The Role of Naturalness in Concept Learning: A Computational Study

Igor Douven

The Role of Naturalness in Concept Learning: A Computational Study*

Igor Douven
IHPST / CNRS / Panthéon–Sorbonne University
igor.douven@univ-paris1.fr

Abstract

This paper studies the learnability of natural concepts in the context of the conceptual spaces framework. Previous work proposed that natural concepts are represented by the cells of optimally partitioned similarity spaces, where optimality was defined in terms of a number of constraints. Among these is the constraint that optimally partitioned similarity spaces result in easily learnable concepts. While there is evidence that systems of concepts generally regarded as natural satisfy a number of the proposed optimality constraints, the connection between naturalness and learnability has been less well studied. To fill this gap, we conduct a computational study employing two standard models of concept learning. Applying these models to the learning of color concepts, we examine whether natural color concepts are more readily learned than nonnatural ones. Our findings warrant a positive answer to this question for both models employed, thus lending empirical support to the notion that learnability is a distinctive characteristic of natural concepts.

Keywords: concepts; design; learning; naturalness; optimality; similarity spaces; simulations.

1 Introduction

In the conceptual spaces framework, as developed by Gärdenfors (2000, 2014) and others, concepts can be represented geometrically, specifically as regions, or sets of regions, in so-called similarity spaces. For example, the concept of redness can be identified with a certain region in perceptual color space (see below), and the concept of tartness, with a certain region in taste space (e.g., Churchland, 2012). However, not every region in a similarity space can represent a concept, at least not a *natural* one, where natural concepts are concepts like BLUE, or GREEN, or TIGER, or GOLD, which carve nature at its joints (in the words of Plato's Phaedrus) and which have or can have a place in our thinking and theorizing. The question then arises what distinguishes those regions that do or can represent natural concepts from those that cannot.

Gärdenfors (2000, p. 71 ff) suggests that natural concepts are represented by *convex* regions in a similarity space,[1] where a region is convex precisely if for any two points lying in the region, every point between them lies in the region as well. Although both empirical evidence and considerations of cognitive economy support this idea, it is clear that convexity alone is not enough to define naturalness. For instance, it is easy to pick a convex region in perceptual color space that groups together color shades which we would not naturally associate.

*The Supplementary Materials, including the Julia file that was used for the simulations reported, are available in a GitHub repository which can be accessed via this link.

[1]Or by *sets* of convex regions, if we are dealing with a multi-domain concept. We take this qualification to be read from here on.

In response to this problem, Douven and Gärdenfors (2020) argue that natural concepts are ones that are represented by the cells of an optimally partitioned similarity space. As they show, this proposal encompasses, but goes beyond, the aforementioned convexity criterion. At the core of their proposal is a set of constraints that a partitioning must satisfy to be considered optimal. While, as Douven and Gärdenfors (2020) show, there is already support—both empirical and from computational studies—for most of the constraints they propose, this cannot be said for what they call *Learnability*, according to which an optimal partitioning creates concepts that are easy to learn and allow for reliable generalization from a few samples.

Previous studies have shown that people generally learn new concepts quickly. However, most of this research focuses on the role of prototypes in concept learning. It does not specifically address whether learnability is a distinctive characteristic of *natural* concepts, given that much of it concerns artificial concepts—like ones about triangular patterns of dots distorted to varying degrees (Posner, Goldsmith, & Welton, 1967; Posner & Keele, 1968, 1970) or about artificial faces (Reed, 1972)—or involves both natural and artificial concepts (Rosch & Mervis, 1975), without however examining whether naturalness differentially affects learnability.

We present evidence for Learnability in the form of the outcomes of a computational study using two distinct models of concept learning. We use these models to simulate the learning of color concepts, with an eye toward investigating whether natural color concepts (i.e., the mental correlates of basic color terms like "blue," "red," "green," etc.; see Berlin & Kay, 1969) are learned more readily than systems with equal numbers of nonnatural color concepts. It will be seen that, given either model, the answer to this question is positive. Before we delve into our study, we outline the theoretical background that informs our work.

2 Theoretical background

We start by stating the basic tenets of the conceptual spaces framework, then summarize recent work on the problem of how to define naturalness of concepts in that framework, and lastly describe the two methods of concept learning our study will rely on.

2.1 From similarities to concepts

At the core of the conceptual spaces framework is the thought that concepts can be represented as regions in similarity spaces, where a similarity space is a one- or multi-dimensional metric space whose dimensions represent measurable qualities that objects can have to varying degrees. Objects are mapped onto points in such spaces depending on the degree to which they possess these qualities. In principle, there is a great choice of metrics, but the ones most commonly encountered in practice are the Manhattan and the Euclidean metric, which, given an n-dimensional space S, are the instances of the schema below with $p = 1$ and $p = 2$, respectively:

$$\partial_S(x, y) \quad = \quad \left(\sum_{i=1}^{n} | x_i - y_i |^p \right)^{1/p},$$

where $x = \langle x_1, \ldots, x_n \rangle$ and $y = \langle y_1, \ldots, y_n \rangle$. The measure in which two objects representable in S are similar to each other, in the respect corresponding to S (e.g., similar in taste, if S is taste space), is then some inverse function of the distance between them in S, meaning that they are the more similar in the said respect the closer together they are represented within S.

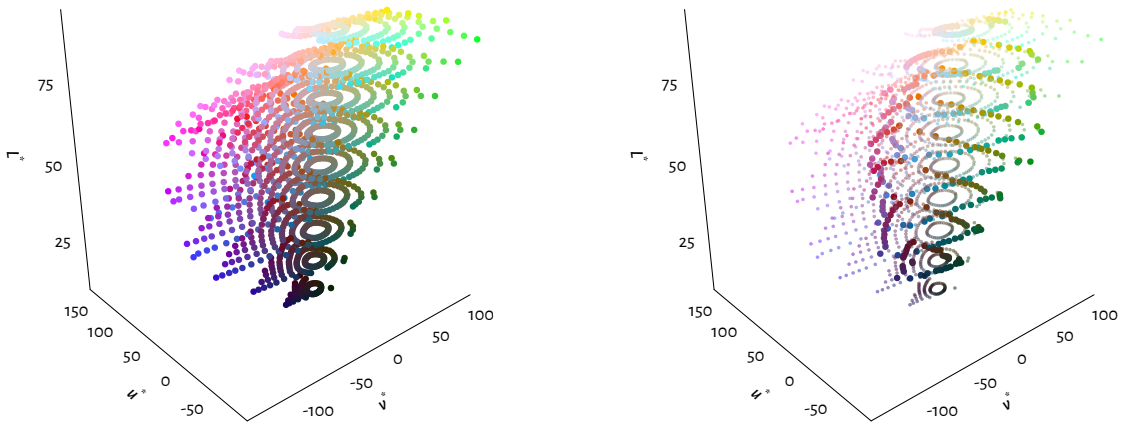

Figure 1: The 1,625 chips from the Munsell Book of Colors shown in CIELUV space (left panel); the 320 chromatic chips used for the World Color Survey highlighted (right panel).

