Journal of Mathematical Psychology 96 (2020) 102357

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Category-based induction in conceptual spaces

Matías Osta-Vélez [a,b,c,*], Peter Gärdenfors [d]

[a] Munich Center for Mathematical Philosophy (LMU), Ludwigstr. 31, D-80539 Munich, Germany
[b] IHPST (Université Paris 1/CNRS), 13 rue du Four, 76006, Paris, France
[c] Instituto de Filosofía (Universidad de la Repú blica), Av. Uruguay 1695, 11200, Montevideo, Uruguay
[d] Philosophy and Cognitive Science Department, Lund University, Helgonavägen 3, Lund, Sweden

ARTICLE INFO

Article history:
Received 10 December 2019
Received in revised form 9 March 2020
Accepted 28 March 2020
Available online xxxx

ABSTRACT

Category-based induction is an inferential mechanism that uses knowledge of conceptual relations in order to estimate how likely is for a property to be projected from one category to another. During the last decades, psychologists have identified several features of this mechanism, and they have proposed different formal models of it. In this article; we propose a new mathematical model for category-based induction based on distances on conceptual spaces. We show how this model can predict most of the properties of this kind of reasoning while providing a solid theoretical foundation for it. We also show that it subsumes some of the previous models proposed in the literature and that it generates new predictions.

1. Introduction

Reasoning and concept representation are two central issues in philosophy and cognitive psychology. Surprisingly enough, they have traditionally been understood as independent research topics and they rarely intersect in the literature. One possible explanation for this is that reasoning studies have been dominated by a *logicist* approach that conceives inference as a purely formal-syntactic processes (i.e., non-semantical), that builds on some set of domain-general and topic-neutral rules of inference. On that view, lexical concepts are seen as inferentially inert, that is, not playing any crucial role in the very process of inference and reasoning.

This view, summarized in Inhelder and Piaget's claim that "[human] reasoning is nothing more than the propositional calculus itself". (Inhelder & Piaget, 1958, p. 305), led cognitive psychologists to consider deductive reasoning as the paradigm of rational inference. In deductive reasoning, drawing a conclusion from a set of premises does not require the agent to exploit semantic knowledge about the premises. Deduction is informationally conservative, the information in the conclusion is implicit in the premises, and the agent is only required to grasp the logical form of arguments and to know how to use logical constants to infer.

However, deductive reasoning is hardly the only or the most important mechanism used in everyday reasoning. Navigating our environment requires us to cope with constant uncertainty and partial information (Oaksford & Chater, 1998). Truth-preservation is not the central concern, but the type of inferential processes needed are those allowing us to make risky predictions about new stimuli exploiting background knowledge and expectations.

Induction is one of these processes. Induction and deduction deal with semantic information in very different ways (Johnson-Laird, 2010). The conclusion of an inductive inference adds semantic information that is not present in the premises. If we infer that all ravens are black from a set of premises about individual ravens, we are inferring under uncertainty, and adding information that is not present in the premises. In this sense, inductive reasoning is not "formal", that is, merely based on the syntactic structure of the premises (see Thagard, 1988, pp. 27–29).

Understanding induction, therefore, requires explaining how background knowledge is exploited in reasoning. One way of doing this (following Gärdenfors, 2000; Thagard, 1984) is to see inference as an activity that is not strictly linguistic-based, but that combines information codified at the symbolic-propositional level, with information encoded at the conceptual level (see Gärdenfors & Stephens, 2018).

During the last decades, psychologists have been studying a kind of cognitive phenomenon that is directly related to this last point. In the pioneering article "Inductive judgments about natural categories", Rips (1975) analyzed a particular type of inductive reasoning that exploits information about individual categories (and about relations among categories) for estimating the probability of property projection among them. For instance, the inference "Dogs have sesamoid bones; thus wolves have sesamoid bones" relies on the conceptual similarities among the categories DOG and WOLF, and not on the logical form of the

* Correspondence to: Ludwig-Maximilians-Universitat, Munchen Center for Mathematical Philosophy (First Floor) Ludwigstr. 31, D-80539 Munich, Germany.
E-mail address: matiasosta@fhuce.edu.uy (M. Osta-Vélez).

https://doi.org/10.1016/j.jmp.2020.102357
0022-2496/

argument or some other propositionally codified property. Such processes, called *category-based inferences* (CBI), are fundamental to our cognitive lives. On the one hand, they are crucial for dealing with uncertainty: they allow us to reason about some unknown input X by exploiting information stored in our conceptual system about things that resemble X. On the other hand, as Feeney (2017, p. 167) observes, they are a clear example of how concepts make our cognition efficient.

Understanding how CBI works, and especially which properties of our conceptual systems it exploits, can shed light on the general problem about the role of concepts in inferences. In this article, we discuss the general features of CBI, and we propose conceptual spaces (Gärdenfors, 2000, 2014) as an explanatory framework. Conceptual spaces is a theory about the structure and organization of conceptual knowledge. We will propose a model of CBI that builds on *distances* in conceptual spaces and show that the model can explain most of the empirical results concerning CBI.

The article is organized as follows. Section 2 presents a basic taxonomy of category-based inductions and reviews the central phenomena associated with them. Section 3 introduces some of the theoretical and technical aspects of conceptual spaces. Section 4 presents our model and explains how the theory of conceptual spaces provides a natural way of modeling the central properties of CBI based on the capacity of the theory for representing similarity and typicality relations among categories. Section 5 compares our model to some of the previous explanations, and Section 6 briefly considers some methodological aspects of the approach we defend.

2. Category-based induction

2.1. The general structure of category-based inferences

Rips' (1975) seminal paper intended to understand the strategies that agents use for reasoning under uncertainty about property projection among biological kinds (such as HAWK, BIRD, EAGLE, etc.). He showed that agents exploit structural properties of categories for estimating the plausibility of property projection. In particular, Rips saw that similarity among categories and the degree of typicality of premise-categories were guiding principles of this kind of reasoning.

Most studies on CBI follow Rips' analysis (for example, Carey, 1985; Heit, 2000; Osherson, Smith, Wilkie, López, & Shafir, 1990; Sloman, 1993). They all assume that inductive reasoning is a process that exploits information at the conceptual level, and not at the propositional level. From a methodological perspective, these studies analyze the inferential dispositions of cognitive agents by studying how people judge the strength of different types of (categorical) inductive arguments. Various phenomena concerning CBI have been identified during the last decades (see Feeney, 2017; Hayes, Heit, & Swendsen, 2010 for reviews). We next explain a basic taxonomy of CBI in order to present the phenomena that characterize CBI.

Category-based inferences are structured as arguments with one or more premises of the form 'X are S' (for example, 'Dogs have sesamoid bones,' and 'Bears love onions') and one conclusion of the same type. We sometimes abbreviate an inference of the form 'X have property S; thus Y have property S' as X → Y. One argument for this abbreviation is that, in almost all studies, subjects typically have little or no knowledge about the property S and therefore it does not influence the strength of the argument.

CBI can be classified in two major ways: according to their number of premises, and according to whether the conclusion is at the same conceptual level as the premises or in some superordinate category. When the premise(s) and conclusion categories are at the same conceptual level, the argument is called *specific*; when the argument involves a generalization (a "jump" to a superordinate conceptual level), then it is called *general.* For instance, arguments with the (categorical) form ROBIN → CROW or TABLE → CHAIR are specific, while arguments that generalize to a superordinate category, like ROBIN → BIRD, ROBIN → ANIMAL or TABLE → FURNITURE, are general. Both specific and general arguments can be composed of one or multiple premises (see Table 1). .

Table 1
Basic taxonomy of category-based inferences.

	Single premises		Multiple premises	
Specific	(1)	*Foxes have property S* Wolves have property S	(2)	Penguins have property S Pigeons have property S *Ostriches have property S* Sparrows have property S
General	(3)	*Robins have property S* Birds have property S	(4)	Polar bears have property S *Grizzly bears have property S* Bears have property S

In what follows, we will review the main properties of CBI as described by the empirical literature. The idea is that these phenomena say something about what kind of categorical or conceptual relations people exploit when judging category-based inductive arguments.

2.2. Premise–conclusion similarity

The primary categorical relation guiding category-based inferences is similarity. Similarity has been considered as a crucial criterion for induction since at least Hume (1999, p. 20). Quine famously argued (1969, 1974) that similarity might be a fundamental psychological principle in a wide range of cognitive phenomena, like learning, concept formation, and reasoning. In psychology, the notion of similarity has proved to be fruitful since the 70 s. Since the pioneering work of Shepard (1974) and Tversky (1977), formal models of similarity were developed for explaining concept formation, categorization, and even induction. And since Rosch's (1973) work on prototypes, similarity was developed as the central criterion for explaining category structure.

Not surprisingly, the empirical literature has shown that the most robust criterion used in CBI is similarity among categories (Carey, 1985; López, Gelman, Gutheil, & Smith, 1992; Osherson et al., 1990; Rips, 1975). This can be formulated as that our expectations regarding property projection among two categories X and Y is a positive function of their similarity. For instance, arguments like "Ostriches are S, then emus are S" are generally seen as stronger than arguments like "Ostriches are S, then blue jays are S", since *sim*(OSTRICH, EMU) > *sim*(OSTRICH, BLUEJAY), where *sim*(X,Y) denotes a measure of the similarity between the categories X and Y.

2.3. Typicality

Similarity, as a criterion for category-based inferences, can only be used for categories at the same conceptual level; but it is not useful in arguments that generalize a property from the category premise to the category of the conclusion. In those cases, *typicality* is what guides categorical inferences.[1]

Roughly put, the *typicality effect* is the finding that individuals respond more quickly to typical examples of a category than they do to cases that are considered atypical. For instance, when asked to name a bird, an individual is much more likely to respond

[1] Typicality is deeply related to similarity (see Rips, 1989).

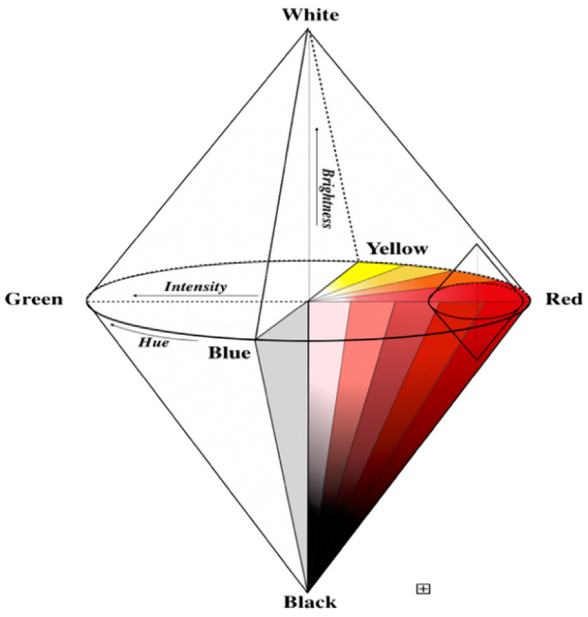

Fig. 1. RED is a subregion of the color space.

with 'robin' than with 'penguin'. The idea was proposed and tested by Rosch (1973), and it suggests that conceptual structures (especially natural kinds) are articulated around *prototypes.* In fact, most categories seem to have a *graded structure* (see Barsalou, 1987; Decock & Douven, 2014), which means that different members of the category are perceived with different degrees of typicality. For instance, cows are generally seen as very typical representatives of the MAMMAL category, while mice are moderately typical, and whales are highly atypical members.

