision,” was executed. Within the revision step, two object types were selected at random, and the probabilities assigned to the two object types were redistributed in order to maximize the closeness of the entropy of the revised distribution to \( \delta \). More precisely, the algorithm attempted a revision that resulted in the largest possible swing in entropy in the direction of \( \delta \). If such a revision would result in an entropy level that overshot the interval \([\delta - .01, \delta + .01]\), then a revision that was one half as great was attempted, and then a revision that was one third as great, etc., until a revision that did not overshoot the interval \([\delta - .01, \delta + .01]\) was found.

10. This correlation is not strict for fitted classes, as illustrated by Fig. 10, in the case where the entropy of the underlying distribution is close to 1. Centroid placement in this case has the following pattern: Three atomic properties, \( \varphi_1, \varphi_2, \varphi_3 \), are selected. Centroid placement corresponds to the eight vertices of the subspace for these properties, with values near .5 for the other five properties. This results in complete homogeneity with respect to \( \varphi_1, \varphi_2, \) and \( \varphi_3 \), and more or less complete heterogeneity for the remaining five atomic properties, resulting in exactly 1.5 typical atomic properties per fitted class (for all typicality thresholds). For slightly lower entropy distributions, levels of homogeneity are typically less extreme, resulting in fewer typical properties for high typicality thresholds (e.g., for \( r = .9 \), as presented in Fig. 10).

11. Fig. 13 does not represent the fact that fitted base classes allowed more than four times as many inferences as unfitted base classes.
References

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: Reidel.

Adams, E. W. (1986). On the logic of high probability. *Journal of Philosophical Logic, 15*, 255–279.

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning, 75*, 245–249.

Bacchus, F. (1990). *Representing and reasoning with probabilistic knowledge*. Cambridge, MA: MIT Press.

Benferhat, S., Cayrol, C., Dubois, D., Lang, J., & Prade, H. (1993). Inconsistency management and prioritized syntax-based entailment. In IJCAI’93, 640–647.

Bindel, D., & Goodman, J. (2009). *Principles of scientific computing*.

Bonevac, D. (2003). *Deduction: Introductory symbolic logic* (2nd ed). Malden, MA: Blackwell.

Brewka, G. (2009). *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambridge University Press.

Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2009). Why stereotypes don’t even make good defaults. *Cognition, 103*, 1–22.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer & P. M. Todd the ABC Research Group. (Eds.), *Simple heuristics that make us smart* (pp. 97–118). Oxford: Oxford University Press.

Douven, I., & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language, 35*, 313–334.

Dua, D., & Taniskidou, E. K. (2017). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science.

Elio, R., & Pelletier, F. J. (1994). On relevance in nonmonotonic reasoning: Some empirical studies. In R. Greiner & D. Subramanian (Eds.), *Relevance: AAAI 1994 Fall Symposium Series* (pp. 64–67). Palo Alto, CA: AAAI Press.

Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science, 3*, 85–100.

Gabbay, D. (1984). Theoretical foundations for non-monotonic reasoning in expert systems. In K. R. Apt (Ed.), *Logics and models for concurrent systems* (pp. 439–458). Berlin: Springer.

Gagné, C., & Spalding, T. (2014). Subcategorisation, not uncertainty, drives the modification effect. *Language, Cognition and Neuroscience, 29*, 1283–1294.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Goldszmidt, M., Morris, P., & Pearl, J. (1990). A maximum entropy approach to nonmonotonic reasoning. In Proceedings of American Association for Artificial Intelligence Conference, 646–652.

Goldszmidt, M., & Pearl, J. (1991). System-Z+: A formalism for reasoning with variable-strength defaults. In AAAI’91 Proceedings of the 9th National Conference on Artificial Intelligence - Volume 1, 399–404.

Goldszmidt, M., & Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision and causal modeling. *Artificial Intelligence, 84*, 57–112.

Halpern, J. (2009). *Reasoning about uncertainty*. Cambridge, MA: MIT Press.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition, 15*, 55–71.

Hampton, J. A., Passanisi, A., & Jönsson, M. L. (2011). The modifier effect and property mutability. *Journal of Memory and Language, 64*, 233–248.

Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C*, 28, 100–108.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning — Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Hawthorne, J., & Makinson, D. (2007). The quantitative /qualitative watershed for rules of uncertain inference. *Studia Logica, 86*, 247–297.

Horty, J. (2007). Defaults with priorities. *Journal of Philosophical Logic, 36*, 367–413.
Horty, J., Thomason, R., & Touretzky, D. (1990). A sceptical theory of inheritance in nonmonotonic semantic networks. *Artificial Intelligence*, 42, 311–348.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2, 283–304.

Inaba, M., Katoh, N., & Imai, H. (1994). Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. In Proceedings of 10th ACM Symposium on Computational Geometry, 332–339.

Jäger, G. (2007). The evolution of convex categories. *Linguistics and Philosophy*, 30, 551–564.

Jönsson, M. L., & Hampton, J. A. (2008). On prototypes as defaults (comment on Connolly, Fodor, Gleitman and Gleitman, 2007). *Cognition*, 106, 913–923.