We find a great variety of similarity spaces discussed in the cognitive science literature, including the already mentioned taste space, auditory spaces (Petitot, 1989), olfactory space (Castro, Ramanathan, & Chennubhotla, 2013), various shape spaces, such as a space for shells (Gärdenfors, 2000, pp. 142–150), for human faces (Valentine, Lewis, & Hills, 2016), and for container objects (Douven, 2016), various animal spaces (e.g., Henley, 1969), as well as various action spaces (e.g., Gärdenfors & Warglien, 2012), moral spaces (e.g., Peterson, 2017; Verheyen & Peterson, 2021), and social spaces (e.g., Deauvieau et al., 2014; Bendifallah et al., 2023). Perhaps the best known similarity space, and certainly the one most easily accessible for experimentation, is perceptual color space. Because it is also the space to be used in our study, we will describe this in more detail.

There are in effect *two* perceptual color spaces, CIELUV space and CIELAB space, where the former is assumed to best represent similarity judgments concerning colors perceived on screen while the latter is assumed to best represent such judgments when colors are shown on cloth or paper (Malacara, 2002, pp. 86–90; Fairchild, 2013, Ch. 10). The spaces look very similar: both are three-dimensional, spindle-like Euclidean spaces. The left panel of Figure 1 visualizes CIELUV space by placing in it all 1,625 chips from what is commonly known as "the Munsell book of colors" (Munsell, 1941).[2] The right panel of this figure highlights the 320 chromatic Munsell chips that have been widely used in color naming studies, most famously those reported in Berlin and Kay (1969) and the World Color Survey (Cook, Kay, & Regier, 2005). These chips will also serve as the main materials in our computational study.

There are different ways to obtain a conceptual space from a similarity space, but the currently dominant approach combines prototype theory with the mathematical technique of Voronoi tessellations (Gärdenfors, 2000, 2014). Locating the points representing prototypes in a given space S, we can let them generate a Voronoi tessellation of S by associating each point in S with the prototype or prototypes closest to it. This will result in a partitioning of S, the cells of which can be considered to rep-

[2] The RGB coordinates of the Munsell chips were downloaded from the website of the Munsell Color Science Laboratory of the Rochester Institute of Technology and converted to CIELUV coordinates using the Colors.jl package for the Julia language.

resent concepts, while the points which are equally close to two or more prototypes form the concept boundaries.

We briefly mention also a different approach to representing concepts in similarity spaces, which is advocated by Nosofsky (1986, 1987). In his Generalized Context Model, concepts are conceived not as regions of points but as sets of exemplars, or individual instances of a concept, represented in a similarity space. Our focus will be on Gärdenfors' model, which offers a richer mathematical structure, thereby facilitating the formalization of more complex cognitive operations and relations.[3] Nosofsky's GCM is still relevant to our study, inasmuch as it may provide an important complementary viewpoint concerning concept *learning*, specifically, by framing this as a comparative process of assimilating new items based on their similarity to items stored in memory, in a way to be detailed shortly.

2.2 Naturalness and optimality

The study we are to report is directly related to the question of what distinguishes natural from non-natural concepts. What, for instance, distinguishes the natural color concepts from ones obtained by picking a number of points in CIELUV space randomly and letting those generate a Voronoi tessellation of that space? Given that any Voronoi tessellation of a Euclidean space partitions that space into convex regions (Okabe et al., 2000, p. 58), the convexity criterion is not going to help answering this question. As mentioned in the introduction, Douven and Gärdenfors (2020) propose to analyze the notion of a natural concept in terms of optimally partitioned similarity spaces. More exactly, natural concepts are those that are represented by a cell in an optimally partitioned similarity space.

In doing so, these authors took their cue from work in cognitive science relating categorization to principles of rationality and optimality (e.g., Rosch, 1978; Marr, 1982; Anderson, 1990, 1991; Oaksford & Chater, 1994; Chater & Oaksford, 1999; Goodman et al., 2008; Frank & Goodman, 2012; Griffiths, Lieder, & Goodman, 2015) as well as from recent work in biology, which explains the occurrence of certain biological traits and processes by reference to what they would look like had they been designed according to principles of good engineering; for instance, Alon (2003) argues that many biological networks look *as if* they had been constructed by someone who obeyed the same principles of modularity, use of recurring circuit elements, and robustness of component tolerances that guide the design of engineered networks.

The main part of Douven and Gärdenfors' (2020) proposal consists of a set of design principles that a good engineer would want to follow if tasked with designing one or more conceptual spaces for creatures with our perceptual and cognitive make-up. More exactly, these authors formulate a number of design principles which they believe to define optimal design for representational systems generally, regardless of which format one prefers for concepts (so regardless of whether one commits to the conceptual spaces framework). But then their main concern is to show what these principles imply for the design of conceptual spaces, specifically, for when such spaces are optimally designed—optimally designed for creatures whose memories have limited storage capacity, whose discriminatory capacities are limited as well, and who are to survive in a world with scarce commodities where competition for those commodities can be fierce.

One of the principles is what Douven and Gärdenfors call *Informativeness*, according to which a partitioning should offer fine distinctions across the similarity space. But the authors also stress the need for *Parsimony*, cautioning against partitioning a space so finely that it overloads memory. Two further principles—*Contrast* and *Representation*—have to do with the placement of prototypes in the space.

[3] For instance, Gärdenfors' model has been used to derive prior probabilities for items falling under specific concepts (Decock, Douven, & Sznajder, 2016) and to formalize vagueness (Douven et al., 2013; Douven, 2016; Douven et al., 2017) as well as analogical reasoning (Douven et al., 2022). It is not clear to us how any of that could be accomplished using the GCM.

According to the former, the partitioning should allow a placement of the prototypes such that they can easily be told apart from one another (i.e., prototypes of different concepts should be at a sufficiently large distance from each other in the space). According to the latter, the placement should at the same time be such that the prototypes are good representatives of the items falling under the concept they are the prototype of (i.e., they should not be too distant from any of the items falling under the concept). The final principle on Douven and Gärdenfors' list is *Learnability*, which, as already mentioned in the introduction, demands that the concepts resulting from the partitioning should be easy to learn, in particular, that concept learners should be able to safely (even if not infallibly) generalize to unlabeled instances from a relatively small number of labeled instances.

Douven and Gärdenfors explicitly leave open the possibilities that their list of principles is incomplete as well as that it contains redundancies. They further recognize that different ones of their principles can pull in different directions. Therefore, they define an optimal partitioning of a similarity space to be one that strikes the or a best balance between the different desiderata.