Typicality plays a crucial role in CBI.[2] The most robust effect found in the empirical literature is that expectations of property projection are a positive function of premise-typicality. For instance, the inference "Robins have enzyme E; thus ostriches have enzyme E" is often judged as stronger than "Penguins have enzyme E; thus ostriches have enzyme E". This is due to the fact that robins are prototypical birds and as such they better represent the category than penguins (which are atypical). To a lesser extent, conclusion-typicality also seems to be a factor in category-based inferences. Hampton and Cannon (2004) have shown that arguments with prototypical conclusion-categories (like CHICKEN → ROBIN) are judged as stronger than arguments with non-typical conclusion categories (like CHICKEN → VULTURE).

Furthermore, the typicality effect produces what is called *asymmetry*, that is, the fact that switching the categories from the premises and conclusion often changes the expectations of property projection, according to the degree of typicality of the category in the premise(s). For instance, arguments like "Cows have enzyme E; thus otters have enzyme E" is considered stronger than arguments like "Otters have enzyme E; thus cows have enzyme E" since cows are more typical mammals than otters.

2.4. Conclusion homogeneity and premise diversity

Another important aspect is that agents assume a common superordinate category of the premises when making inferences or judging the strength of this kind of argument. Sometimes this superordinate category appears explicitly in the conclusion (as in general arguments); some other times, it is just considered implicitly. For instance, consider the arguments in Fig. 1. In (1), the implicit superordinate category is MAMMAL, while in (2) it is BIRD. In (3) and (4), the superordinate category appears explicitly in the conclusion. Four important phenomena related to such an evoked superordinate category have been studied in the empirical literature: *homogeneity, monotonicity, nonmonotonicity, and premise diversity*.

(i) Homogeneity refers to the idea that the more abstract and less homogeneous the category in the conclusion is, the weaker the argument. For instance, arguments like "Robins are S and blue jays are S; thus all birds are S" are judged as stronger than "Robins are S, and blue jays are S; thus all animals are S". This is not surprising at all. As we said before, when evaluating arguments or making inferences that involve generalizations, we deal with different degrees of uncertainty. The more abstract the category in the conclusion, the more information we need from the premises to cover it. Hence category-based inductions with abstract categories (like ANIMAL or LIVING BEING) involve higher degrees of uncertainty and are more difficult to cover by the information from the premises.

Studies of categorization (especially in the prototype tradition) provide some insight into this phenomenon. Categories may have different degrees of generality (e.g. DOG, MAMMAL, ANIMAL, LIVING THING), and these degrees are related to the computational cost of using them in categorization. Categories with an intermediate level of specificity are preferred in terms of cognitive efficiency. These are called *basic-level categories* (DOG instead of MAMMAL; CHAIR instead of FURNITURE), and studies have shown that they are central for carrying out several cognitive tasks (Mervis & Rosch, 1981). Inductive inference seems to follow the same principle. We have a preferred level of induction (Sloman & Lagnado, 2005, p. 106) that coincides with basic-level categories.

A possible way of explaining this is by referring to similarity and typicality as the two main criteria for using categories. Basic level categories are more homogeneous. As such, it is easier for us to apply criterion of similarity among their members. Abstract categories are more diverse and less homogeneous, so comparing their members in terms of similarity is more complex (for instance, the category ANIMAL include highly dissimilar subcategories, such as ELEPHANT and STARFISH). Along the same line, basic categories have clear prototypes, while it is complicated for us to construct prototypes for abstract categories (see Ungerer & Schmid, 2006, Ch. 2 for an introductory explanation). In this sense, typicality, considered as a criterion for using categories, is stronger in basic-level categories than in abstract ones.

(ii) Monotonicity refers to the fact that the addition of premises, as long as their categories are included in the evoked superordinate category, strengthen the argument (Osherson et al., 1990). For instance, an argument of the form (ROBIN & HAWK) → BIRD is weaker that an argument of the form (ROBIN & HAWK & PIGEON) → BIRD. However, if we add to the premises a category that is not from the evoked superordinate category, then the argument becomes weaker. This is called *nonmonotonicity* (iii). For instance, an argument with the categories (PEACOCK & CROW) → BIRD (or (PEACOCK & CROW) → PIGEON) is stronger than an argument that goes from (PEACOCK & CROW & RABBIT) → ANIMAL (or (PEACOCK & CROW & RABBIT) → PIGEON).

(iv) Finally, there is the *diversity phenomenon* (Feeney & Heit, 2007; Osherson et al., 1990). Empirical studies have shown that having diverse categories tends to promote expectations regarding property projections. For instance, arguments like "Horses have an ulnar artery and seals have an ulnar artery; thus all mammals have an ulnar artery" are considered as stronger than the argument "Horses have an ulnar artery, and cows have an ulnar artery; thus all mammals have an ulnar artery". The less similar

[2] For a discussion on the role of prototypicality in reasoning in general, see Cherniak (1984).

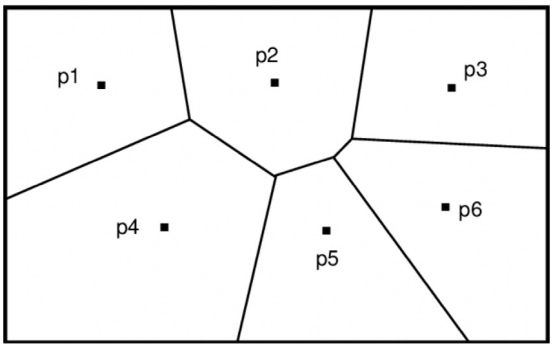

Fig. 2. Example of a Voronoi tessellation from Gärdenfors (2000).

the categories in the premises are, the stronger the argument tends to be.

An interesting way of understanding this phenomenon builds on the idea of "category coverage" (Osherson et al., 1990). As we mentioned before, when performing or evaluating categorical inductions, we take as a reference (implicitly or explicitly, according to whether we deal with a specific or general argument) some superordinate category that includes all the categories in the premises. The strength of the argument will depend, to some extent, upon how the categories in the premises *cover* this superordinate category. For instance, similar categories like HORSE and COW have less coverage of the superordinate category than dissimilar categories like HORSE and SEAL. In this sense, coverage can be described in terms of similarity. We will discuss this idea further in Section 5.

Sloman (1993, pp. 253–254) pointed out that diversity has a limit: if highly dissimilar categories are used in the premises, this can weaken the argument instead of making it stronger. For instance, the argument "German shepherds have sesamoid bones and elephants have sesamoid bones; thus moles have sesamoid bones" seems stronger than "German shepherds have sesamoid bones, and blue whales have sesamoid bones; thus moles have sesamoid bones". This indicates that in arguments that include highly atypical categories in some premise (BLUE WHALE is a highly atypical mammal), then diversity becomes negative regarding argument strength.

In the following section, we present the basic framework for conceptual spaces. This will allow us to explain later how we propose to use this framework as a model of the phenomena presented above and other aspects of CBI.

3. Conceptual spaces

3.1. Defining conceptual spaces, properties and concepts

Conceptual Spaces (CS) (Gärdenfors, 2000, 2014) is a research program in cognitive science for modeling several cognitive phenomena involving concepts and conceptual structures (semantic processing, learning, reasoning, categorization, concept formation, etc.). Unlike the dominant computational tradition in philosophy and cognitive science, CS does not assume that language (or some language-like structure like Fodor's "language of thought") is the fundamental representational system supporting high-level cognition. Instead, CS builds upon the fundamental hypothesis that there exists an intermediate representational system that encodes semantic information in spatial structure.

CS is an heir to the geometrical models of conceptual representations inaugurated by Shepard (1987) in psychology, and a development of the notions of "quality spaces" in Quine (1960), "attributes spaces" in Carnap (1971), and "logical spaces" in Stalnaker (1981). Just like in the other geometrical models in psychology, the fundamental idea behind conceptual spaces is that concept formation and representation take place in some psychological space — in which similarity can be represented in terms of distances determined from some metric (see Eliot, 1987).

CS builds on two fundamental notions: *quality dimensions* and *domains*. Quality dimensions are the building blocks of conceptual spaces. They are geometrical structures able to represent some "quality" of objects, for example, *brightness*, *height*, *time*, *weight*, *pitch*. Each of these qualities of stimuli can be represented by a particular geometrical structure (see Gärdenfors, 2000). For instance, *weight* can be represented by a line isomorphic to the non-negative real numbers.

Dimensions can be *integral* or *separable*. Dimensions are integral when you cannot assign to an object a value in one dimension without assigning another value in another dimension (Maddox, 1992). For instance, it is not possible to attribute a value to *pitch* of a tone without attributing one to *loudness*. When quality dimensions are not integral, they are called separable.

A set of integral dimensions that are separable from all other dimensions is called a *domain*. The classic example of a domain is the color spindle. It is composed of three integral dimensions *hue, saturation*, and *brightness*. The geometrical representation of hue is the color circle. Saturation or intensity is represented as an interval of the real line, while brightness varies from white to black and is thus a linear dimension with endpoints. These three integral dimensions together, one with a circular structure and two with a linear structure, make up the *color space* (see Fig. 2).[3]

Domains serve to represent different qualities of objects through their geometrical and topological properties. A central such property is distance, that allows us to represent similarity among different properties: The closer they are in space, the more similar they are.[4] For instance, within the color space, predicates like RED, BLUE or ORANGE correspond to regions of the domain. And the relationships among them correspond to their position in the color spindle. For instance, the distances in the color domain allow us to represent why ORANGE and RED are more similar than RED and GREEN.

The domains of a conceptual space are related in various ways since the properties of the objects modeled in the spaces co-vary. For example, in the fruit domain, the ripeness and color dimensions co-vary, as well as size and weight. These co-variations are crucial for inferential procedures that exploit conceptual properties.

A *conceptual space* is a collection of one or more domains with a distance function (metric) that represents properties, concepts, and their (similarity) relationships. The distance function can vary, but the most common one is the Euclidean distance function, which makes conceptual spaces Euclidean spaces (Johannesson, 2003). Objects are seen as instances of concepts and are mapped into points of the space, and concepts are represented as *regions*. Similarity among concepts and objects can then be easily estimated since it is a monotonically decreasing function of their distance within the space (Shepard, 1987).

A concept is generally defined as a region of some conceptual space. But within single domains, concepts are called *properties*. According to Gärdenfors (2000), natural properties are characterized by the following criterion:

[3] It is worthwhile to mention that the figures in this article have only an illustrative purpose. They do not come from actual data about the conceptual spaces they are supposed to represent.

