The Rationality of Vagueness

Igor Douven

Sciences, Normes, Décision (CNRS)

Sorbonne University

Abstract: Vagueness is often regarded as a kind of defect of our language or of our thinking. This paper portrays vagueness as the natural outcome of applying a number of rationality principles to the cognitive domain. Given our physical and cognitive makeup, and given the way the world is, applying those principles to conceptualization predicts not only the concepts that are actually in use, but also their vagueness, and how and when their vagueness manifests itself (insofar as the concepts are vague).

Keywords: concepts, design, prototypes, rationality, vagueness.

Introduction

Vagueness is often characterized in terms of the existence of so-called borderline cases. To say that a predicate is vague is to say that there are cases to which the predicate neither clearly applies nor clearly fails to apply. Such a situation of semantic indecision can easily appear unfortunate, and the study of vagueness may seem to some like studying a pathology of language. Too bad linguists and logicians were not around when our ancestors stumbled upon the possibility of verbal communication, tens of thousands of years ago; then we would not now be saddled with a language that

more often than not resists formalization in the systems taught in introductory and even not so introductory courses on logic.

On reflection, however, one realizes that this view on the matter is too harsh. Linguists and logicians could have done little or nothing to enhance human discriminatory capacities, nor could they have made the world "gappier." Thus, researchers interested in vagueness tend to conceive of it as an unfortunate phenomenon, but also as one that is inevitable, given our physical and cognitive makeup and the way the world is.

In this paper, I take a less apologetic approach to vagueness. I propose that the occurrence of vagueness is not just excusable, or understandable, or inevitable; the "structure" of vagueness—by which I mean, which predicates are vague and where their borderlines are found—can be seen as following from principles of rational design applied in the cognitive realm.

The design principles to be discussed actually concern *concepts*—the cognitive correlates of predicates—which in this paper are taken to be the primary bearers of vagueness. In a popular framework for modeling concepts, concepts reside in so-called similarity spaces (see below). The main claim is that if such spaces are optimally designed, then we should *expect* to see vagueness arise when and where, as a matter of empirical fact, it does arise.

Section 2 briefly describes the conceptual spaces framework, which is an increasingly popular format for theorizing about concepts and will serve as a background for the proposals made in this paper. I should say at the outset that, while popular, it is still

unclear exactly how broadly applicable the framework really is. So far, it has been used with great success for the representation of perceptual concepts, and important progress has recently been made in applying the framework to more abstract—especially scientific—concepts. We should not be surprised, however, if it turns out that the framework is limited in terms of the types of concepts it can represent. I set this issue aside, however, for the remainder of this paper.

In Section 3, I outline my favorite theory of vagueness, which relies heavily on the machinery of the conceptual spaces framework. It relies just as heavily on a proposal made in Gärdenfors (2000), which gives pride of place to prototypes and so-called Voronoi tessellations for arriving at specific structures of spaces. I show how this proposal yields a formally precise account of vagueness, if we are willing to take on board a seemingly uncontroversial assumption about prototypes and slightly modify the technique of Voronoi tessellations.

Gärdenfors' proposal, and concomitantly the account of vagueness based upon it, leaves some important questions unanswered. Most notably, while Gärdenfors stresses from the outset that his proposal is meant to pertain to *natural* concepts—concepts that have, or could plausibly have, a role in our everyday thinking and theorizing, and that therefore we might care to name in our language—he admits that he has probably not managed to characterize the relevant notion of naturalness in a fully satisfactory way. And indeed, on his original account, too many concepts qualify as natural, including ones that pre-theoretically are clearly *not* natural.

Although not exactly presented in this way, Gärdenfors' attempt to differentiate natural concepts from non-natural ones appeals to a kind of rationality principle (or design principle). Natural concepts—the claim runs—are those that resulted from carving up a conceptual space in the most rational manner. As mentioned, this proposal does not quite do the job. But as argued in Douven and Gärdenfors (2018), this is not because it is a mistake to try to connect naturalness with rational design, but rather because not enough rationality principles are invoked to determine a carving-up of any given space. Douven and Gärdenfors list a number of rationality principles that—they argue—together might just suffice to yield carvings-up of conceptual spaces that result in truly natural concepts. Their claim, in slogan form, is that natural concepts are concepts represented by the cells of an optimally (i.e., most rationally) partitioned similarity space. This proposal is summarized in section 4.