In their paper, Douven and Gärdenfors muster a range of experimental and computational results reported in the scientific literature which support various of their principles. For example, Informativeness and Parsimony receive strong support from Regier, Kay, and Khetarpal (2007). Following an idea put forward by Jameson and D'Andrade (1997), these authors apply a computational clustering algorithm implementing principles that, at bottom, amount to Informativeness and Parsimony to the Munsell chips used in the WCS, finding that the algorithm produces clusterings that closely match how those same chips are carved up into categories by various languages spoken across the globe.[4] Douven (2019) provided evidence specifically for Contrast and Representation by first experimentally determining the locations in CIELUV space of the basic color prototypes and then showing via simulations that the constellation of those prototypes is a near-to-optimal trade-off between the two designated constraints. As said, however, evidence supporting Learnability is still sparse, which motivated the present paper.

2.3 Concept learning

Laboratory studies involving human participants might seem ideal to examine whether natural concepts are acquired more readily than nonnatural ones. However, such studies are likely to be confounded by the fact that the participants are already familiar with the natural concepts. Whichever difficulties they might have acquiring nonnatural concepts, at least as compared to natural concepts, could be due to the fact that the concepts to be acquired (as part of the experimental task) clash with the ones the participants have been using for most of their lives. A way around this might be to conduct an experiment in which toddlers in the earliest stages of language development are trained on nonnatural concepts (say, nonnatural color concepts) prior to their having acquired the natural ones. But it would probably be difficult to get such an experiment approved by the ethics committee of one's university.

At any rate, in this paper we take a computational rather than an experimental approach to investigating concept learning. More exactly, we consider two algorithms that are suited for modeling concept learning, supposing concepts are represented as regions in similarity spaces. One is a computational implementation of Gärdenfors' (2001) Approximate Prototype Model (APM), which is his account of concept learning set within the conceptual spaces framework. The other is the k-Nearest Neighbors (KNN) algorithm, which is a popular machine learning algorithm and which, as Nosofsky (2011, p. 22)

[4] For similar results, see Kemp and Regier (2012), Xu and Regier (2014), Xu, Regier, and Malt (2016), Zaslavsky et al. (2019), and Mollica et al. (2021).

notes, is closely connected, or even identical (depending on the exact functional relationship between similarity and distance), to his GCM.

According to the APM, we learn a family of concepts (e.g., color concepts) by going through the following steps:

1. we are shown examples of each of the concepts in the family, where these examples are *labeled*, meaning that it is given, for each, under which concept it falls;
2. we estimate the prototype of each concept by averaging the examples of the concept, meaning that, where these examples are given as points in an n-dimensional similarity space, we first group the examples according to their labels and then calculate, for each group and each dimension i ($1 \leqslant i \leqslant n$), the average i-th coordinate of the members of the group (this yields the group's so-called center of gravity in the space);
3. we use these estimates to generate a Voronoi tessellation of the relevant similarity space;
4. when new labeled examples come in, we repeat steps 2 and 3.

The hope is that, if all goes well, the Voronoi tessellation generated by our best guesses of prototypes will converge toward the *actual* conceptual structure, as we come to learn more and more labeled examples. How accurately you will be able to classify *unlabeled* examples at a later point in time will depend on how close, at that point in time, the best-guess structure is to the actual one, and is a measure of the extent to which you can be said to possess the concepts represented in the relevant space. Figure 2 illustrates the process in the abstract.

The GCM-based model of learning is simpler. In this model, you learn a family of concepts by storing in memory more and more exemplars (i.e., labeled examples) of each of the concepts. Upon encountering an unlabeled example, you find the exemplars most similar to it, and you label it on the basis of those exemplars' labels. This leaves some details to be filled in. For instance, if all the most similar exemplars have the same label, then the decision how to label the new example will be straightforward—but what if they have different labels? And how many most-similar exemplars are we to consider in order to determine the label of the new example?

To start with the latter question, the k-Nearest Neighbors algorithm can be regarded as implementing the GCM-based model of concept learning, and as its name suggests, this algorithm considers the k most similar exemplars. It allows the number k to be set by the user, though a common recommendation is to set k equal to the integer nearest to the square root of the total number of exemplars. With k fixed, the k most similar exemplars vote on the label to be given to the new example, where this vote can be weighted depending on the distances in the relevant similarity space between the exemplars and the new example. For instance, suppose $k = 5$ and three of the exemplars most similar to a new example are labeled l_1 while two are labeled l_2. Then the unweighted majority vote would yield l_1 as the label for the new example. If, however, we take weights into account, this could be different. For example, if the l_2-exemplars are much closer to the new example than any of the l_1-exemplars, then the new example could end up being labeled l_2. Most packages providing KNN allow users to set a great variety of weighting functions.[5]

The two learning methods to be used in our study have in common that they are both forms of *supervised* learning, meaning that they both require *labeled* data to learn from. But they are different in the sense that the APM is a form of what in machine learning is called *greedy learning*, while KNN is a form of *lazy learning*. That is to say that, in learning, the former creates a *map*—in our case a conceptual structure—and then labels new examples according to their place on the map. KNN, on

[5] In Nosofsky's GCM, there are parameters to model response bias, memory strength, and salience of dimensions, which help to model recency and context effects. For our computational implementation, these do not matter. For instance, the order in which data are stored in computer memory makes no difference to how easily they can be retrieved.