[4] It should be noted that not all spaces have a metric. For example, some dimensions only have an ordering structure.

Criterion P: A *natural property* is a convex region in some domain.

Convexity exploits the geometric properties of quality dimensions. A region is convex when for every pair of points x and y in the region, all points between them are also in the region (see Fig. 2). In this way, criterion P assumes that the notion of betweenness is meaningful for the relevant domain.

Gärdenfors (2000) conjectures that color terms, being natural properties, have to respect the structure of the conceptual space in which they are grounded across different languages. That means that it would not be possible for any language to have one single word for the extension of two colors like RED and GREEN since they are disjoint in the color conceptual space. This conjecture has been confirmed for 110 different languages by Jäger (2010).

Within this framework, concepts are represented as regions of some set of interweaved domains. Gärdenfors (2000) defines concepts according to the following criterion:

Criterion C: A *concept* is represented as a set of convex regions in a number of domains, together with information about how the regions in different domains are correlated.

Fruit categories are good examples of natural concepts. For instance, the concept APPLE comprises regions in domains like color, shape, taste, texture, nutrition, and smell. The APPLE concept has a strong positive correlation between sweetness in the taste domain and sugar content in the nutrition domain, and a weaker positive correlation between redness and sweetness.

In this article we assume that for each of the domains that belong to a concept, there exists a distances measure. However, it is also assumed that these measures can be weighted together to create an overall measure for the space. As we will see in Section 3.3 this weighting is, in general, context dependent. In most of the experiments on categorical induction, however, the properties studied are presented without any specific context, which makes it reasonable to assume that the similarity judgments of the subjects can be used to estimate a common space and a metric that can be used in testing the model we propose.

3.2. Prototypically and conceptual spaces

One important advantage of the conceptual spaces framework is that it can represent prototypes of concepts. In that sense, it fits very well with the prototype theory of categorization (Gärdenfors, 2000; Lakoff, 1987; Mervis & Rosch, 1981; Rosch, 1975, 1978).

Criteria P and C allow for a natural way of representing the prototype effects. Within convex regions, one can take some specific point (or set of points) as the prototype of a category.[5] As a result, and using the built-in metric of the space, one can get a measure of typicality of any member of the category by estimating its distance with the prototype. For example, focal colors are generally considered in cognitive science and linguistics as prototypes of the color space (Douven & Gärdenfors, 2019; Rosch, 1971).

Assuming the prototypical structure of concepts does not require that there is an actual object that represents the prototype. Describing concepts as convex regions of a conceptual space allows us to represent the complete space of possible objects that would fall under the concept. In particular, a prototype may be represented as a partial vector, where only the values of the dimensions that are most relevant for the concept have been determined. For example, the general shape of the prototypical bird would be included in the vector, but its color or age presumably would not.

It is possible to argue in the converse direction, too, and show that, if prototype theory is adopted, then the representation of concepts as convex regions is to be expected. To obtain a prototypically structured conceptual space, we start from a set of prototypes p1, …, pn of the categories to be represented (for example, different bird species). These are the central points in the categories they represent. If it is then assumed that for every point p in the space one can measure the distance from p to each of the pi's and stipulate that p belongs to the same category as the *closest* prototype pi, then it can be shown that this rule will generate a partitioning of the space that consists of convex areas (convexity is here defined in terms of an assumed distance measure). This is the so-called *Voronoi tessellation*, a two-dimensional example of which is illustrated in Fig. 3. Thus, assuming that a metric is defined on the subspace that is subject to categorization, a set of prototypes will by this method generate a unique partitioning of the subspace into convex regions.

Because of the role that prototypes have within this theoretical framework, typicality is an independent variable. As Gärdenfors shows, this particular spatial configuration of the space has several advantages in terms of the economy of cognitive processing (ibid, pp. 123–126).

3.3. Context, domain salience, and dynamic conceptual spaces

An important phenomenon that any theory of concepts must account for is that psychological similarity is a variable measure that is dependent on the context (Goodman, 1972). In particular, as noticed by Nosofsky (1986), conceptual similarity is modulated by attention to specific domains of the compared concepts. For instance, apples are seen (generally) as more similar to tomatoes than to dates. However, in the context of choosing dessert, in which "sweetness" is a salient feature, it is expected for this similarity judgment to change.

The contexts in which concepts are used are crucial in the modulation of the similarity measure. Context-effects have been extensively studied in the psychological literature (see Goldstone, Medin, & Halberstadt, 1997; Keßler, Raubal, & Janowicz, 2007), and geometrical models of similarity have been often criticized because of their limitations at the moment of accounting for them (Decock & Douven, 2011; Tversky, 1977 for a review). The conceptual spaces model, however, does not suffer these shortcomings (Johannesson, 2003). In this theory, the context-sensitive character of psychological similarity is accounted for in terms of a weighted distance measure. For instance, within the context of a Euclidean metric, the distance measure will include attention-weights w_i that modify the salience of dimension i in the conceptual space

$$d(x, y) = \sqrt{\sum_i w_i(x_i + y_i)^2}$$

When a larger value is given to a weight w_i the conceptual space is magnified along that dimension, which means that dimension i will become more important when determining the similarity between categories (Gärdenfors, 2000, p. 20). As we will see later in the article, this weighted-distance function will have a central role for accounting for the role of context in CBI, and in particular, for dealing with the influence on non-blank properties.

In summary, conceptual spaces, thanks to their particular geometrical and topological structures, allow us to represent similarity and typicality, which are two main properties that characterize concepts and conceptual systems. Modeling CBI by conceptual spaces makes use of these two properties.

[5] It should be noted that this does not necessarily lead to being central in the regions they are assigned.

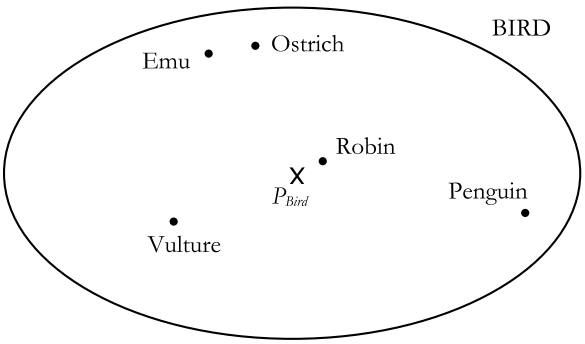

Fig. 3. "Bird space" representing the positions of the different bird categories relative to a prototype.

3.4. Inference and conceptual spaces

As we mentioned earlier, the formalist view of reasoning based on logic and probability has dominated philosophy and cognitive science (Mercier & Sperber, 2017). In general, this approach assumes that reasoning is based on propositions, and can be described by some set of topic-neutral and domain-general rules. As a consequence, concept structures are immediately dismissed as inferentially irrelevant.

However, this view has been criticized because of its psychological implausibility (see Johnson-Laird, 2010; Oaksford & Chater, 1991) and most cognitive scientists assume that concepts play a constitutive role in inferential processes (Carey, 2009; Evans, 1989).

We believe that conceptual spaces can offer insights into the role of conceptual knowledge in reasoning. In this framework, inference is not conceived as a process that takes place (exclusively) at the propositional level, but one that supposes the interaction between the conceptual and the symbolic levels (Gärdenfors, 1997). In particular, inference exploits properties of the conceptual structure, and since conceptual structure is represented geometrically, inference can be understood as exploiting different geometrical properties of concepts.

As a simple example, consider the inference "the car is red; thus the car is not green". This inference is intuitively valid (not formally) for any subject who grasps the basic color concepts. In conceptual spaces, having that concept involves being able to represent an object in the RED region of the color spindle (Fig. 2), something which immediately implies that the object is not located in the other regions of the spindle (GREEN, YELLOW, etc.).

Furthermore, Gärdenfors (2000, 2008) showed that conceptual spaces are useful for modeling the non-monotonic character of rational inference, and inferences associated with the metaphorical use of language. And recently, Schockaert and Prade (2013) used conceptual spaces to model interpolative reasoning. We next turn to use conceptual spaces in our model of CBI.

4. Category-based inferences modeled in conceptual spaces

4.1. Inferences and expectations

CBI has been analyzed from different perspectives, depending on what kind of categorical relation it is assumed to explicate. Class-inclusion (Inhelder & Piaget, 1964), shared features (Sloman, 1998), and similarity (Osherson et al., 1990), have been the most explored ones in the literature. However, reasoning about categories seems to be a complex mechanism that involves combining all these relations and probably other more sophisticated heuristics. It is a challenge to present a model that can account for all of them. In this section, we propose a general model based on conceptual spaces that, among other things, offers a natural (and relatively simple) way of explaining similarity and typicality which are the two main categorical relations that are central to CBI.

For our model, we will talk about *expectations* of property projection among categories instead of argument strength. As Gärdenfors has argued (1991, 1994), expectations play a crucial role in everyday reasoning. For instance, the sentence "John got a new pet" comes associated with a large set of expectations related to the lexical concepts in the sentence. After hearing that sentence, I would expect John's new pet to be prototypical, that is a dog, or a cat, or (less likely) a bird. But if I later hear the sentence "John's new pet is grazing in the garden", I would then discard the aforementioned expectations and infer that John's new pet is probably some pet-size grazing animal. It has been argued that much of nonmonotonic reasoning can be modeled in terms of a semantic framework based on expectations (Gärdenfors, 1994; Gärdenfors & Makinson, 1994). In relation to CBI, the idea is that the agent's inferential dispositions are determined – to a large extent – by her expectations about regularities in the world, which are codified in the agent's background knowledge.

In our modeling of CBI, we use the expression $ExpS(X \rightarrow Y)_Z$ to stand for *the expectation that the property S is projected from category X to category Y,* with Z as the lowest-level superordinate category that contains both X and Y. We will start our analysis by focusing on the simplest case of category-based inference: single premises/specific arguments. For this kind of inductive inference, we want $ExpS(X \rightarrow Y)_Z$ to satisfy the following criteria:

1. It is positively correlated with sim(X, Y).
2. It is positively correlated with sim(X, P_Z), where P_Z is the prototype of Z.
3. It is positively correlated with sim(Y, P_Z)

The rationale for the first condition is that the more similar the categories X and Y are, the more expected will it be that Y has property S if X has it. Regarding condition (2), the intuition is that the more prototypical X is, the more expected it is that another category Y has property S, given that X has it. Condition (3) is motivated by Hampton and Cannon' (2004) conclusion-typicality: the more prototypical Y is the more expected it is that Y has property S, if X has it.

4.2. A simple model

To illustrate the basic idea of our approach with a simple model, let us, for the time being, assume that X and Y are small regions so that we can identify them with points in a conceptual space. Then, given a conceptual space representing the categories X, Y, and Z and a distance function d of the space, we can account for the three conditions above by the following equation:

$$ExpS(X \rightarrow Y)_Z = (d(X, Y) \cdot d(X, P_Z)^a \cdot d(Y, P_Z)^b)^{-1} \quad (1)$$

where a and b are positive constants such that $a > b$. This assumption expresses that premise typicality contributes more to expectations than conclusion-typicality since, according to the literature, the former is a more prevalent phenomenon than the latter. The values of both a and b must be empirically determined from data about CBI judgments. We return to this point in Section 6.