Douven and Gärdenfors (2018) explicitly state that they leave aside the issue of how to accommodate vagueness within their new proposal. That issue is taken up in section 5 of this paper. Broadly speaking, the answer will be that the account of vagueness mentioned above is not only consistent with the new proposal, but that the two fit together very naturally, and can be seen as both being part of a three-step theory of conceptualization, with rationality playing a key role at every stage.

The conceptual spaces framework

In the 1960s and 1970s, cognitive psychologists started using so-called dimensionality reduction techniques, such as principal component analysis and multi-dimensional scaling, to give low-dimensional representations of people's similarity judgments

concerning (often) high-dimensional data (see, e.g., Krumhansl 1978; Borg and Groenen 2010). The outcomes of these procedures are commonly referred to as "similarity spaces."

To illustrate, consider the example of faces. Faces can vary along multiple dimensions: height, width, shape of the nose, eyes, eyebrows, color of the skin, and on and on. When two faces strike us as looking very similar, they probably differ very little along all the relevant dimensions, or at least along most of them. One would expect, then, that if we are to have a formal model that allows us to predict with some accuracy whether or not people will judge two given faces to be similar, the model must have separate parameters for all or most of the dimensions along which faces can vary.

Surprisingly, this turns out not to be the case. Low-dimensional models—so-called "face spaces"—have been developed that tend to yield precise predictions about people's similarity judgments of faces. Indeed there exist three-dimensional face spaces that have proven highly successful in this respect; see Valentine, Lewis, and Hills (2016). What this means is that faces can be located in a three-dimensional space in such a way that measuring the distance in the space between any pair of faces gives a good indication of how similar to each other those faces will be deemed. Specifically, the prediction—which in the meantime has been supported by a wealth of evidence—is that the closer two faces are in the space, the more similar they will be judged to be.[1] The same is known to hold for colors, odors, sounds, tastes, various

[1] Different people will in general give somewhat different similarity judgments. For that reason, one might want to construct a similarity space for each person

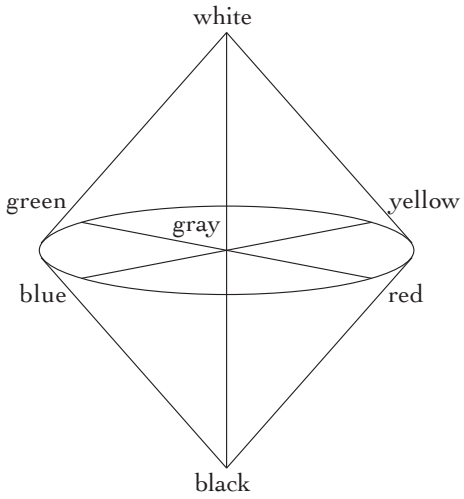

Figure 1: An approximate representation of color space.

types of actions, and various types of shapes. Accordingly, we find in the scientific literature references to color spaces, odor spaces, taste spaces, vowel spaces, consonant spaces, action spaces, shape spaces, and others.

In the first instance, these are all similarity spaces: geometric models to predict people's similarity judgments with regard to whichever respect the given space is meant to capture. But most of these spaces can be, and have been, further developed into *conceptual* spaces. That requires designating particular regions in a similarity space as representing specific concepts. So, for instance, SWEET is a specific region in taste space, and VASE is a specific region in a particular shape space. Concepts, in this view, are geometric objects in similarity spaces.

---

individually. However, that is often impractical, and non-individualized spaces—spaces typically constructed on the basis of similarity judgments coming from different people, and not pretending to represent anyone's personal similarity space—tend to work well enough for many purposes; see Douven (2016).

To make this more concrete, consider the representation of color space given in Figure 1. Color space is three-dimensional, with hue being represented by the color circle that one sees in the middle of the space (this goes through all the colors in the rainbow), lightness being represented by the vertical axis (which goes from white to black through all shades of gray), and saturation (the "depth" of a color) being represented by distance from the vertical axis.[2] The labels in the figure are not to be interpreted as indicating the locations where the named colors are to be found in the space. Rather, they give an indication of roughly where the corresponding color *concepts* are to be found, which are regions—*sets* of points—in this space, not single points.