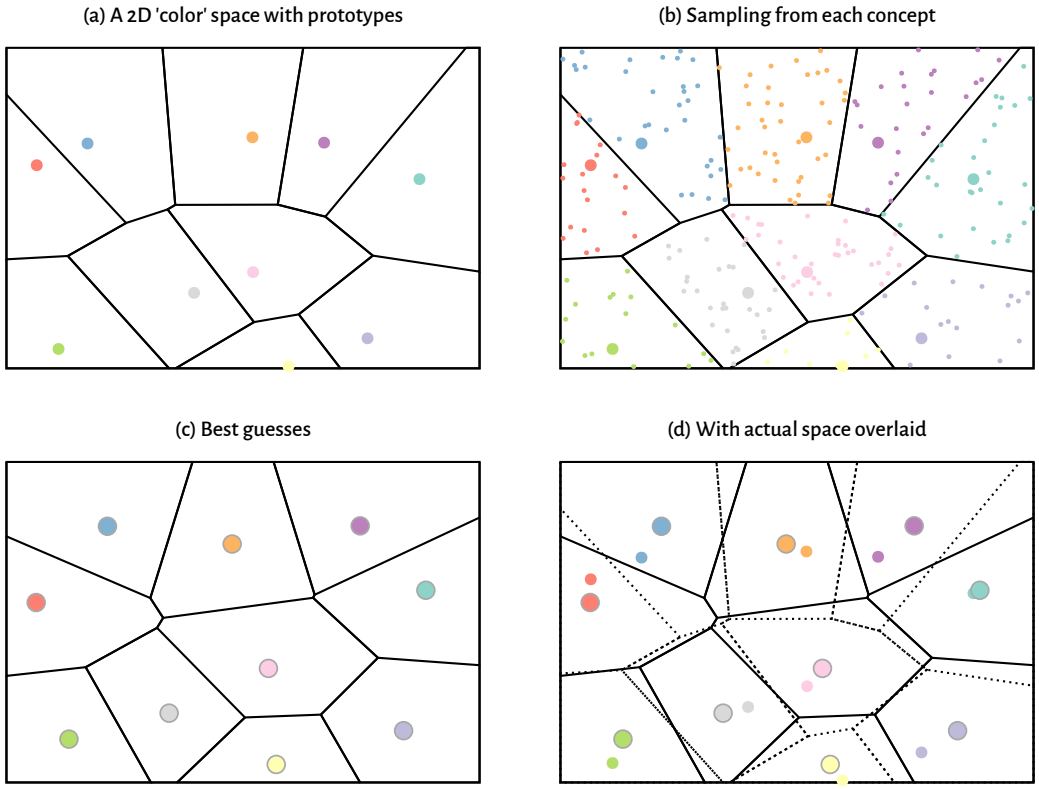

Figure 2: Illustration of the Approximate Prototype Model of concept learning, using a two-dimensional "color" space, shown with prototypes (panel a); based on the sampling from each of the concepts (panel b), a best guess is made of the locations of the various prototypes in the space, to which corresponds a best guess of the conceptual structure, which is the Voronoi tessellation generated by the prototype estimates (panel c); this is compared with the actual space (panel d), the latter shown in dotted lines.

the other hand, runs one and the same routine each time a new example is received in order to determine the label of that example, without ever creating a map. Of course, if we want, we can let it create a map by approximation, by asking for a fine grid of points in a given space how the algorithm would classify them. To show how this would work, we can use the same space and sample of examples that was used to illustrate the APM and let KNN predict the label for each point in a grid of 500×500 points, spaced uniformly in each dimension. The result is shown in Figure 3. Just glancing at the results from the two illustrations, it would seem that KNN offers a somewhat better approximation of the real conceptual structure than the APM does. Needless to say, though, this is a toy example of a space and so we should be cautious to infer anything in general from the relative performance of the two learning methods.

The main focus of this paper is on the question of how to distinguish natural from nonnatural concepts, not on the different learning methods (though we will briefly compare their performance in our study). It is still worth commenting on what could appear as problematic aspects of the methods—problematic, at least, from the standpoint of the conceptual spaces framework.

The first concerns the APM and is illustrated in Figure 4. As can be seen in that figure, some of the examples that underlie the estimate of the conceptual structure end up, not in the concept they were supposed to be an example of, but in a neighboring one. The problem was already mentioned, though only in passing, in Gärdenfors (2001, p. 176 n), where it is speculated that these problem cases

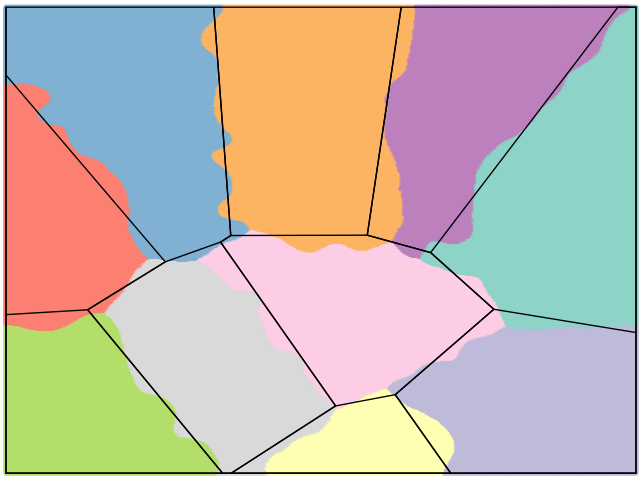

Figure 3: Map created by means of the k-Nearest Neighbors algorithm on the basis of the same sample of exemplars used for the illustration in Figure 2.

might actually give rise to the introduction of new concepts. A different response, which we want to propose here, points to the fact that the issue of vagueness is typically "idealized away" in presentations of the conceptual spaces framework. The framework *can* model vagueness, as was shown in Douven et al. (2013), Decock and Douven (2014), Douven and Decock (2017), Verheyen and Égré (2018), and Douven (2020), but "fuzzifying" concept boundaries in a realistic manner can be computationally costly. A quick and dirty way of representing vagueness uses the notion of silhouette width, as developed by Kaufman and Rousseeuw (2005), which helps to distinguish between instances that are more central to a concept and those that are more peripheral to it. Given a family of concepts $\mathscr{C} = \{C_1, \ldots, C_k\}$ representable in some similarity space S, with x an instance of concept C_i, its silhouette width is defined

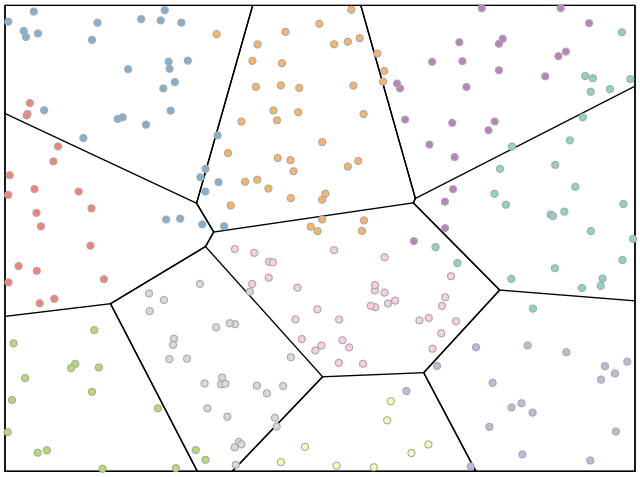

Figure 4: Best APM guess of conceptual structure and sample on which the guess is based: some sampled exemplars end up in a concept other than the one they exemplify.

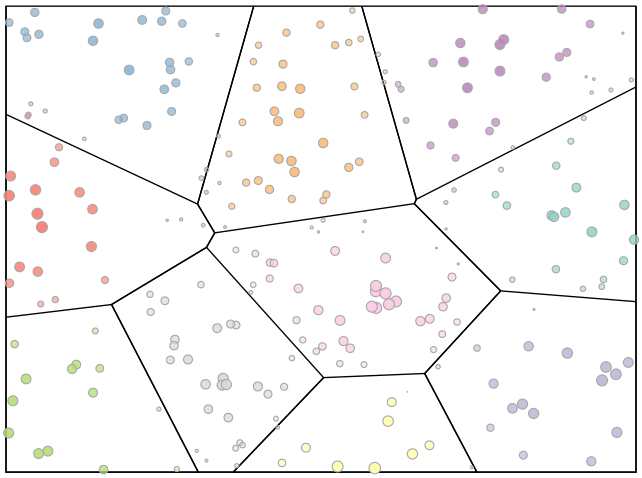

Figure 5: Markers sized according to the silhouette width of the exemplars.

to be $\big(b(x) - a(x)\big) \,/\, \max\big\{a(x), b(x)\big\}$, with $a(x)$ the average distance in S of x to the other instances of C_i and $b(x)$ the smallest average distance in S of x to the instances falling under $C_j \in \mathscr{C}$, for $j \neq i$. Sizing the point markers according to the silhouette width of the exemplars they represent, as is done in Figure 5, suggests that the problem cases might all turn out to be borderline cases, once vagueness is taken into account.