Now, following Shepard's (1987) universal law of generalization, which claims that similarity is an exponentially decreasing function of distance, we can take the logarithm of (1) and obtain:

$$logExpS(X \rightarrow Y)_Z = sim(X, Y) + a \cdot sim(X, P_Z) + b \cdot sim(Y, P_Z) \quad (2)$$

By convention, for any two categories X and Y, $0 \leq sim(X, Y) \leq 1$ and $sim(X, Y) = 1$ if and only if $X = Y$.

Now, Eq. (2) captures the basic idea that for single-premise specific arguments the expectations of property projection among categories are determined by a weighted sum of three factors: premise–conclusion similarity, premise-typicality, and conclusion-typicality.

Eq. (1), applied to a set of prototypes for categories, captures similarity, premise and conclusion typicality and asymmetry effects in CBI. For instance, when considering the sentence "emus have property S", people expect more that ostriches also have property S than that penguins have it. This is due to the similarity effect since *sim*(EMU, OSTRICH) > *sim*(EMU, PENGUIN). If we construct a "bird space" through some set of prototypes, this inequality would be immediately represented by the relative positions in the space of the two pairs ⟨EMU, OSTRICH⟩, and ⟨EMU, PENGUIN⟩ (see Fig. 4). And it can be measured via the distance function of the space. Since *sim*(EMU, OSTRICH) > *sim*(EMU, PENGUIN), if follows from (1) that $ExpS(EMU \rightarrow OSTRICH)_{BIRD} > ExpS(EMU \rightarrow PENGUIN)_{BIRD}$.

As we mentioned, this model also predicts asymmetry and premise and conclusion-typicality. For instance, $ExpS(ROBIN \rightarrow EMU)_{BIRD} > ExpS(EMU \rightarrow ROBIN)_{BIRD}$ since *sim*(ROBIN, P_{BIRD}) > *sim*(EMU, P_{BIRD}) and $a > b$. Regarding conclusion-typicality assume, following the bird space in Fig. 4, that *sim*(OSTRICH, VULTURE) ≈ *sim*(OSTRICH, ROBIN) and that *sim*(OSTRICH, P_{BIRD}) ≈ *sim*(VULTURE, P_{BIRD}). Then $ExpS(OSTRICH \rightarrow ROBIN)_{BIRD} > ExpS(OSTRICH \rightarrow VULTURE)_{BIRD}$ since *sim*(ROBIN, P_{BIRD}) is significantly larger than *sim*(VULTURE, P_{BIRD}).

4.3. A more general model

In general, concepts are represented as *regions* of conceptual spaces, not points. We now turn to a more general model to account for this. Our idea then is that the *volumes* of the regions that represent concepts in some conceptual space (areas in the case of a 2-dimensional space), can be taken as predictors of expectations, that is, argument strength in CBI.[6] The volume of a concept in a conceptual space depends on the metric that is assigned to the space and it is defined in a standard way. The volume depends on the variability of properties ("values") that can be given to an object falling under that concept in each domain. For instance, it is expected that the concept DOG has a larger volume than the concept TIGER, since dogs can be of several different colors, shapes and, sizes, while tigers have less variability in these domains. The immediate consequence of this is that the more heterogeneous the concept is, the larger its volume will be in a conceptual space.

Coming back to expectations, we assume that $ExpS(X \rightarrow Y)_Z$ is positively correlated with the volume V(X) of X and negatively correlated to the volume V(Y) of Y. The positive correlation is due to the fact that the larger V(X), the more it "covers" – or is more *representative* of – the superordinate category Z. For example, $ExpS\ (BEAR \rightarrow WOLF)_{MAMMAL}$ should be larger than $ExpS\ (POLARBEAR \rightarrow WOLF)_{MAMMAL}$ (see Fig. 4).

The negative correlation holds because the smaller the region Y is, the more likely it is for Y to have property S in the inductive argument. If X and Y cover overlapping regions of the space, then the relative sizes of their non-overlapping regions X – Y and Y – X, that is, V(X – Y)/V(Y – X) should be considered. Building on (1), and considering the above ideas we propose the following:

$$ExpS(X \rightarrow Y)_Z = \left(d(P_X, P_Y)^{\frac{V(X-Y)}{V(Y-X)}} \cdot d(P_X, P_Z)^a \cdot d(P_Y, P_Z)^b\right)^{-1} \quad (3)$$

[6] An alternative idea is to introduce explicitly distances between regions as a function of distances between their points (see for example Niiniluoto, 1987). It would be a matter of empirical testing to determine which model would give the best results.

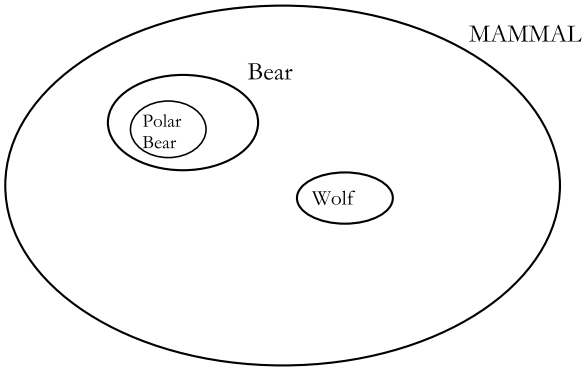

Fig. 4. "Mammal space" representing the difference in volume of BEAR, POLAR BEAR and WOLF.

Again, taking the logarithm and considering the relation between distance and similarity, we obtain)

$$logExpS(X \rightarrow Y)_Z = \frac{V\,(X-Y)}{V\,(Y-X)} sim(P_X, P_Y) + a \cdot sim(P_X, P_Z) + b \cdot sim(P_Y, P_Z) \quad (4)$$

In cases when X and Y are disjoint regions, which are the most typical ones, the quotient reduces to V(X)/V(Y). And in cases when X and Y are represented by small non-overlapping regions, we can take V(X)/V(Y) = 1 and then Eqs. (3) and (4), respectively, will have (1) and (2) as limiting cases.

Just as (3), this new equation predicts premise-similarity, premise and conclusion-typicality, and asymmetry. To see an example of how it works, consider the conclusion-typicality effect. As mentioned in Section 2.3, some experiments show a robust effect of conclusion typicality in CBI (Hampton & Cannon, 2004). For instance, an argument with categories KOALA → GUINEA PIG should be seen as weaker than an argument like KOALA → TIGER, since TIGER is a more typical mammal than GUINEA PIG. Assume, for the sake of argument that *sim*(KOALA, GUINEA PIG) ≈ *sim*(KOALA, TIGER), and that V(GUINEA PIG) ≈ V(TIGER). Then, using (4) we will have that

$$\frac{V\ (\text{KOALA})}{V\ (\text{GUINEA PIG})}.sim\ (P_{KOALA}, P_{\text{GUINEA PIG}}) + a \cdot sim\ (P_{KOALA}, P_{MAMMAL}) + b \cdot sim\ (P_{GUINEAPIG}, P_{MAMMAL}) < \frac{V\ (\text{KOALA})}{V\ (TIGER)} \cdot sim\ (P_{\text{KOALA}}, P_{TIGER}) + a \cdot sim\ (P_{KOALA}, P_{MAMMAL}) + b \cdot sim\ (P_{GUINEAPIG}, P_{MAMMAL})$$

since

$$b \cdot sim\ (P_{GUINEAPIG}, P_{MAMMAL}) < b \cdot sim\ (P_{TIGER}, P_{MAMMAL})\ .$$

Then it follows that

$$ExpS(KOALA \rightarrow GUINEAPIG)_{MAMMAL} < ExpS(KOALA \rightarrow TIGER)_{MAMMAL}$$

(see Fig. 5).[7]

Note that (4) is not defined when Y ⊂ X, since in that case Y – X = ∅. However, in this case the expectation of property projection is maximal, and we can define it to be some arbitrary high

[7] It is possible that a concept with greater volume is less typical than a concept with a smaller volume. For example, FISH may have a greater volume than CAT, but being less typical as a PET. However, in Eq. (3), the expectation value is not only determined by the volume of a concept but also its typicality. So even though FISH may have a larger volume than CAT, the greater typicality of CAT will counterweigh this.

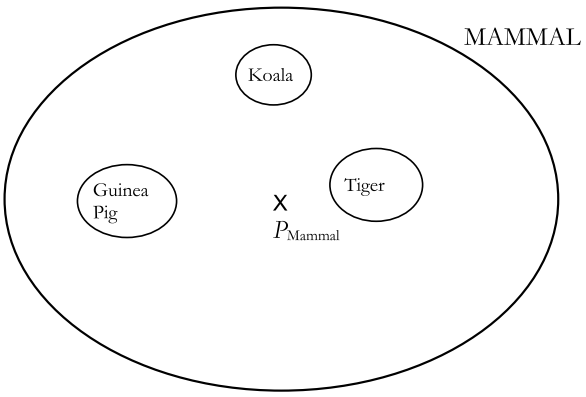

Fig. 5. – Mammal space for categories KOALA, TIGER and GUINEA PIG.

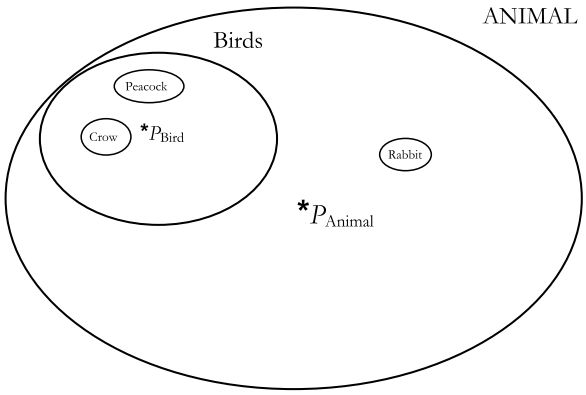

Fig. 6. *Animal space* including the subspace BIRD.

value.[8] For general judgments, for example $ExpS(VULTURE \rightarrow BIRD)_{BIRD}$, we have Y = Z and X ⊂ Y, and since *sim*(BIRD, BIRD) = 1, (4) will consequently take the following form:

$$logExpS(X \rightarrow Z)_Z = a \cdot sim\,(P_X, P_Z) + b \tag{5}$$

This coheres with the idea that single premise/general arguments depend, mostly, on premise-typicality relations, that is, on the idea that agents represent a category with a certain degree of typicality in the context of a more abstract superordinate category. This is not a minor point. In our model, we assume that agents cannot compare categories from different conceptual levels directly in terms of similarity (like comparing COLLIE with MAMMAL). Instead, the categorical relation that works in these cases is typicality, which comes by default for all categories in a conceptual space, given that conceptual spaces are constructed and articulated around prototypes. As we will explain later, we consider that this is an advantage over the two classical models of CBI, which have more difficulties representing typicality relations among categories.