If concepts are regions in a similarity space, is every region in a similarity space a concept? Given that "concept" is a somewhat technical notion, we are probably free to answer the question in the positive. But what this question means to ask, of course, is whether every such region corresponds to a concept that might be of practical relevance to us, one that we would have real use for and that we would want to have a predicate for in our language. Consider that, mathematically speaking, any connected set of points in color space is a region in that space. And surely the vast majority of regions—"almost all," in the technical sense of measure theory—are not worth having

[2] "Distance" in color space is usually taken to mean "Euclidean distance." Not all conceptual spaces are Euclidean spaces, but the issue of which metric is appropriate for which space need not detain us here. For details, see Gärdenfors (2000, Ch. 3) or Douven (2016).

a name for or otherwise singling out; the vast majority strike us as corresponding to very unnatural and gerrymandered concepts at best.

This raises the further question of what distinguishes those concepts that pre-theoretically strike us as natural from those that do not do so. In Gärdenfors (2000, 71), we find the following tentative proposal:

> CRITERION P
>
> A *natural concept* is a convex region of a conceptual space.

To say that a region is convex is to say that, for any two points in the region, the line segment connecting those points lies in its entirety in the region as well. Gärdenfors (p. 70) defends CRITERION P as "a principle of *cognitive economy*; handling convex sets puts less strain on learning, on your memory, and on your processing capacities than working with arbitrarily shaped regions." He also notes that there is already a fair amount of evidence supporting the thought that the concepts we use do correspond to convex regions of similarity spaces.

However, none of this helps in answering the question of whether convexity is also a *sufficient* condition for naturalness. Gärdenfors (*ibid*.) is not entirely convinced that it is, and explicitly remarks that he "only view[s] the criterion as a *necessary* but perhaps not sufficient condition on a natural property." In fact, the word "perhaps" here understates the extent to which CRITERION P falls short of capturing our intuitive notion of naturalness vis-à-vis concepts.

To illustrate the problem, consider again the color space shown in Figure 1 and note that any plane intersecting that space divides it into two regions, both of which are convex. Moreover, it can be shown that whenever the intersection of several convex sets is taken, the result is itself convex (Douven et al. 2013, 147). Hence, carving up color space by means of any random pick of planes intersecting that space will result in a partitioning of the space with only convex cells. This is so even though some of these cells could contain shades that any normal observer would deem as falling under different color concepts. If convexity were sufficient for naturalness, we would nevertheless be forced to say that each of those cells corresponds to a natural concept.

We leave the question of the characterization of naturalness open for now, but will return to it in section 4. There it will be argued that, while CRITERION P on its own is not capable of defining naturalness, there are other principles very much like it, both in spirit and in motivation, such that together these principles stand a good chance of doing the job. First, though, we turn to an account of vagueness that was built on the conceptual spaces framework as presented in this section.[3]

Conceptual spaces and vagueness

[3] Note that, to date, it is still largely unknown how broadly applicable the conceptual spaces framework really is. It has been successfully used to model perceptual concepts and to some extent also more abstract concepts (e.g., Gärdenfors and Zenker 2013). Whatever the framework's limitations may be, it is clear that the current proposal, by relying so heavily on this framework, will face the same limitations.

If concepts are to be conceived as regions in similarity spaces, then how are we to accommodate the fact that many concepts are vague? If RED is a region in color space, this suggests that, for any shade, we can say that it is either red (it is represented by a point lying in the region representing RED) or not red. But that runs counter to everyday experience: there are shades that we cannot classify so well, that strike us as being neither quite red nor quite some other color. It would seem that we need something like *blurry* regions to represent color concepts, or other vague concepts. However, mathematically it is not clear what blurry regions might amount to.

Douven et al. (2013) present a formal account of vagueness cast within the conceptual spaces approach. The account actually builds on more than the framework of conceptual spaces by combining Gärdenfors' (2000) proposed combination of that framework with prototype theory (Rosch 1973) and the mathematical technique of Voronoi tessellations.

Not all readers will be familiar with Voronoi tessellations, but the basic idea behind them is very simple. Let S be a space and $P = \{p_1, \ldots, p_n\}$ a set of points in S. Then the Voronoi tessellation of S generated by P is the set of cells $\{c_1, \ldots, c_n\}$ with each c_i containing all and only points in S that are at least as close to p_i as they are to p_j, for all $j \neq i$. (For an example, see Figure 2, left panel.)

Suppose we are given locations in color space of the prototypical colors. Then we can let those locations act as points generating a Voronoi tessellation of color space, which would gather the red shades around the RED prototype, the green shades around the GREEN prototype, and so on. An interesting observation Gärdenfors makes is that this

would automatically yield convex concepts, given that, necessarily, any cell of a Voronoi tessellation is convex (Okabe et al. 2000, 58).[4]

It might be thought that this answers the previous question regarding what makes a concept natural, the putative answer being that it is not convex concepts *per se* that are natural, but convex concepts resulting from a Voronoi tessellation generated by the relevant prototypes. But this cannot be right. After all, prototypes are not supposed to be prior to concepts. Rather, a prototype is said to be the best representative, the most typical instance, of a concept (Rosch 1973, 330; see also Rosch 1975). Thus, the concept must be there before the prototype can come into existence.