To see why this is relevant, note that a parent trying to teach a child the concept of redness would probably not want to pick red–orange, or red–pink, or red–blue borderline cases. That does not mean she will only pick examples that are typically red, as this might fail to give a good impression of the extension or width (in Carnap's, 1980, sense) of the concept. But she will *either* want to leave out borderline instances completely *or* explicitly single them out as being, for instance, reddish-orangish. Note also that, in other spaces, the boundary regions between concepts may not even *have* instances. For instance, mammal space, as represented in Henley (1969), can plausibly be thought of as having continuous dimensions, but for biological reasons, there may be gaps in this space. Thus, one diagnosis of the problem cases as highlighted in Figure 4 is that we have been assuming that examples are sampled uniformly from a space, which in reality may often be false. For pedagogical purposes, we may want to select examples that are rather central to a concept, or we may be *bound* to select such examples, because there are no borderline cases, for biological or more broadly scientific reasons.

The second problem concerns KNN, or rather, KNN as a method for concept learning, where concepts are understood as per the conceptual spaces program. After all, it was said that even if convexity is not a sufficient condition for naturalness, it is generally agreed among proponents of the framework that it is a necessary one. And from Figure 3 it is evident that if we learn concepts via KNN, we cannot expect them to be convex. Note, though, that the deviations from convexity might in their entirety fall within the boundary regions of the concepts and thus could, in a realistic setting, be nothing but artefacts of the idealizing assumption that concepts have sharp borders. This is particularly plausible in view of the results reported in Douven (2016), Douven et al. (2017), and Douven et al. (2018), which show for different real similarity spaces that boundary regions can have substantial volume, relative to the total volume of the space.

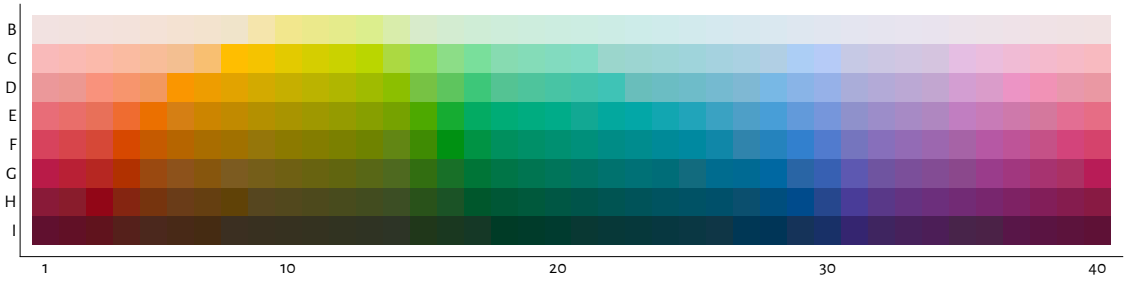

Figure 6: The 320 chromatic Munsell chips used for the World Color Survey.

3 Computational study

As previously mentioned, perceptual color space is readily available to experimenters. CIELUV and CIELAB space can both be accessed via dedicated packages in a great number of popular computer languages, which is true for none of the other similarity spaces known from the psychological literature (Sect. 2.1). For our study, we have used the Colors.jl package for the scientific computing language Julia (Bezanson et al., 2017) to work with these spaces. The results to be reported in the following were obtained in CIELUV space, but interested readers are invited to use the code provided in the Supplementary Materials to verify that rerunning the procedures to be stated below in CIELAB space yields qualitatively identical outcomes.

Given our focus on color space, the specific research question we aim to answer becomes this: Is the system of color concepts we use, which is widely regarded as a prime example of a system of natural concepts (e.g., Rosch, 1973; Kripke, 1980; Hardin, 1988), more easily (more quickly, more reliably) learnable than systems of color concepts that we would regard as nonnatural? Given a number of exemplars of all color concepts, can we more reliably generalize from those exemplars to the conceptual structure if that structure is the natural one (i.e., the carving-up of perceptual color space into natural concepts) than if it is a nonnatural structure? To further clarify, consider a paradigmatic situation of concept learning in which a parent teaches her child color names by showing the child shades of blue, shades of red, shades of green, and so on, while naming those shades. We can then ask how accurately the child will be able to predict the color names of shades it has not yet seen or in any case has not been taught to name. We can further ask whether, if our system of color concepts consisted of nonnatural concepts, the learning process would be slower in that the child would have to see more exemplars of each color to achieve a similar level of accuracy in naming so-far-unseen color shades. To repeat, we are not going to use human participants to test the hypothesis that the answer will turn out in favor of the natural color concepts, as Douven and Gärdenfors (2020) predict it will, but rather try to model the learning process as best we can by computational means.

The materials for the computational study consisted in the 320 chromatic chips from the WCS, which were already highlighted in the right panel of Figure 1 and which Figure 6 shows in a chart, in the way they are usually presented in publications related to the WCS.[6] The study is in two parts, each part centering on one of the learning methods described in the previous section. The procedure is the same

[6] As an anonyous referee remarked, it would be interesting to conduct the simulations to be presented in the following also for the full set of 1,625 Munsell chips, especially given that the chips for the WCS were all selected precisely because they all show highly saturated colors, thus raising the question whether we would obtain the same results if also chips showing less saturated colors were included. A practical problem here is that we currently lack data on how people would carve up the full set of Munsell chips into the eleven basic color categories. Jraissati and Douven (2018) did use the full set in their study, but presented their participants with a free-naming task, meaning that they could use any name they liked for any of the chips they were shown.

in both parts and involves simulating color concept learning for 10,000 randomly generated systems of nonnatural color concepts and comparing the accuracy achieved in the learning process with that achieved in a simulation of color concept learning for the system of natural color concepts. Specifically, each part of our study consisted of the following steps:

1. Split up the materials into the natural color concepts and sample randomly from the chips in each concept, where the number sampled from each concept is greater than 0 but otherwise random (sampled uniformly from the number of chips in the given concept), thus obtaining a set of pairs $\left\{\left\langle\langle L^*_c, u^*_c, v^*_c\rangle, C_c\right\rangle\right\}_{c \in s}$ representing the CIELUV coordinates of Munsell chip c in sample s as well as its label C_c indicating the color concept under which it falls.[7]
2. Apply the learning method (APM or KNN) to the set of pairs and on that basis predict the labeling of the chips in the materials that were *not* sampled.
3. Measure the accuracy of the prediction—the best guess of what the system of concepts is—by comparing it with the actual system of natural color concepts.
4. Repeat the foregoing steps 50 times.
5. Run the same procedure for 10,000 systems of color concepts, each obtained by randomly picking 10 or 11 (see below) points in CIELUV space as prototypes and using these to generate a Voronoi tessellation of the space (so meaning that now the 320 chips in the materials are split up according to the random system).
6. Compare the accuracy scores obtained for the natural system with those obtained for the nonnatural systems.