Now, Sloman (1993) observes that subjects exhibit an 'inclusion fallacy' since the argument "Robins have property S; thus birds have property S" is judged to be stronger than "Robins have property S; thus ostriches have property S" despite the fact that ostriches form a subset of birds. Our model can explain this phenomenon. To see how, note that (since V(ROBIN – BIRD) = ∅),

$$logExpS(ROBIN \rightarrow BIRD)_{BIRD} = a \cdot sim\,(P_{ROBIN}, P_{BIRD}) + b,$$

and that

$$logExpS(ROBIN \rightarrow OSTRICH)_{BIRD} = \frac{V\,(ROBIN)}{V\,(OSTRICH)}.sim(P_{ROBIN}, P_{OSTRICH}) + a.sim\,(P_{ROBIN}, P_{BIRD}) + b.sim\,(P_{OSTRICH}, P_{BIRD})\,.$$

Then

$$logExpS(ROBIN \rightarrow BIRD)_{BIRD} > logExpS(ROBIN \rightarrow OSTRICH)_{BIRD}$$

as long as

$$(a \cdot sim\,(P_{ROBIN}, P_{BIRD}) - a.sim\,(P_{ROBIN}, P_{BIRD})) + (b - sim\,(P_{OSTRICH}, P_{BIRD})) > \frac{V\,(ROBIN)}{V\,(OSTRICH)}.sim(P_{ROBIN}, P_{OSTRICH}),$$

which would typically be the case since $sim\,(P_{OSTRICH}, P_{BIRD})$ is small. This shows that our model, unlike the similarity-coverage model (Osherson et al., 1990), also predicts results that are not valid under all conditions, but only under certain circumstances.

Our model can also predict the conclusion-specificity phenomenon (Osherson et al., 1990, p. 187), which says that people tend to judge arguments with more specific categories as stronger than argument with more abstract categories. For instance, an argument like CROW → BIRD will be judged as stronger than an argument of the form CROW → ANIMAL. This is easily explained by our model because the more abstract the conclusion category is, the bigger its volume in the conceptual space, and the further the prototype of this category will be from the prototype of the premise-category. Considering the above example, we have that

$$logExpS(CROW \rightarrow BIRD)_{BIRD} = a \cdot sim\,(P_{CROW}, P_{BIRD}) + b$$

and that

$$logExpS(CROW \rightarrow ANIMAL)_{ANIMAL} = a \cdot sim\,(P_{CROW}, P_{ANIMAL}) + b.$$

Since

$$d\,(P_{CROW}, P_{ANIMAL}) > d\,(P_{CROW}, P_{BIRD})\,,$$

then

$$a.sim\,(P_{CROW}, P_{BIRD}) > a.sim\,(P_{CROW}, P_{ANIMAL})\,.$$

Making

$$a.sim\,(P_{CROW}, P_{BIRD}) > sim\,(P_{CROW}, P_{ANIMAL})\,,$$

it follows that

$$ExpS(CROW \rightarrow BIRD)_{BIRD} > ExpS(CROW \rightarrow ANIMAL)_{ANIMAL}$$

(see Fig. 6).

4.4. Arguments with multiple premises

In single-premise arguments the focus is on the relation between the premise and the conclusion categories; but when dealing with multiple premises, we must also account for premise–premise categorical relations. The main phenomenon to model in these cases is *diversity,* i.e., the more different are the categories in the premises, the stronger the argument. For instance, the argument (i) "Jaguars and leopards have property P; thus otters have property P", is weaker than the argument (ii) "Jaguars and elephants have property P; thus otters have property P". That is because the categories JAGUAR and LEOPARD are similar and they

[8] A reason for this assignment is that if Y is a region that partially overlaps X and then shrinks to become a subset of X, then V(Y – X) will be smaller and smaller, which means that $ExpS\,(X \rightarrow Y)\,z$ will approach infinity. From a psychological perspective, we hypothesize that these cases require agents using an inferential mechanism based on class-inclusion, like property-inheritance. If Y ⊂ X, members of Y inherit all properties of X, thus $ExpS\,(X \rightarrow Y)\,z$ is maximal.

provide less "coverage" of the superordinate category MAMMAL than JAGUAR and ELEPHANT.

The diversity phenomenon suggests that argument strength is a negative function of premise–premise similarity. One possible way of modeling this would be to compute the pairwise similarity of category premises, but this would represent a significant increase in the computational complexity of the process, in particular when we have arguments with more than two premises. Our proposal tries to avoid this by considering all the categories of the premises as part of one large inclusive set. In a sense, the premise categories can be seen as "exemplars" of a more general category. More precisely, we can model n-premises arguments by considering the convex hull of the categories X_1, X_2,..., X_n in the premises. A convex hull of a set S – denoted by C(S) – is the smallest convex region containing all elements in S (see Devadoss & O'Rourke, 2011 for a detailed explanation).

Convex hulls are also convex regions of n-dimensional spaces with the same geometrical properties as regions in conceptual spaces. The size of their volumes is positively correlated to the number of convex regions they include, as well as to the distances among these regions. For instance, in a conceptual space in which all the categories have similar volumes, the volume of the hull of two contiguous regions is going to be smaller than the volume of two non-contiguous regions of the space.[9] This is precisely the kind of property of interest to represent the diversity phenomena. For example, if we consider the argument described above, in an animal space the categories JAGUAR and LEOPARD would be represented by contiguous (or very close) regions in the space, while JAGUAR and ELEPHANT would be far from each other. As a consequence, the volume of C(JAGUAR∪LEOPARD) would be smaller than the volume of C(JAGUAR∪ELEPHANT), and then it would provide less coverage of the MAMMAL category (see Fig. 7).

However, one problem of this approach is that we do not have a "natural" prototype – like P_X in (3) – for the premise anymore. Our solution is to consider an "artificial" prototype P_C, at the centroid of the convex hull $C(X_1 \cup X_2 \cup, \ldots, \cup X_n)$.[10] For convex hulls, we can then reformulate (4) for multiple premises in the following way:

$$logExpS(X_1, \ldots, X_n \rightarrow Y)_Z = \frac{V(C(X_1 \cup X_2, \cup, \ldots, \cup X_n) - Y)}{V(Y - C(X_1 \cup X_2, \ldots, \cup X_n))} \cdot sim(P_C, P_Y) + a \cdot sim(P_C, P_Z) + b \cdot sim(P_Y, P_Z) \quad (6)$$

To see how this formula predicts diversity, consider the example at the beginning of this section. According to (6),

$$ExpS(\text{JAGUAR}, LEOPARD \rightarrow OTTER)_{MAMMAL} = \frac{V\,(C\,(\text{JAGUAR} \cup \text{LEOPARD}))}{V\,(OTTER)} \cdot sim(P_C, P_{\text{OTTER}}) + a.sim(P_C, P_{MAMMAL}) + b.sim(P_{OTTER}, P_{MAMMAL}).$$

This is smaller than

$$ExpS(\text{JAGUAR}, ELEPHANT \rightarrow OTTER)_{MAMMAL}$$

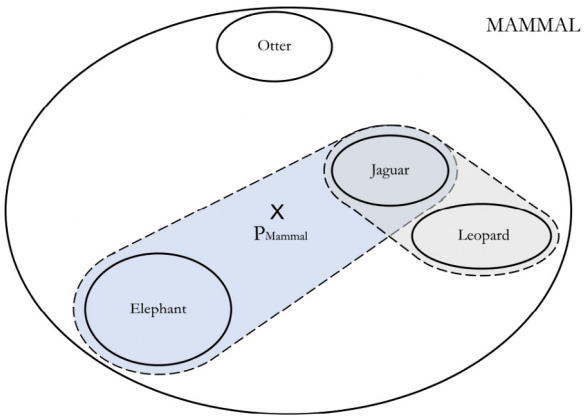

Fig. 7. *Mammal space* illustrating that the volume of (ELEPHANT∪JAGUAR) is larger than the volume of (JAGUAR∪LEOPARD).

$$= \frac{V\,(C\,(\text{JAGUAR} \cup \text{ELEPHANT}))}{V\,(OTTER)} \cdot sim(P_{C*}, P_{OTTER}) + a \cdot sim(P_{C*}, P_{MAMMAL}) + b \cdot sim(P_{OTTER}, P_{MAMMAL}),$$

since V(C(JAGUAR∪ELEPHANT)) > V(C(JAGUAR∪LEOPARD)), which makes

$$\frac{V\,(C\,(\text{JAGUAR} \cup \text{ELEPHANT}))}{V\,(OTTER)} \cdot sim\,(P_{C*}, P_{OTTER}) > \frac{V\,(C\,(\text{JAGUAR} \cup \text{LEOPARD}))}{V\,(OTTER)} \cdot sim\,(P_C, P_{\text{OTTER}}),$$

when $sim\,(P_{C*}, P_{OTTER}) \geq sim\,(P_C, P_{OTTER})$.
The result also follows in the case in which $sim\,(P_{C*}, P_{OTTER}) < sim\,(P_C, P_{OTTER})$, when the difference between $\frac{V(C(\text{JAGUAR} \cup \text{ELEPHANT}))}{V(OTTER)}$ and $\frac{V(C(\text{JAGUAR} \cup \text{LEOPARD}))}{V(otter)}$ is enough to make $\frac{V(C(\text{JAGUAR} \cup \text{ELEPHANT}))}{V(OTTER)} \cdot sim\,(P_{C*}, P_{OTTER}) > \frac{V(C(\text{JAGUAR} \cup \text{LEOPARD}))}{V(OTTER)} \cdot sim\,(P_C, P_{\text{OTTER}})$. Again, this is a conclusion that is not always valid, but depends on the relations between the categories involved.

If Y is a subregion of $C(X_1 \cup X_2 \cup, \ldots, \cup X_n)$, then $Y - C(X_1 \cup X_2 \cup, \ldots, \cup X_n) = \varnothing$, so (6) is undefined. For the same reasons as before, we can set this to some maximal value. For example, if BUZZARD belongs to the convex hull of EAGLE, KITE, and HARRIER, it would follow that $ExpS(EAGLE, KITE, HARRIER \rightarrow BUZZARD)_{BIRD}$ would be maximal. This is the first example of a prediction that emerges from our model. It is an interesting empirical problem, whether this would correspond to the judgment of real subjects. As far as we are aware, this phenomenon has not been tested.

If Eq. (6) is applied to multiple premise *general* arguments then $C(X_1 \cup X_2 \cup, \ldots, \cup X_n) - Y = \varnothing$, since $C(X_1, X_2, \ldots, X_n) \subset Y$. Then, (6) reduces to

$$logExpS(X_1, \ldots, X_n \rightarrow Y)_Z = a \cdot sim\,(P_C, P_Z) + b \cdot sim\,(P_Y, P_Z) \quad (7)$$

Note that when $Y = Z$, $sim\,(P_Y, P_Z) = 1$ and (7) reduces to

$$logExpS(X_1, \ldots, X_n \rightarrow Z)_Z = a \cdot sim\,(P_C, P_Z) + b \quad (7')$$

A problem with this expression is that is does not account for the diversity of X_1, X_2,..., X_n, but only the prototype P_C. One way of solving this problem is to let the constant a depend on the proportion of Z that is covered by X_1, X_2, ..., X_n, that is, $V(C(X_1, X_2, \ldots, X_n))/V(Z)$. However, since empirical evidence for this case seems to be lacking, we will not pursue this theme here.