At first, it might also seem that Gärdenfors' add-ons to the conceptual spaces framework answer, by themselves, the earlier question about vagueness. Looking again at the left panel of Figure 2, we see that on Gärdenfors' account concepts clearly have borderline cases: these are what constitute the borderlines between the regions, the points that are equally far from the prototypes of at least two different concepts. However, proposing this as an answer to the question of where vagueness comes from would lead to what Douven et al. (2013) have termed "the thickness problem." Open on your computer any drawing program that you are familiar with and go to the color menu. Then move the sliders (or dials, or whatever exactly the interface uses) to set the

[4] This result holds for Euclidean spaces, such as color space and many other similarity spaces, but not for all spaces. Again, these details are left aside here.

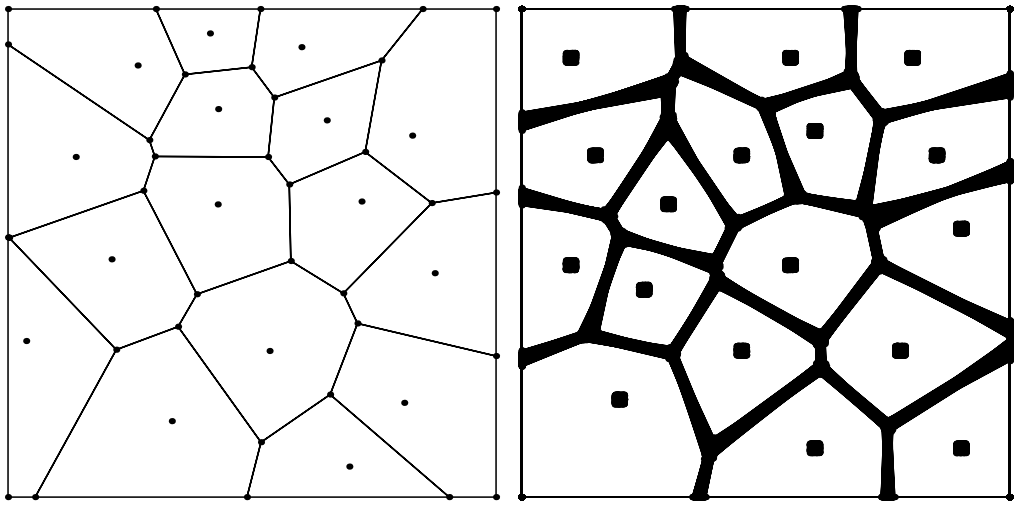

Figure 2: Simple Voronoi tessellation (left) and collated Voronoi tessellation (right) on a two-dimensional space.

color coordinates in whichever color space the program uses (probably RGB space[5]) so that the resulting color strikes you as being borderline red/orange. Then, *very* probably, moving any of the sliders just a tiny bit in either direction will result in a color that *still* strikes you as being borderline red/orange. That would not be possible, however, if we were to build our account of vagueness directly on Gärdenfors' proposal. For then "almost all" (again in the measure-theoretic sense of that expression) small adjustments of the sliders would have to yield either a clear case of red or a clear case of orange. Put graphically, borderline cases tend to be surrounded by borderline cases, which would seem to require border *regions* instead of border*lines*.

[5] To forestall misunderstanding, it is to be noted that RGB space and other color spaces frequently used in applications software are not *similarity* spaces; they are not intended to represent human similarity judgments concerning color, and they are also as a matter of fact known *not* to represent such judgments.

Douven et al. (2013) have come up with an arguably non-ad hoc way of equipping conceptual spaces with such border regions. The non-ad hocness derives from the fact that their proposal starts from an independently plausible observation and then extends the technique of Voronoi tessellations in a natural way immediately motivated by that observation.