Jönsson, M. L., & Hampton, J. A. (2012). The modifier effect in within-category induction: Default inheritance in complex noun phrases. *Language and Cognitive Processes*, 27, 90–116.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. Hoboken, NJ: Wiley.

Kern-Isberner, G. (2001). *Conditionals in nonmonotonic reasoning and belief revision – Considering conditionals as Agents*. No. 2087 in LNCS. Springer.

Koons, R. (2017). *Defeasible reasoning*. Stanford Encyclopedia of Philosophy.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44, 167–207.

Kyburg, H., & Teng, C. (2001). *Uncertain inference*. Cambridge: Cambridge University Press.

Lehmann, D. (1989). What does a knowledge base entail? Proceedings of 1st International Conference on Knowledge Representation, 212–222.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 128–137 [originally a technical report, Bell Laboratories, 1957].

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297. University of California Press.

McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence*, 25, 41–72.

McGrew, T. (2001). Direct inference and the problem of induction. *The Monist*, 84, 153–174

Ng, M., Li, M., Huang, Z., & He, Z. (2007). On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 503–507.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Santa Mateo, CA: Morgan Kaufmann.

Pearl, J. (1990). System Z. In: *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, 121–135. Santa Mateo, California.

Pelletier, F. J., & Elio, R. (1997). What should default reasoning be, by default? *Computational Intelligence*, 17, 165–187.

Pollock, J. (1990). *Nomic probability and the foundations of induction*. Oxford: Oxford University Press.

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, 36, 27–47.

Poole, D. (1994). Default logic. In D. M. Gabbay, C. J. Hogger, & J. A. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming* (Vol. 3, pp. 189–215). New York: Oxford University Press.

Reichenbach, H. (1949). *A theory of probability*. Berkeley, CA: Berkeley University Press.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13, 81–132.

Reiter, R. (1987). Nonmonotonic reasoning. *Annual Review of Computer Science*, 2, 147–186.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104, 192–233.

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

Salem, S., Naouali, S., & Sallami, M. (2017). Clustering categorical data using the k-means algorithm and the attribute’s relative frequency. In: World Academy of Science, 19th International Conference on Machine Learning and Applications, 691–696.
Schurz, G. (1997). Probabilistic default reasoning based on relevance and irrelevance assumptions. In D. M. Gabbay, R. Kruse, A. Nonnengart, & H. J. Ohlbach (Eds.), *Qualitative and quantitative practical reasoning (LNAI 1244)* (pp. 536–553). Berlin: Springer.

Schurz, G. (1998). Probabilistic semantics for Delgrande’s conditional logic and a counterexample to his default logic. *Artificial Intelligence, 102*, 81–95.

Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint. *Synthese, 146*, 37–51.

Schurz, G., & Thorn, P. D. (2012). Reward versus risk in uncertain inference: Theorems and simulations. *Review of Symbolic Logic, 4*, 574–612.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review, 63*, 129–138.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12*, 485–527.

Strasser, C., & Antonelli, G. A. (2019). *Non-monotonic logic*. Stanford Encyclopedia of Philosophy.

Strößner, C., & Schurz, G. (2020). The role of reasoning and pragmatics in the modifier effect. *Cognitive Science, 44*, 1–35.

Tan, S., & Pearl, J. (1995). Specificity and inheritance in default reasoning. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI’95) - Volume 2, 1480–1486.

Thorn, P. D. (2012). Two problems of direct inference. *Erkenntnis, 76*, 299–318.

Thorn, P. D. (2014). Defeasible conditionalization. *Journal of Philosophical Logic, 43*, 283–302.

Thorn, P. D. (2015). The inheritance of defaults in the case of exceptional subclasses. *The Reasoner, 9*, 93.

Thorn, P. D. (2017). On the preference for more specific reference classes. *Synthese, 194*, 2025–2051.

Thorn, P. D., Eichhorn, C., Kern-Isberner, G., & Schurz, G. (2015). Qualitative probabilistic inference with default inheritance. In C. Beierle, G. Kern-Isberner, M. Ragni, & F. Stolzenburg (Eds.), *Proceedings of the KI 2015 Workshop on Formal and Cognitive Reasoning* (pp. 16–28).

Thorn, P. D., & Schurz, G. (2014). A utility based evaluation of logico-probabilistic systems. *Studia Logica, 102*, 867–890.

Thorn, P. D., & Schurz, G. (2016). Qualitative probabilistic inference under varied entropy levels. *Journal of Applied Logic, 19*, 87–101.

Touretzky, D., Horty, J., & Thomason, R. (1987). A clash of intuitions: The current state of monotonic multiple inheritance systems. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence, 476–482.

Wikipedia Contributors. K-means clustering. Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title=K-means_clustering&oldid=900149325

Wobcke, W. (1995). Belief revision, conditional logic and nonmonotonic reasoning. *Notre Dame Journal of Formal Logic, 36*, 55–102.

Venn, J. (1866). *The logic of chance*. New York: Chelsea.