Let us now see how this model can deal with monotonicity. As we mentioned before, the monotonicity effect states that adding premises to a CBI argument increases expectations of property

[9] For cases in which this condition does not hold, it is possible that the volume of the hull of two large contiguous regions is larger than the hull of two distant small regions. An empirical study of this fact could be a way of testing the fruitfulness of the notion of *volume* of a category for the analysis of CBI.

[10] This assumption is not meant to be psychologically realistic. Prototypes are hardly centroids of the convex regions that represent them, even for natural categories (see Douven, 2019). However, according to the empirical literature, the typicality effect holds for multiple-premise arguments as a compound measure of the degree of typicality of some of the categories in the premises. Considering the lack of robust evidence about how these degrees of typicality interact, we introduced the centrality of the artificial prototype as a formal idealization that seems to respond well to the classical examples.

projection when the new premise-categories are also included in the original evoked superordinate category of the argument. For instance, adding a premise with the category PIG to the argument (FOX, WOLF)→ MAMMAL is going to strengthen it . Note that adding a premise-category to an argument will increase the volume of the convex hull of the premises. And in most cases, the volume of that set is negatively correlated to the distance between P_C and P_Y, that is, the more V(C(X_1, X_2,…, X_n)) approximates V(Y), the closest P_C is to P_Y. Then, for the above arguments we have that if P_C is the centroid of C(FOX ∪ WOLF) and P_{C*} is the centroid of C(FOX ∪ WOLF ∪ PIG), since V(C(FOX ∪ WOLF ∪ PIG)) > V(C(FOX ∪ WOLF)) then $sim\,(P_{C*}, P_{MAMMAL}) > sim\,(P_C, P_{MAMMAL})$, and as a consequence $ExpS(FOX, WOLF, PIG \rightarrow MAMMAL)_{MAMMAL} > ExpS(FOX, WOLF \rightarrow MAMMAL)_{MAMMAL}$.

This model can also predict Sloman's (1993) observation that diversity has a limit. To analyze his example (mentioned in Section 2.4), note that$sim\,(P_C, P_{MOLE})$, where P_C is the prototype of GERMAN C(SHEPHERD ∪ ELEPHANT), is considerably larger than $sim\,(P_{C*}, P_{MOLE})$, where P_{C*} is the prototype of C(SHEPHERD ∪ BLUE WHALE). Similarly $sim\,(P_C, P_{MAMMAL}) > sim\,(P_{C*}, P_{MAMMAL})$. Then it typically follows that

$$\left(\frac{V\,(C\,(\text{GERMAN SHEPHERD} \cup \text{ELEPHANT}))}{V\,(\text{MOLE})}.sim\,(P_C, P_{MOLE}) + a \cdot sim\,(P_C, P_{MAMMAL}) + b \cdot sim\,(P_{MOLE}, P_{MAMMAL})\right) > \left(\frac{V\,(C\,(\text{GERMAN SHEPHERD} \cup \text{BLUE WHALE}))}{V\,(\text{MOLE})}.sim\,(P_{C*}, P_{MOLE}) + a \cdot sim\,(P_{C*}, P_{MAMMAL}) + b \cdot sim\,(P_{MOLE}, P_{MAMMAL})\right).$$

Next, suppose that we add to some premise-set a new premise with a category that is not included in Z. What will happen is that the new modified argument will have a different (and more abstract) evoked superordinate category Z* such that Y ⊂ Z* with Z ⊂ Z*. According to the empirical literature, subjects should perceive the new argument as weaker than the original one, making CBI *nonmonotonic*. Remember the example of nonmonotonicity that we gave in Section 2.4: the argument (PEACOCK & CROW)→BIRD is stronger than the argument (PEACOCK & CROW & RABBIT)→BIRD (see Fig. 6). According to our analysis, adding the premise RABBIT change the evoked superordinate category (Z) from BIRD to ANIMAL. Then, our model correctly predicts that $logExpS(CROW, PEACOCK \rightarrow BIRD)_{BIRD} > logExpS(CROW, PEACOCK, RABBIT \rightarrow BIRD)_{ANIMAL}$ since

$$a.sim\,(P_C, P_{BIRD}) + b > a.sim\,(P_{C*}, P_{ANIMAL}) + b \cdot sim\,(P_{BIRD}, P_{ANIMAL})$$

$$a.sim\,(P_C, P_{BIRD}) > a.sim\,(P_{C*}, P_{ANIMAL}) \text{ and } b.sim\,(P_{BIRD}, P_{ANIMAL}) < b.$$

4.5. Knowledge effects and nonblank properties

The model presented so far only focuses on two types of semantic relations among premise and conclusion categories, namely similarity and typicality. However, newer experimental results on CBI have shown that there are other cognitive mechanisms that also influence inductive judgments. Beyond similarity and typicality relations, different kinds of knowledge about premise and conclusion categories (Coley & Vasilyeva, 2010; Shafto, Coley, & Vitkin, 2007) or different reasoning heuristics (Rehder, 2006) might shape inductive inferences. For instance, there is evidence that knowledge about thematic relations of the categories involved in the arguments (Coley, Shafto, Stepanova, & Baraff, 2005), as well as expertise in some domain related to the topic of the arguments (Proffitt, Medin, & Coley, 2000), can play an important role in the agent's expectations of property projection. Furthermore, in most cases of CBI, people also use knowledge (or make hypotheses) about the property projected for estimating the strength of the argument. In any case, a full model of CBI should account for the effects of background knowledge and consider nonblank properties. We believe that the conceptual spaces approach is rich enough to deal with the most studied knowledge effects involved in CBI concerning non-blank properties. In what follows we briefly explain how our model can be developed in this direction.

An influential criticism of the similarity-based models of CBI was presented by Heit and Rubinstein (1994). They showed that it is not possible to account for some knowledge effects that influence inductive inferences using only a single similarity measure. In particular, they showed that in category-based arguments with nonblank properties, the agents' knowledge about the property S that was projected modulated the similarity measure that was used for comparing the categories in the premise and conclusion. For instance, they showed that arguments of the form CHICKEN→HAWK are judged as stronger than arguments of the form TIGER→HAWK when the property projected is *anatomical* (such as "has a liver with two chambers"). But the opposite holds when the property projected is *behavioral* (such as "prefer to feed at night").

For explaining this phenomenon, we propose an extension of our model that includes a similarity measure that puts larger weights on the categories involved in the projected properties. This focus would be determined by the *dimensions* of the nonblank property in the arguments. When the agent has little knowledge of the property *S* that is projected (which is by definition the case for a blank property), she will compare categories using a general similarity measure. However, if the agent has more precise knowledge about *S* (at least about what kind of property *S* is), it is expected for her to use a similarity measure that gives more weight to the dimensions related to *S*.

Formally this can be done by using a weighted distance function like the one introduced in Section 3.3. Eq. (1) would be reformulated in the following way: $ExpS(X \rightarrow Y)_Z = d^{(S)}(d\,(X, Y) \cdot d^{(S)}(X, P_Z)^a \cdot d^{(S)}(Y, P_Z)^b)^{-1}$, where the function $d^{(S)}(x, y)$ is defined as the distance between *x* and *y* when the domains relative to *S* are salient. In the example from Heit and Rubinstein (1994), when the projected category refers to anatomical properties ("has a liver with two chambers"), the model will predict that the argument CHICKEN→HAWK will be judged as stronger than TIGER→HAWK. On the other hand, if the projected category refers to behavioral properties ("prefer to feed at night"), then the model will predict the converse relation — since now the weight to the behavioral domain will make TIGER more similar to HAWK, just as it was found in the experiments by Heit and Rubinstein.

Medin, Coley, Stroms, and Hayes (2003) showed that property effects also show up in arguments with blank properties. In particular, they discovered a *non-diversity effect by property reinforcement* that occurs when some salient feature shared between the premise-categories leads the agent to produce hypotheses about the nature of the property S that is projected (Shafto et al., 2007). As a result, the agent will use a weighted similarity measure that can override normal diversity effects. For instance, according to what we saw so far, the argument "Polar bears and antelopes have property S; thus all animals have property S" should be considered weaker than the argument "Polar bears and penguins have property S; thus all animals have property S", since the first premise set is less diverse than the second. However, in this case, the fact that both polar bears and penguins inhabit cold areas, leads agents to hypothesize that property S is related to this shared feature. That will weaken the argument, since properties of this kind are atypical regarding animals in general.

Our interpretation of this example is that the conjunction of POLAR and PENGUIN evokes a new (non-taxonomic) minimal superordinate, namely ANIMAL IN COLD AREAS and thereby that the property S somehow is related to this superordinate. The superordinate ANIMAL IN COLD AREAS generates a new way of classifying the similarity animals, that is a new distance function d*. As a result after applying Eq. (7), we expect that $logExpS(POLA\ RBEAR, ANTELOPE \rightarrow ANIMAL)_{ANIMAL} > logExpS(POLAR\ BEAR, PENGUIN \rightarrow ANIMAL)_{ANIMAL}$, since $a \cdot sim(P_C, P_{ANIMAL})+b$ will be bigger than $a{\cdot}sim(P_{C*}, P_{ANIMAL})+b$, because the distance between P_C and P_{ANIMAL} will be smaller than the distance between P_{C*} and P_{ANIMAL} since ANIMAL IN COLD AREAS is a rather small and atypical region of ANIMAL

Finally, similar ideas can be applied for explaining some of the effects of *expertise* in CBI (Proffitt et al., 2000). For a non-expert, the two inferences "Dutch elms have disease A; thus, gingko trees have disease A" and "River birches have disease A; thus, gingko trees have disease A" would, for lack of knowledge, be judged to be equally strong. For a tree expert, however, the knowledge that gingko trees are more similar to Dutch elms when it comes to which diseases affect them would make the first inference stronger than the second. In brief, for experts, the distance measure $d^{(S)}$ in the model would be dependent on that S relates to diseases, while this would not affect the non-experts' judgments.

These examples of how knowledge effects can be handled by our model show that it is able to cover a wide variety of experimental findings from the literature.[11]

5. Previous models of CBI

5.1. The similarity-coverage model

The first formal model of CBI was the *similarity-coverage model* (SCM), proposed by Osherson et al. (1990). In this model, argument strength in CBI is judged on the basis of two factors: (i) premise–conclusion similarity, and (ii) the degree of *coverage* that the premise's category has regarding the lowest superordinate category that includes both the category of the premise and the category of the conclusion.

For specific arguments, argument strength depends only on (i). If the argument has multiple-premises, the model uses a *maximum* rule that estimates premise–conclusion similarly by focusing on the premise with the most similar category to the conclusion's category. For instance, for an argument like "Horses and bats have property S; thus cows have property S", argument strength will be determined by $Maxsim(HORSE, COW); (BAT, COW)$, which will return $sim(HORSE, COW)$.