The observation is that, for many concepts, it is at best a convenient fiction to suppose that it has exactly one prototype, one best representative. Open your drawing program again, and adjust the color coordinate sliders until the color you see on the canvas strikes you as typically red. Now ever so slightly move one of the sliders. Whichever slider you move, and whether you move it to the left or to the right, chances are that the color on the canvas will still strike you as typically red. What goes for red goes for any other color, and indeed, *mutatis mutandis*, for most or even all other perceptual concepts. And it is not, or not necessarily, a matter of discriminability. I have no difficulty telling apart the taste of vanilla ice cream and that of strawberry ice cream, and yet both strike me as typical for the concept SWEET. (I am talking here of our local ice cream; where you live, the strawberries may not be as sweet.)

It thus appears that, for many conceptual spaces, we are better off thinking of them as having prototypical *regions* rather than prototypical *points*. If this is accepted, however, then what remains of the other part of Gärdenfors' proposal, the technique of Voronoi tessellations? Standard Voronoi tessellations have so-called generator *points*, not generator *regions*.

Douven et al. (2013) argue that there is still a use for Voronoi tessellations, provided that we slightly change the technique. In a nutshell, the modification they propose is this: Consider the set *V* of all possible selections of one single point from each prototypical region in a space and note that every member of that set generates a Voronoi tessellation on the space. Then, instead of looking at those tessellations one at a time, imagine how they are all simultaneously present in the space. Call the resulting construct "the collated Voronoi tessellation generated by *V*" on the given space. Douven et al. (2013) show that this construct still partitions the space into convex regions. For the purposes of explaining vagueness in the conceptual spaces framework, this constitutes an important step forward in that collated Voronoi tessellations have "thick" boundary regions, meaning that "almost all" borderline cases are surrounded by other borderline cases. For an illustration, see the right panel of Figure 2.

Even if this picture yields a phenomenologically more plausible account of borderline cases and therefore of vagueness, the fact that, as the collated Voronoi tessellation in Figure 2 shows, borderline cases are still sharply delineated from non-borderline cases (i.e., clear cases) could give cause for concern, simply because it seems unrealistic. We could again use a drawing program to verify that we do not tend to experience sudden transitions from, say, borderline cases of red to clear cases of red: where the clear cases end and the borderline cases begin is, for most or all vague concepts, itself a vague matter. In other words, the new proposal still seems to face the so-called problem of higher-order vagueness.

This problem can actually be dealt with in more than one way. Douven et al. (2013) account for it by reference to the imprecision of psychological metrics. But a more

satisfactory solution to the problem is given in Decock and Douven (2013), which develops a measure of graded membership, and which explains our experience of higher-order vagueness by showing that there are no abrupt transitions from clear cases to borderline cases, given that the borderline cases neighboring the clear cases fall under the given concept to a degree still very close to 1 (and thus very close to the degree of any of the clear cases; see also Douven and Decock 2017).

To end this section, I mention that Douven et al. (2013) do note that their proposal has clear empirical content, and that in recent years the proposal has been put to the test and so far has passed all tests with flying colors. See Douven (2016), Douven et al. (2017), and Verheyen and Égré (2018).

Naturalness and Design

It is all well and good to have an account of vagueness in the conceptual spaces framework, even one that already enjoys some empirical support, but there remains a cloud over the account as long as it has not been shown that the conceptual spaces framework itself is tenable. As we saw in section 2, a question that looms large over the framework is whether it has the resources to differentiate between natural and non-natural concepts. If ultimately we had to admit that, given this framework, any convex region in a conceptual space, however bizarre and gerrymandered it may appear pre-theoretically, represents a concept as natural as, say, GREEN, we would want to reject the framework, and the account of vagueness described in the previous section would go out of the window with it.

Douven and Gärdenfors (2018) have recently addressed the question of how to characterize natural concepts within the conceptual spaces framework. To do so, they take their cue from work on design thinking in theoretical biology. Biologists have been arguing that many traits and features of organisms, and also many structures of cells or even structures of molecules within cells, are best understood from a design perspective, specifically as being the best solutions to engineering problems. Douven and Gärdenfors argue that this same design thinking also applies at the cognitive level and promises to answer the question of what natural concepts are. Notably, their proposal is that natural concepts are concepts represented by the cells of an *optimally designed* conceptual space. Most of their paper then consists of (i) spelling out the notion of optimal design as it pertains to conceptual spaces, and (ii) illustrating and supporting the proposal by discussing a number of recent results concerning categorization from the cognitive sciences.[6]