For KNN, we followed the customary recommendation of setting k equal to the square root of the size of the sample, rounded to the nearest integer. Of course, given that the size of the sample was always random, k could have a different value in every simulation. Also, given a sample s of chips and a chip c not in the sample, the weight each of the k chips in s nearest to c had in voting on c's label was set equal to the inverse of their squared Euclidean distance (as measured in CIELUV space) to c.

Both parts of the study were conducted twice over, once for 10 color concepts and once for 11. This was because some participants in color naming studies for English and French used all basic color terms in describing the colors of the 320 Munsell chips used for our study while other participants used only 10, leaving out "gray" (see Claidière, Jraissati, & Chevallier, 2008). Given that in the same color naming studies there was quite some interpersonal variability in how the chips were named, we used the k-means clustering algorithm to obtain a kind of objective approximation of the natural color concepts (see Douven, 2017). The results are shown in the first two rows of Figure 7, while the bottom row of that figure shows an example of a randomly generated system of 10 color concepts.[8]

Accuracy was measured in terms of Normalized Mutual Information (NMI), which quantifies the mutual dependence of two variables. In our case, in which we are comparing clustering results, the two variables are two different clusterings, and the NMI quantifies how similar they are. That the measure is *normalized* means that its values are between 0 and 1, with 0 indicating no shared information (the clusterings are completely dissimilar) and 1 indicating perfect agreement (the clusterings are identical), making NMI values easy to interpret and compare. Another benefit of using the NMI measure for comparing clusterings is that it takes into account both the amount of shared information and the total

[7] Note that because a *random* number of chips was sampled randomly from each concept, there was no fixed sample size. It was empirically determined that the sample size was, on average, 165.02 ($\pm$ 31.98).

[8] The nonnatural concepts have the color resulting from averaging the CIELUV coordinates of the chips that fall under the concept. Also, all color concepts that occurred in the study—both the natural ones and the nonnatural ones—were convex, by construction. That the charts shown in Figure 6 might suggest otherwise is simply because these charts are not meant to represent a perceptual color space.

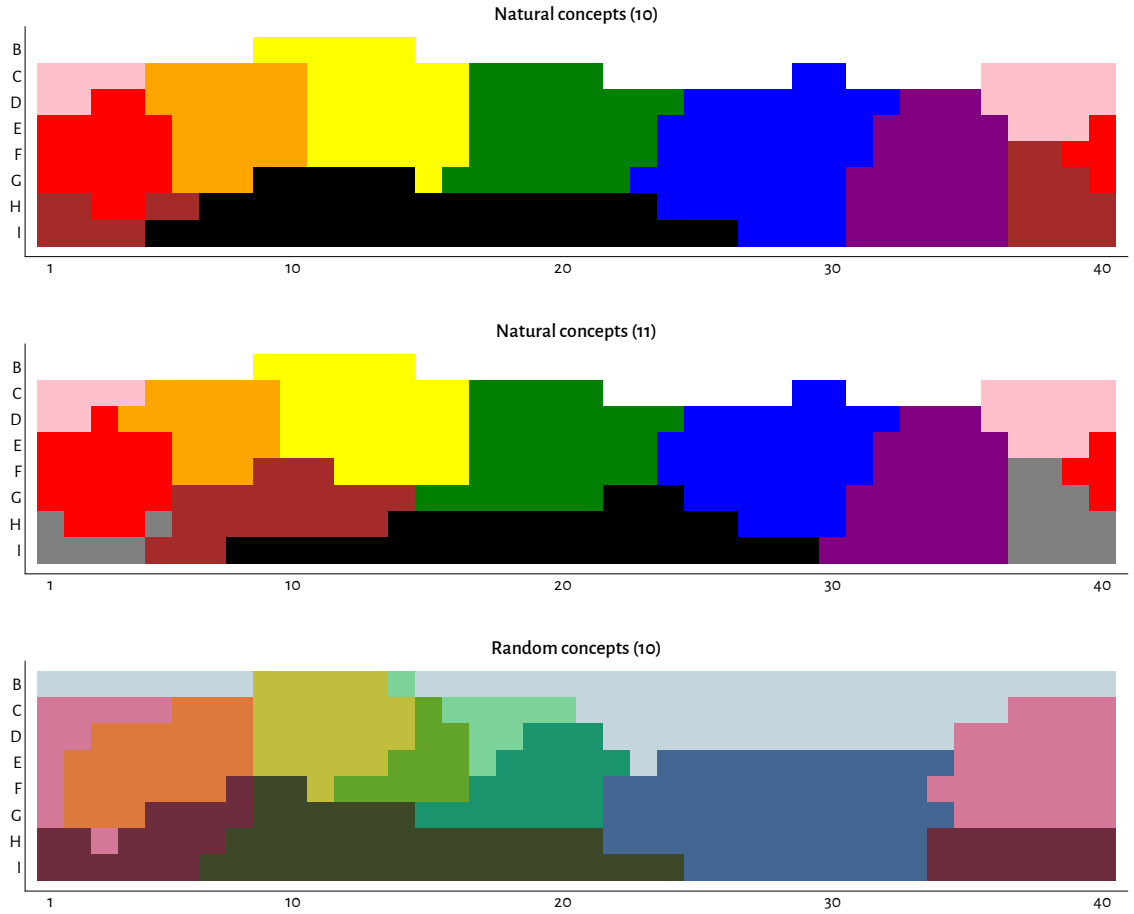

Figure 7: Clusterings of the 320 chromatic WCS chips into 10 and 11 natural color concepts (top and middle row) and clustering into 10 concepts according to randomly generated system of color concepts (bottom row; clusters have been colored by taking the center of mass of the CIELUV coordinates of the chips in the cluster).

information in each clustering, in contrast to most of the simpler measures (see, e.g., Pfitzer, Leibbrandt, & Powers, 2009).

As mentioned above, we ran the study both for 10 and for 11 color concepts. As also explained, in each run, we sampled 50 times from each system of color concepts (the natural one plus the 10,000 randomly generated ones), the sample each time yielding a random number of examples from each concept in the system. And applying the learning method (APM in one part of the study, KNN in the other) gave us a clustering of the 320 chips from Figure 6, which could then be compared via the NMI measure with what the clustering was according to the system from which the sample was taken. Thus, for each system of concepts, we obtained 50 NMI scores per learning method, which we summarized via their means and standard deviations.