Coverage is a more complex notion. The model assumes – as we do – that CBI with natural categories involves "evoking" an implicit superordinate category that includes all the categories in the argument. Coverage is then a relation between the premises' categories with that superordinate category, and it is also explained in terms of similarity. More specifically, coverage is an average measure of several pairwise similarity judgments that compare the premise's category with members – "examples" – of the superordinate category in question; and it is a negative function of similarity among premises. For instance, consider the following arguments:

(a)	Horses have sesamoid bones *Cows have sesamoid bones* Mammals have sesamoid bones	(b)	Horses have sesamoid bones *Rats have sesamoid bones* Mammals have sesamoid bones

(a) is weaker than (b) because the pair (HORSE, COW) provides less coverage of MAMMALS than the pair (HORSE, RAT). In particular, the degree of coverage can be associated with the extension of the set which includes all the categories similar to those of the premises. In (a), that set is relatively small because most categories that are similar to HORSE are also similar to COW. In (b), however, that set is bigger, since most categories similar to RAT are not in the set of categories that are similar to HORSE.

Coverage is also related to typicality. The SMC assumes that typical categories are associated with larger sets of similar categories (of the same conceptual level) than atypical ones. For instance, the argument "Horses have property S; thus mammals have property S" is stronger than "Bats have property S; thus mammals have property S" because the set of mammals similar to horses is larger than the set of mammals similar to bats.

Despite being a very successful model thanks to its predictive power, the SCM has various limitations. One of them is that it does not build on a psychologically grounded notion of similarity. For instance, the model does not include any precise notion of similarity relations among categories. It uses *similarity* as an empty notion that can be filled out with different specific measures. As we mentioned before, it is desirable for a theory about CBI to build on some fundamental theory of conceptual knowledge; one that provides a basic notion of conceptual similarity and that can be used to give a unified explanation of the diversity of concept-based cognitive phenomena (categorization, concept formation, language-learning, etc.). Furthermore, as observed by Tenenbaum, Kemp, and Shafto (2007), the SCM lacks a systematic mathematical foundation. This is also related to the previous point. The formal structure of this model is not based on any formal model of inference or categorical relation, but it was directly designed to model the properties of CBI as described by the empirical studies.

Our approach, while it is close to the SMC model in various respects, does not suffer from the aforementioned problems since both the formal and the psychological foundations of our model come from the general theory of conceptual spaces. Furthermore, our model can account for the same range of CBI phenomena than the SMC model, while also predicting some results that are valid in special cases, something that the SMC model cannot do.

5.2. The feature-based model

The other well-known model of CBI was proposed by Sloman (1993) as an alternative to the SCM. Sloman started by criticizing the assumption that reasoning with categories involves the necessary representation of their hierarchical structure. He argued that inclusion fallacies in reasoning form strong evidence against that idea. As an alternative, he proposed to understand categorical relations as based on the overlap of features. "Features", Sloman claims, "represent a large number of interdependent perceptual and abstract attributes. In general, these values may depend on the context in which categories are presented" (1993, p. 237).

Sloman develops his *feature-based* model within a connectionist framework, in which categories are represented as sets of features described by vectors of real numbers from the [0,1]

[11] One area that we will not consider in this article is the influence of *causal relations* between the concepts involved. Various experimental studies have shown that causal knowledge is important in CBI, sometimes overriding standard similarity and typicality relations (Bright & Feeney, 2014; Medin et al., 2003; Rehder & Hastie, 2001; Shafto et al., 2007). For example, "Grass has enzyme E; thus cows have enzyme E" is judged to be stronger than "Cows have enzyme E; thus grass has enzyme E" since there may exist a causal link from the enzyme of the grass to the enzyme of the cows. One possible way to use our model also for these phenomena is that causal connections may introduce a different kind of 'typicality' relations between the concepts so that the presence of the enzyme is more typical for grass than for cows.

interval. With it, he is able to explain ten of the patterns explained by SCM and three new ones, not treated by Osherson et al. (1990). He also presents empirical support for the new patterns. The central idea of this model is that argument strength is positively correlated with the proportion of features in the conclusion category that are also included in the premise categories. For instance, in the simple case "All Xs are S; thus all Ys are S", the premise category X, and the conclusion category Y can be represented by two vectors $F(X)$ and $F(Y)$ of feature values. The strength of the inductive argument is determined by the following expression: $\frac{F(X)\cdot F(Y)}{|F(Y)|^2}$, where $F(X)\cdot F(Y)$ can be seen as a measure of the overlap of the features of X, Y, and $|F(Y)|^2$ a measure of the magnitude of the conclusion category vector.[12]

Unlike the SCM, the feature-based approach does not have foundational issues, since it is developed within a connectionist framework.[13] One could think that this "connectionist" background leaves no space to our conceptual spaces approach. However, as it has been argued before (Gärdenfors, 1997; Lieto, Chella, & Frixione, 2017), conceptual spaces is compatible with connectionist approaches.

In general, the main ideas of Sloman's model are not in contradiction with our conceptual space approach. In fact, they could be *translated* into our framework. The theory of conceptual spaces also assumes that concepts are represented as collections of properties from different domains. The feature-overlap measure that Sloman's use to determine argument strength could be replaced by a similarity measure in a conceptual space covering the dimensions of the feature vector.

One important advantage of the conceptual space model over the feature-based approach concerns the representation of typicality relations. In Sloman's model, there is no specific mechanism for accounting for typicality. Both typicality and similarity relations are reduced to feature-overlap. The model can account for typicality effects in general arguments because it assumes that typical categories (such as APPLE) share more features with their immediate superordinate category (FRUIT in this case) than non-typical categories. However, this model cannot account for independent premise-typicality effects in specific arguments. For instance, if we have three categories A, B and C, and A is more typical than B but both categories A and B have the same feature overlap with C, then the model would predict the arguments A $\rightarrow$ C and B $\rightarrow$ C to be equally strong (Heit, 2002, p. 586). The conceptual space approach does not have this limitation since it is able to explicitly represent independent typicality relations both in general and specific arguments. To give an example, consider two arguments of the form QUINCE $\rightarrow$ PINEAPPLE and APPLE $\rightarrow$ PINEAPPLE. The categories APPLE and QUINCE have the same feature overlap with PINEAPPLE, but since APPLE is a more typical fruit than QUINCE, $sim\,(P_{APPLE}, P_{FRUIT})$ is going to be significantly larger than $sim\left(P_{QUINCE}, P_{FRUIT}\right)$, and as a consequence $ExpS(APPLE \rightarrow PINEAPPLE)_{FRUIT} > ExpS(QUINCE \rightarrow PINEAPPLE)_{FRUIT}$. This is a second example of a new prediction that follows from our model. As far as we are aware, it has not been empirically tested.

In general, the two models presented here provide different insights into CBI. One interesting thing about our conceptual spaces model, is that it combines the two main features of the SMC and Sloman's model: it is a similarity-based model that includes a feature-based view of categories. Furthermore, our approach has an important theoretical advantage regarding these other two models; it inherits from the theory of conceptual spaces an explanation of how knowledge domains are formed and structured, and how they are grounded on perception an action. In that way, our formal model is grounded on a systematic psychological theory about the nature and structure of conceptual systems. At the same time, this psychological theory comes with a formal model of some of the main cognitive mechanisms behind conceptual processes. As we mentioned before, our model leverages this formal model and, in that sense, builds on a solid mathematical foundation. Another difference is, as Feeney (2017, pp. 172–173) notes, that neither SMC, nor the feature-based model can explain the conclusion effect reported by Hampton and Cannon (2004).

5.3. Bayesian models

Besides these two classical models of CBI, Bayesian accounts have recently become influential in the literature. The first proposal in this area was advanced by Heit (1998) and consisted of a computational-level analysis that puts the agent's knowledge about *properties* at the center stage of the process of CBI. His idea is that while evaluating a CBI argument, the agents estimate the probability of property projections among categories based on her estimation of the *range* of the property projected (i.e., the set of categories for which the property is true and the set of categories for which the property is false). For doing so, the agent exploits prior knowledge about other familiar properties, under the assumption that the property projected is *distributed* in a similar manner.

For instance, in an argument of the form "X has property S; thus Y has property S", the agent will reason from a set of four basic hypotheses about the possible range of S: (1) S is true of X and Y, (2) S is true of X and false of Y, (3) S is false of X and true of Y, and (4) S is false of X and Y. The prior probability distribution for these hypotheses may vary according to the similarity between X and Y or other categorical relations. Then, using the premise of the argument as evidence, the agent will update their beliefs about the set of hypotheses and estimate the probability of the conclusion using Bayes' theorem.

Heit showed that his approach predicts as many properties of CBI as Osherson's and Sloman's models. However, it has also an important drawback: it does not include any mechanism for estimating the prior distribution over the hypotheses about the range of the property. This is mainly because, unlike the other models (including ours), the Bayesian approach is centered on property-relations instead of categorical-relations.

Tenenbaum et al. (2007) followed Heit's approach and made some important improvements regarding the above problem. Their strategy consists of defining a set of structures with information about the agent's knowledge of categorical relations in different domains and knowledge about the compatibility of different properties with these relations. These structures will determine the prior probability that some property P may be projected from one category X to a category Y from the same domain. Then, these probabilities may be updated according to standard Bayesian rules when considering specific category-based arguments.

This approach can work with different types of knowledge structures. Taxonomic systems of categories, causal structures, or spatial knowledge are some of the knowledge structures that have been studied for CBI in the Bayesian tradition. This represents an important advantage over the SCM and Sloman's model, which have serious troubles for dealing with forms of inductive reasoning that do not involve natural categories.

There are, however, considerable drawbacks of Bayesian models of CBI. One is that there is no natural way to represent similarity and typicality in these models. Another is that probabilistic

[12] $F(X)\cdot F(Y)$ is the inner product of the two vectors, defined $\sum_i F(X)_i \cdot F(Y)_i$ and $|F(Y)|^2$ is the inner product of $F(Y)$ with itself, defined as $\sum_i F(Y)_i^2$.

[13] See Rogers and McClelland (2004) for a connectionist approach to semantic cognition.

reasoning is very resource-demanding when implemented computationally. In our opinion, these drawbacks make the Bayesian models psychologically unrealistic. (see Jones and Love (2011) for a general criticism of the use of Bayesian models in cognitive science).

6. A proposal for a new methodology

A major challenge that researchers on CBI face is to develop quantitative tests for the available models. The framework presented in Section 4 opens up for a new methodology of investigating category-based induction. The distance measure and the similarity and betweenness it generates will allow new and more precise quantitative predictions. We have already mentioned the prediction when Y is a subregion of $C(X_1 \cup X_2 \cup \cdots \cup X_n)$, then the prediction is that $ExpS(X_1, X_2 \rightarrow Y)_Z$ should be maximal. For regions Y, X_1 and X_2 say that Y lies *between* X_1 and X_2 if for every y in Y there are points x_1 and x_2 in X_1 and X_2 respectively, such that y is between x_1 and x_2. Given this definition, a special case of the prediction above is that if Y lies between X_1 and X_2, then $ExpS(X_1, X_2 \rightarrow Y)_Z$ should be maximal. A second new prediction concerns explicit representations of independent typicality relations as was discussed in Section 5.2.