[6] Douven and Gärdenfors' approach is somewhat akin to Anderson's (1990, Ch. 3) rational analysis of categorization, which tries to understand categorization as the outcome of a procedure aimed at maximizing predictive success and in particular also invokes optimality considerations to identify what Anderson calls "basic level categories." Apart from the fact that Anderson is not concerned with conceptual spaces, another important difference is that he takes the world to be carved up independently of any human mental activity. That objective structure is to be discovered by us, and that is where—on his account—optimality comes into play. In contrast, Douven and Gärdenfors are concerned with showing what the natural concepts are, not—in first instance—how we latch onto those concepts. It is in

To accomplish (i), Douven and Gärdenfors state a number of (what they call) design criteria and constraints, where design *criteria* hold generally for conceptual structures, independently of the representational format we choose for concepts (so, in particular, independent of the conceptual spaces format), while design *constraints* pertain explicitly to one specific type of conceptual structure, to wit, conceptual spaces. The criteria and constraints are all supposed to pertain to conceptual structures for creatures that, like us, have memories with limited storage capacity; are not able to detect arbitrarily small differences between stimuli; and seek to thrive in a world in which commodities are scarce and competition for those commodities can be fierce. The criteria and constraints are meant to help endow such creatures with conceptual structures that increase their chances of long-term survival and of reproduction.

Douven and Gärdenfors list the following design criteria:

> PARSIMONY
>
> The conceptual structure should tax the system's memory as little as possible.
>
> INFORMATIVENESS
>
> The concepts should be informative, meaning that they should jointly offer good and roughly equal coverage of the domain of classification cases.

---

answering the first question that they appeal to optimality considerations. (Thanks to Daniel Lassiter for pressing me on this.)

REPRESENTATION

The conceptual structure should be such that it allows the system to choose for each concept a prototype that is a good representative of all items falling under the concept.

CONTRAST

The conceptual structure should be such that prototypes of different concepts can be so chosen that they are easy to tell apart.

LEARNABILITY

The conceptual structure should be learnable, ideally from a small number of instances.

As mentioned, the creatures to whose conceptual structures these criteria are supposed to pertain have limited memory capacity, which explains the presence of PARSIMONY on the list.[7] On the other hand, conceptual structures should be useful; in particular, they should allow the creatures to make sufficiently fine-grained distinctions between whichever items they may encounter, as their success will depend on this capacity—which explains INFORMATIVENESS. REPRESENTATION and CONTRAST are both motivated by the creatures' limited discriminatory capacities in combination with

[7] Marzen and DeDeo (2017) formally show how cost considerations in organisms motivate parsimonious representational systems, even though this tends to come at the cost of greater inaccuracy.

the fact that the creatures will need to avoid making classification errors as much as possible. Finally, the creatures will have to be able to use the conceptual structures after a relatively short time period following birth, which is why LEARNABILITY is on the list.

As for design *constraints*, CRITERION P is among them, now taken to derive from REPRESENTATION (convex regions make it easy to pick representative prototypes: just pick their centers, or small regions surrounding and including their centers) and PARSIMONY (recall that Gärdenfors presented CRITERION P as a principle of cognitive economy). There is also the constraint (adapted from the machine-learning literature on clustering; e.g., Kaufman and Rousseeuw 1990) that a conceptual space should be such that items falling under any of its concepts are maximally similar to each other and maximally dissimilar from the items falling under any of the other concepts represented in the same space. Following Regier, Kay, and Khetarpal (2007), Douven and Gärdenfors label this constraint WELL-FORMEDNESS.

To make the foregoing more concrete, and to show how the above principles may help to fix a furnishing of a similarity space with concepts, suppose that a group of engineers are to endow an artificial creature with a system of concepts. They are asked to work under the assumption that the creature's limitations are relevantly similar to ours (the same limitations on storage capacity, discriminatory capacities, available time to familiarize itself with the conceptual system) and also that it has to operate under much the same pressures that we face, although at the point of design no specifics are known about the environment that it will inhabit. Thus the engineers are asked to solve a *constrained optimization* problem.

They tackle the task in a piecemeal fashion, starting with the subtask of providing the conceptual structure for one particular perceptual domain. By eliciting from the creature similarity judgments concerning the given domain, and by applying one of the statistical dimensionality-reduction techniques briefly mentioned in section 2, they are able to ascertain that the corresponding similarity space is a perfectly round disk with a Euclidean metric defined on it.

In trying to determine the optimal conceptual structure for this space—optimal given the system's limitations—they pay heed to the design criteria and constraints stated above. It is clear to the engineers that these criteria and constraints will sometimes pull in different directions, so that their goal must be to find a conceptual structure that does best *on balance*.