Learning via the APM method yielded an average NMI score of .92 (± 0.04; range = [.83, .97]) for the system of 10 natural color concepts (shown in the top row of Fig. 7) and also an average NMI score of .92 (± 0.03; range = [.84, .97]) for the system of 11 natural color concepts (middle row of Fig. 7). The same learning method achieved, on average (i.e., averaged over the 10,000 systems), an average (i.e., averaged over the 50 samplings per system) NMI score of .72 (± 0.04; range = [.53, .85]) for the nonnatural systems with 10 color concepts and an average NMI score of .73 (± 0.03; range = [.57, .84]) for the nonnatural systems with 11 color concepts. The left panel of Figure 8 shows a density plot of the average

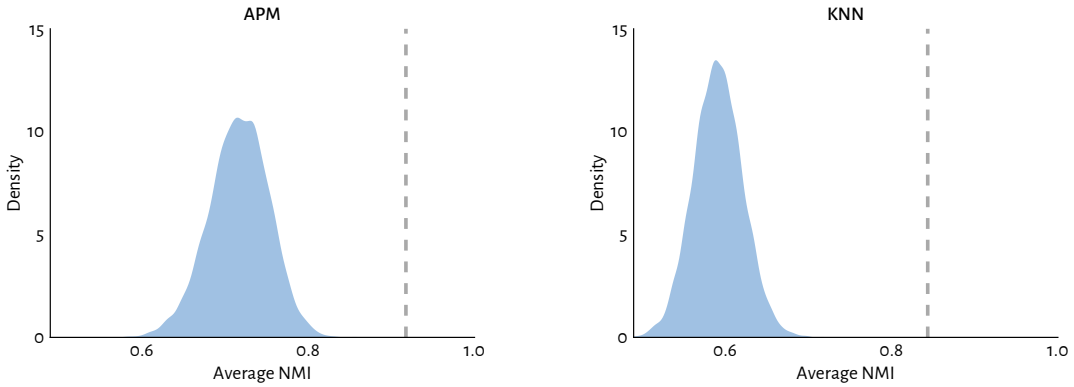

Figure 8: Density plots of average NMI scores for 10-concept systems obtained for the APM method (left) and the KNN method (right), each panel also showing the average NMI score for the corresponding system of natural color concepts (dashed line).

NMI scores for 10-concept nonnatural systems, with the average score for the natural system shown as a dashed line. The corresponding plot for 11-concept systems looks virtually identical and is not shown here.

For the KNN method, we registered an average NMI score of .84 (± 0.06; range = [.70, .93]) for the system of 10 natural color concepts and an average NMI score of .83 (± 0.06; range = [.72, .96]) for the system of 11 natural color concepts. The average score for the 10-concept nonnatural systems was .59 (± 0.03; range = [.47, .70]), on average, while for the 11-concept nonnatural systems it was .61 (± 0.03; range = [.48, .71]). The right panel of Figure 8 shows again the density plot of these scores for the 10-concept systems, together with the average score for the system with 10 natural color concepts. Here, too, the corresponding 11-concept-systems plot is omitted, given that it looks almost the same.

For neither method need we run a statistical test to see that naturalness has a significant effect on learning: the accuracy with which the algorithms were able to predict under which concepts the chips fall that were not sampled was significantly better if the training sample came from the natural concepts than when it came from some system of nonnatural color concepts. While that is a positive for Learnability, it will be recalled from the previous section that Douven and Gärdenfors (2020) left open the possibility that their list of criteria contained redundancies. And in the same section, it was mentioned that Douven (2019) found support for the claim that the system of natural color concepts strikes a near to optimal balance between Representation and Contrast, which are meant to jointly guarantee that prototypes can be placed in a space such that they are (i) good representatives of the other items falling under the concept they are the prototype of and (ii) easily distinguishable from each other. It is thus legitimate to ask whether the fact that the algorithms learned color concepts most accurately when those concepts were the natural ones could be because those concepts satisfy criteria on Douven and Gärdenfors' (2020) list other than Learnability.

To address this question, specifically to examine whether Contrast and Representation might already predict learning accuracy, four linear regressions were conducted, one for each combination of number of concepts (10 or 11) and learning algorithm (APM or KNN). For this, we first calculated for each of the 10,000 nonnatural 10-concept systems as well as for each of the 10,000 nonnatural 11-concept systems their Contrast and Representation scores, following Douven's (2019) proposal to operationalize Contrast as the sum of the Euclidean distances among all the generating points in CIELUV space and Representation as the sum of the distances of those from the center of gravity of the concept that

they represent (where, in this case, the center of gravity is calculated by taking the average L^*-, the average u^*-, and the average v^*-coordinate of those of the 1,625 Munsell chips that fall under the given concept). These Contrast and Representation scores then did duty as the predictors in the said regression analyses, all of which had average NMI score as dependent variable.

For the APM method, the β-coefficients for Contrast and Representation were, respectively, 0.13 ($t = 14.80, p < .0001$) and −0.50 ($t = 58.57, p < .0001$) for the 10-concept systems and 0.12 ($t = 14.23$, $p < .0001$) and −0.48 ($t = 55.12, p < .0001$) for the 11-concept systems. For the KNN method, the corresponding β-coefficients are −0.01 ($t = 0.57, p = .5674$) and −0.23 ($t = 23.32, p < .0001$) for the 10-concept systems and −0.02 ($t = 1.61, p = .1066$) and −0.21 ($t = 21.14, p < .0001$) for the 11-concept systems. Thus, for the APM method, Contrast and Representation were both highly significant predictors for accuracy and hence for the learnability of conceptual systems. Also note that they predicted accuracy in the way one would expect: an increase in Contrast, as measured by an increase of the sum of distances among the points generating the system, led, on average, to a significant increase of accuracy, while an increase in Representation, as measured by a *decrease* of the sum of distances of the chips falling under a concept to the center of gravity of that concept, led to a significant increase of accuracy as well. For the KNN method, we find that only Representation is a highly significant predictor of the accuracy achieved by the learning method; Contrast is not significant.

Our primary aim was to shed light on the relation between naturalness and learnability. However, the learning methods we implemented computationally are of independent interest, so it is worth briefly comparing their performance in the above study. In the toy example we used previously to illustrate the methods, one got the impression that the KNN method did somewhat better than the APM method. Comparing the average accuracy scores reported above, and also the two panels in Figure 8, gives a different impression. Indeed, in our study the APM method did significantly better than the KNN method, as was confirmed by running four t-tests on the results obtained by the two methods for each of the following: the 10-concept natural system, the 10-concept nonnatural systems, the 11-concept natural system, and the 11-concept nonnatural systems. For the natural 10-concept system, we had $t(98) = 7.67, p < .0001$, with a value for Cohen's d of 1.51 (which counts as large); for the nonnatural 10-concept systems, the results were $t(19998) = 266.84, p < .0001$, Cohen's $d = 3.77$; for the natural 11-concept system, $t(98) = 8.94, p < .0001$, Cohen's $d = 1.76$; and finally, for the nonnatural 11-concept systems, we had $t(19998) = 273.42, p < .0001$, with Cohen's $d = 3.87$.

4 Conclusion

Douven and Gärdenfors (2020) hypothesized that what distinguishes natural from nonnatural concepts is that the former, but not the latter, are represented by the cells of an optimally partitioned similarity space. They defined optimality in terms of a list of engineering constraints. In this paper, we focused on the constraint that optimally partitioned similarity spaces result in easily learnable concepts, a correlation for which empirical evidence was still missing, in contrast to most of the other constraints on the list. This motivated the study reported in this paper.