Some other predictions are related to the introduction of the notion of *volume* of a category. First, our model predicts that premise-specificity is negatively correlated to argument strength. More formally, it is to be expected that for categories Y, X_1 and X_2, if $X_1 \subset X_2$ then $ExpS(X_2 \rightarrow Y) > ExpS(X_1 \rightarrow Y)$ because $V(X_2) > V(X_1)$. For instance, arguments of the form GERMAN SHEPHERD $\rightarrow$ COW (or MAMMAL) should be considered as weaker than arguments of the form DOG $\rightarrow$ COW (or MAMMAL). Second, our model predicts that for categories X_1, X_2 and Y, if it is the case that X_1, X_2 are equally typical, but that $V(X_1) > V(X_2)$ then$ExpS(X_1 \rightarrow Y) > ExpS(X_2 \rightarrow Y)$. These two predictions hold *ceteris paribus*.

To test these types of predictions, operational procedures for determining betweenness, similarity and distances are needed. There are general methods for estimating psychological distances, such as Multi-Dimensional Scaling (MDS) (see Hout, Papesh, & Goldinger, 2013 for a review) and Principal Component Analysis (PCA) (Abdi & Williams, 2010). For example, by asking subjects to judge the similarities of a number of different categories, the data can be analyzed by MDS or PCA in order to generate a low-dimensional conceptual space with a distance measure. Once the distance is established, similarity and betweenness can be determined, and the predictions presented above can be tested.

As an example of the relevant type of data collection, Hampton and Cannon (2004) asked subjects to rate the premise typicality, conclusion-typicality and premise–conclusion similarity on a seven-graded scale. This data could also have been used to estimate an underlying distance measure that would have allowed Eqs. (1) or (2) to be tested.[14]

7. Conclusion

Category-based induction is a cognitively central form of inductive inference that has become a focus of research only during the last decades. In this paper we have presented a mathematical model of such inferences that can explain almost all of the available empirical data. The model subsumes the earlier similarity-coverage model by Osherson et al. (1990) and Sloman's (1993) feature-based model and it generates new predictions. Furthermore, it builds on solid formal foundations and it relies on a systematic theory of conceptual knowledge that has been proven successful in the explanation and modeling of several concept-based cognitive phenomena.

We believe that the approach of analyzing inference in terms of how concepts are represented in conceptual space can be extended to other forms of reasoning such as inference based on analogies and metaphors (Gärdenfors, 2000, 2008), interpolative inference (Schockaert & Prade, 2013), and to nonmonotonic reasoning expectations-based reasoning (Gärdenfors, 1994; Gärdenfors & Makinson, 1994). If such a program can be worked through, it would form a unified basis for human inference that considerably extends the classical logicist and probabilistic approaches.

[14] Also Rips (1975) uses data from MDS for analyzing CBI arguments.

Acknowledgments

The authors would like to thank Max Kistler and two anonymous reviewers for helpful comments on an earlier version of this paper.

This research has received support from the DAAD (Germany) and the ANII (Uruguay).

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Computational Statistics*, *2*(4), 433–459.

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Bright, A., & Feeney, A. (2014). Causal knowledge and the development of inductive reasoning. *Journal of Experimental Child Psychology*, *122*, 48–61.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

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

Carnap, R. (1971). A basic system of inductive logic, part 1. In R. Carnap, & R. C. Jeffrey (Eds.), *Studies in inductive logics and probability, Vol. 1* (pp. 35–165). Berkeley: University of California Press.

Cherniak, C. (1984). Prototypicality and deductive reasoning. *Journal of Verbal Learning and Verbal Behavior*, *23*(5), 625–642.

Coley, J., Shafto, P., Stepanova, O., & Baraff, E. (2005). Knowledge and category-based induction. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: essays in honor of Douglas L. Medin* (pp. 69–85). Washington, DC: American Psychological Association.

Coley, J., & Vasilyeva, N. (2010). Generating inductive inferences premise relations and property effects. *Psychology of Learning and Motivation*, *53*, 183–226.

Decock, L., & Douven, I. (2011). Similarity after goodman. *Review of Philosophy and Psychology*, *2*(1), 61–75.

Decock, L., & Douven, I. (2014). What is graded membership? *Nous*, *48*(4), 653–682.

Devadoss, S., & O'Rourke, J. (2011). *Discrete and computational geometry*. Princeton University Press.

Douven, I. (2019). Putting prototypes in place. *Cognition*, 193.

Douven, I., & Gärdenfors, P. (2019). What are natural concepts? A design perspective. *Mind and Language*, 1–22.

Eliot, J. (1987). *Models of psychological space*. New York, NY: Springer-Verlag.

Evans, J. S. B. T. (1989). Concepts and inference. *Mind and Language*, *4*(1–2), 29–34.

Feeney, A. (2017). Forty years of progress on category-based inductive reasoning. In *International handbook of thinking and reasoning* (pp. 167–185). London: Routledge.

Feeney, A., & Heit, E. (2007). *Inductive reasoning*. Cambridge University Press.

Gärdenfors, P. (1991). Nonmonotonic inference, expectations, and neural networks. In R. Kruse, & P. Siegel (Eds.), *Lecture notes in computer science: vol. 548, Symbolic and quantitative approaches to uncertainty. ECSQARU 1991*. Berlin, Heidelberg: Springer.

Gärdenfors, P. (1994). The role of expectations in reasoning. In M. Masuch, & L. Pólos (Eds.), *Logic at work 1992. lecture notes in computer science: vol. 808, Knowledge representation and reasoning under uncertainty* (pp. 1–16). Berlin, Heidelberg: Springer.

Gärdenfors, P. (1997). Symbolic, conceptual and subconceptual representations. In V. Cantoni, V. Di Gesu, A. Setti, & D. Tegolo (Eds.), *Human and Machine Perception* (pp. 255–270). Boston, MA: Springer US.

Gärdenfors, P. (2000). *Conceptual spaces*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2008). Reasoning in conceptual spaces. In J. E. Adler, & L. J. Rips (Eds.), *Reasoning: studies of human inference and its foundations* (pp. 302–320). Cambridge University Press.
Gärdenfors, P. (2014). *The geometry of meaning: semantics based on conceptual spaces*. Cambridge, MA: MIT Press.
Gärdenfors, P., & Makinson, D. (1994). Nonmonotonic inference based on expectations. *Artificial Intelligence*, *65*(2), 196–245.
Gärdenfors, P., & Stephens, A. (2018). Induction and knowledge-what. *European Journal for Philosophy of Science*, *8*(3), 471–491.
Goldstone, R., Medin, D., & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, *25*(2), 237–255.
Goodman, N. (1972). Seven strictures on similarity. In *Problems and projects* (pp. 437–446). Indianapolis/New York: Bobbs-Merrill.
Hampton, J. A., & Cannon, I. (2004). Category-based induction: An effect of conclusion typicality. *Memory and Cognition*, *32*(2), 235–243.
Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 278–292.
Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford, & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford: Oxford University Press.
Heit, E. (2000). Features of similarity and category-based induction. In *Proceedings of the interdisciplinary workshop on categorization and similarity* (pp. 115–121). University of Edinburgh.
Heit, E. (2002). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, 1–24.
Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *20*(2), 411–422.
Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93–103.
Hume, D. (1999). *An enquiry concerning the human understandIng*. Oxford: Oxford University Press,
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge.
Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York: Harper and Row.
Jäger, G. (2010). Natural color categories are convex sets. In *Logic, language and meaning* (pp. 11–20). Berlin, Heidelberg: Springer.
Johannesson, M. (2003). Geometric models of similarity. In *Lund university cognitive studies* (p. 90). Lund University.
Johnson-Laird, P. (2010). Against logical form. *Psychologica Belgica*, *50*(3), 193–221.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188.
Keßler, C., Raubal, M., & Janowicz, K. (2007). The effect of context on semantic similarity measurement. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *Lecture notes in computer science: vol. 4806, On the move to meaningful internet systems 2007*. Berlin, Heidelberg: Springer.
Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.
Lieto, A., Chella, A., & Frixione, M. (2017). Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, *19*, 1–9.
López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*(5), 1070–1090.
Maddox, W. (1992). Perceptual and decisional separability. In G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 147–180). Hillsdale, NJ: Erlbaum.
Medin, D. L., Coley, J. D., Stroms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*(3), 517–532.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
Mervis, C., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.
Niiniluoto, I. (1987). *Truthlikeness*. Dordrecht: Reidel.
Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, *115*(1), 39–57.
Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind and Language*, *6*(1), 1–38.
Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.
Osherson, D., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *1990*, 185–200.
Proffitt, J., Medin, D., & Coley, J. (2000). Expertise and category-based induction. *Journal of Experimental Psychology*, *26*(4), 811–828.
Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
Quine, W. V. O. (1969). *Ontological relativity and other essays*. Columbia University Press.
Quine, W. V. (1974). *The roots of reference*. LaSalle: Open Court.
Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory and Cognition*, *34*(1), 3–16.
Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology*, *130*(3), 323–360.
Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665–681.
Rips, L. J. (1989). Similarity and the structure of concepts. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York, NY: Cambridge University Press.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition*. MIT Press.
Rosch, E. (1971). Focal color areas and the development of color names. *Developmental Psychology*, *4*, 447–455.
Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language* (pp. 111–144). Academic Press.
Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.
Rosch, E. (1978). Prototype classification and logical classification: the two systems. In E. Scholnik (Ed.), *New trends in cognitive representation: challenges to piaget's theory* (pp. 73–86). Hillsdale: Lawrence Erlbaum.
Schockaert, S., & Prade, H. (2013). Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. *Artificial Intelligence*, *202*(C), 86–131.
Shafto, P., Coley, J., & Vitkin, A. (2007). Availability in category-based induction. In A. Feeney, & E. Heit (Eds.), *Inductive reasoning* (pp. 114–136). Cambridge: Cambridge University Press.
Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*(4), 373–421.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.
Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.
Sloman, S. A., & Lagnado, D. (2005). The problem of induction. In R. Morrison, & K. Holyoak (Eds.), *Cambridge handbook of thinking & reasoning* (pp. 95–116). New York: Cambridge University Press.
Stalnaker, R. (1981). Antiessentialism. *Midwest Studies of Philosophy*, *4*, 343–355.
Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney, & E. Heit (Eds.), *Inductive reasoning* (pp. 167–204). Cambridge University Press.
Thagard, P. (1984). Frames, knowledge, and inference. *Synthese*, *61*(2), 233–259.
Thagard, P. (1988). *Computational philosophy of science*. MIT press.
Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.
Ungerer, F., & Schmid, H.-J. (2006). *An introduction to cognitive linguistics*. London: Pearson-Longman.