The engineers begin by comparing the six conceptual structures in Figure 3. Structure A does very well on INFORMATIVENESS: it endows the space with many concepts, and all parts of the space are covered equally. However, it is more than likely that A will overtax the system's memory and that it will therefore do poorly on PARSIMONY. Structure B does much better in the latter respect; but although the engineers believe that eight concepts may be enough for this similarity space, the coverage offered by B is very unequal: it allows the system to make relatively fine distinctions in the central part of the space but only coarse distinctions in the peripheral parts. Structure C, then, might be an acceptable compromise between INFORMATIVENESS and PARSIMONY.

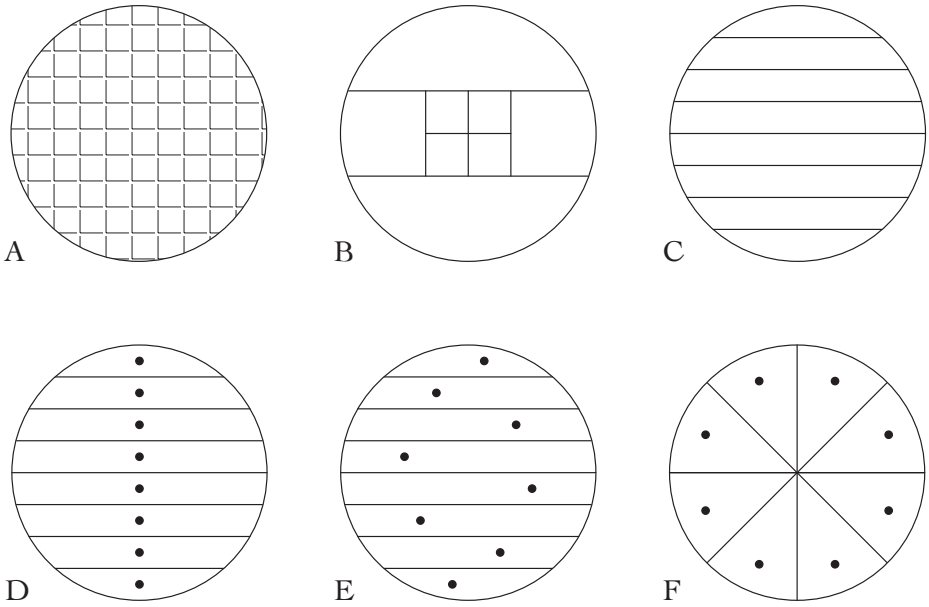

Figure 3: Six candidate conceptual structures for a disk-shaped similarity space.

The engineers realize, however, that structure C is still far from ideal, given that it makes it hard to jointly satisfy REPRESENTATION and CONTRAST. Consider D, which equips each concept in C with a prototype. D locates the prototypes in a perfectly symmetrical fashion, which seems to be the best way to make them representative—it minimizes the average Euclidean distance to all points in the concept corresponding to the given prototype—even though (especially for the concepts represented by the segments in the middle of the space) the prototypes are quite distant from the points that fall under the concept but lie close to the border of the space.

More problematic about D is that prototypes of adjacent concepts lie very close together, and lie much closer to each other than, for instance, to points in the same concept that are close to the border of the space. In fact, the prototypes lying nearest to

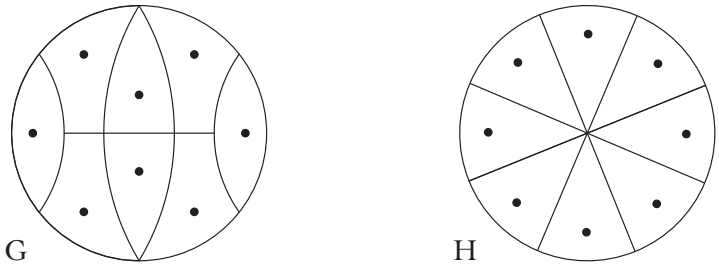

Figure 4: Conceptual structure with some nonconvex concepts (G) and rotation of structure F from Figure 3 (H).

the center of the disk lie closer to almost all other prototypes than to some points in the concepts of which they are the prototypes. So, the prototypes are not contrastive at all.

One could try to mend this defect by slightly shifting each of the prototypes, perhaps as done in E. But while this makes the prototypes a bit more distinct from each other, doing so comes at the expense of representativeness. Consider, for example, that the prototypes of the two middle-segment concepts are closer, and thus more similar, to almost *any* point in all the other concepts than they are to some of the points that lie in their "own" concept.