Our study implemented computationally two plausible models of concept learning, one proposed by Gärdenfors, the other closely connected to Nosofsky's Generalized Context Model. We applied these models to the learning of color concepts in order to examine whether natural color concepts are learned more readily than nonnatural ones. The outcomes of the study confirmed this hypothesis for both employed models, thus lending empirical support to the notion that learnability is a distinctive characteristic of natural concepts.

The results were still not entirely positive for Learnability. While learnability appears to be a feature of naturalness, the learnability of systems of natural concepts may, as far as the results from our study go, be accounted for already by the fact that these systems satisfy other criteria on Douven and Gärdenfors' (2020) list, in particular, Contrast and Representation. As mentioned, this would not necessarily be inconsistent with these authors' proposal, given that they explicitly left open the possibility that their list contains redundancies.

Our study had some clear limitations. First, it only involves *computational* models of learning, which will not fully reflect human learning. For reasons mentioned in Section 2.3, it is not straightforward to compare the learning of natural concepts with the learning of nonnatural concepts by human participants, given that the participants will already be familiar with the former but not with the latter. This is not to say that such experiments are impossible, and we hope that the present results, indicating a clear connection between learnability and naturalness, will inspire cognitive psychologists to come up with a paradigm that will allow the connection to be tested in humans in a way that is methodologically sound. A second limitation concerns the fact that our evidence is restricted to color concepts. In principle, it is easy to rerun the study using conceptual spaces other than color space. In practice, this is more difficult, because the number of conceptual spaces that can be downloaded or are otherwise available for running the kind of simulations we conducted is extremely limited, and we actually know of no conceptual space that is as well validated as CIELUV and CIELAB space. This situation may change, however, given that the conceptual spaces framework is an active area of research, and given also that researchers are now commonly making their data available for others to work with. Once more conceptual spaces are available for experimentation, it will be interesting to see not only whether the link between learnability and naturalness that we found to exist for color concepts generalizes, but also (if it does) whether that link can be more generally accounted for in terms of satisfaction of the Contrast and Representation criteria.[9]

References

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science*, *301*, 1866–1867.

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

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–517.

Bendifallah, L., Abbou, J., Douven, I., & Burnett, H. (2023). Conceptual spaces for conceptual engineering? Feminism as a case study. *Review of Philosophy and Psychology*, in press.

Berlin, B. & Kay, P. (1969). *Basic color terms*. Stanford CA: CSLI Publications.

Carnap, R. (1980). A basic system of inductive logic, part II. In R. C. Jeffrey (ed.), *Studies in inductive logic and probability* (pp. 7–155). Berkeley CA: University of California Press.

Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE*, *8*, e73289, https://doi.org/10.1371/journal.pone.0073289.

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

Churchland, P. M. (2012). *Plato's camera*. Cambridge MA: MIT Press.

Claidière, N., Jraissati, Y., & Chevallier, C. (2008). A colour sorting task reveals the limits of the universalist/relativist dichotomy: Colour categories can be both language specific and perceptual. *Journal of Cognition and Culture*, *8*, 211–233.

[9]I am grateful to Christopher von Bülow and to two anonymous referees for this journal for valuable comments on a previous version.

Cook, R. S., Kay, P., & Regier, T. (2005). The World Color Survey database: History and use. In H. Cohen & C. Lefebvre (eds.), *Handbook of categorization in cognitive science* (pp. 223–242). Amsterdam: Elsevier.

Deauvieau, J., Penissat, É., Brousse, C., & Jayet, C. (2014). Les catégorisations ordinaires de l'espace social français. *Revue Française de Sociologie*, *55*, 411–457.

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

Decock, L., Douven, I., & Sznajder, M. (2016). A geometric principle of indifference. *Journal of Applied Logic*, *19*, 54–70.

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95.

Douven, I. (2019). Putting prototypes in place. *Cognition*, *193*, 104007.

Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, *42*, 137–160.

Douven, I., Elqayam, S., Gärdenfors, P., & Mirabile, P. (2022). Conceptual spaces and the strength of similarity-based arguments. *Cognition*, *218*, 104951.

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

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science*, *41*, 686–722.

Fairchild, M. D. (2013). *Color appearance models*. Hoboken NJ: Wiley.

Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

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

Gärdenfors, P. (2001). Concept learning: A geometrical model. *Proceedings of the Aristotelian Society*, *101*, 163–183.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge MA: MIT Press.

Gärdenfors, P., Jost, J., & Warglien, M. (2018). From actions to effects: Three constraints on event mappings. *Frontiers in Psychology*, *9*, 1391.

Gärdenfors, P. & Warglien, M. (2012). Using concept spaces to model actions and events. *Journal of Semantics*, *29*, 487–519.

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

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Hardin, C. L. (1988). *Color for philosophers: Unweaving the rainbow*. Indianapolis IN: Hackett.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, *8*, 176–184.

Jameson, K. & D'Andrade, R. (1997). It's not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (eds.), *Color categories in thought and language* (pp. 295–319). Cambridge: Cambridge University Press.

Jraissati, Y. & Douven, I. (2018). Delving deeper into color space. *i-Perception*, *9*, 1–27, https://doi.org/10.1177/2041669518792062.

Kaufman, L. & Rousseeuw, P. J. (2005). *Finding groups in data*. Hoboken NJ: Wiley.

Kemp, C. & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kripke, S. A. (1980). *Naming and necessity*. Cambridge MA: Harvard University Press.

Malacara, D. (2002). *Color vision and colorimetry: Theory and applications*. Bellingham WA: SPIE Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt and Co.

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*, e2025993118.

Munsell, A. H. (1941). *A color notation: An illustrated system defining all colors and their relations*. Boston MA: The Hoffman Brothers Co.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge: Cambridge University Press.

Oaksford, M. & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Okabe, A., Boots, B., Sugihara, K., & Chiu, S. N. (2000). *Spatial tessellations* (2nd ed.). New York: Wiley.

Peterson, M. (2017). *The ethics of technology: A geometric analysis of five moral principles*. Oxford: Oxford University Press.

Petitot, J. (1989). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics*, *15*, 25–71.

Pfitzer, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures of pairs of clusterings. *Knowledge and Information Systems*, *19*, 361–394.

Posner, M. I., Goldsmith, R., & Welton, K. E., Jr. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology*, *73*, 28–38.

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

Posner, M. I. & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304–308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences USA*, *104*, 1436–1441.

Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, *85*, 884–892.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (eds.), *Cognition and categorization* (pp. 27–48). Hillsdale NJ: Erlbaum.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 491–502.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, *69*, 1996–2019.

Verheyen, S. & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science*, *42*, 2250–2286.

Verheyen, S. & Peterson, M. (2021). Can we use conceptual spaces to model moral principles? *Review of Philosophy and Psychology*, *12*, 373–395.

Xu, Y. & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1802–1807). Austin TX: Cognitive Science Society.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*, 2081–2094.

Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., & Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, lzac001.