In short, while structure C may be an acceptable compromise between INFORMATIVENESS and PARSIMONY, it scores poorly on the counts of REPRESENTATION and CONTRAST: it does not allow the system to choose prototypes that are both sufficiently representative of their concepts while also being sufficiently different from the prototypes of other concepts.

Now consider structure F, which divides the space into eight concepts, like B and C do, and also offers equal coverage, providing another good compromise between INFORMATIVENESS and PARSIMONY. But F also offers room for making the prototypes both representative of their concept and distinct from each other. Of course, the prototypes can still be placed a bit more toward the center of the disk, which might make them more representative, or placed a bit closer to the border of the space, which would make them a bit more distinct. But that can all be left to the artificial creature itself. The engineers are to make sure that the creature can make such choices in the first place.

Surely, however, there are partitions of the disk-shaped similarity space other than F that have these virtues. For instance, G in Figure 4 may appear to do quite well, too, in terms of PARSIMONY, INFORMATIVENESS, REPRESENTATION, and CONTRAST. But note that if CRITERION P is entailed by PARSIMONY and REPRESENTATION, as Douven and Gärdenfors (2018) argue, then G is ruled out, given that some of the concepts it creates are non-convex.

If we want to go along with Douven and Gärdenfors and claim that, starting from some obvious design principles, we may be able to identify the "natural" conceptual architecture for the disk-shaped similarity space, then structure H shown in Figure 4 seems to pose more of a challenge. After all, while F seems to satisfy the engineers' design principles, so does any rotation in the space of the conceptual structure *cum* prototypes—such as H. We thus seem compelled to admit that there is an abundance of natural structures, and so an abundance of natural concepts (even if not every convex

region corresponds to a natural concept), and that seems just wrong: natural concepts are supposed to be *sparse* (Lewis 1983).

The response to this challenge is two-pronged: First, I claim that the problem of rotational variants stems from the fact that we are considering a fictional, abstract similarity space. It might immediately be countered that this cannot be right, for look again at perceptual color space, as shown in Figure 1, which is not a fictional, abstract similarity space. It seems that design principles of the kind considered so far are fundamentally unable to determine a conceptualization of that space yielding natural concepts, given that any rotation along the luminance axis is structure-preserving in the relevant sense: any non-trivial rotation along that axis of a structure satisfying the design principles introduced so far will satisfy those principles too, although it will yield different color concepts. It is to be stressed, however, that Figure 1 only shows an *approximation* of perceptual color space. The approximation is quite good, yet real color space is not nearly as symmetric as the spindle shown in Figure 1. Just google images of CIELAB space and you will see what I mean.[8] Important work by Regier, Kay, and Khetarpal (2007) suggests that the asymmetry of CIELAB space in conjunction with WELL-FORMEDNESS may even be enough to determine a *unique*, or at least *near-unique*, partitioning of that space into color concepts; see also Douven (2017) and Jraissati and Douven (2017).

[8] Or google images of CIELUV space, which is another, slightly different, perceptual color space; see Malacara (2002, pp. 86–90) for details and for a discussion of how CIELAB and CIELUV spaces differ.

Second, the above example is meant to illustrate how, in the context of the conceptual spaces framework, design considerations may be able to pin down a conceptual structure for a similarity space, once that similarity space has been determined. In other words, the point is to show how (taking into account the constraints under which a creature is to operate) there may be an optimal architecture for each of the creature's similarity spaces, where this architecture can then be said to yield a set of concepts that are natural. Douven and Gärdenfors' claim is *not* that the five design criteria they introduce are guaranteed in every case to be jointly sufficient for the task of identifying the optimal structure. In fact, they leave open the possibility that the design criteria that determine optimality cannot be found strictly through a priori reflection.

As said, Douven and Gärdenfors' 2018 paper consists of two main parts: one in which they propose, and argue for, the design criteria and constraints mentioned above, and another in which they muster experimental findings reported in other publications—typically also in very different contexts—which actually yield support for the design view of naturalness. I have already mentioned Regier, Kay, and Khetarpal's (2007), which shows how WELL-FORMEDNESS helps explain the commonalities we find in color lexicons from cultures across the globe, as registered in the World Color Survey (see Cook, Kay, and Regier 2005). Kemp and Regier (2012), Xu and Regier (2014), and Xu, Regier, and Malt (2016) present similar results concerning kinship categories, numerical systems, and container categories, respectively. Finally, there is work on social learning by Jäger (2007), van Rooij and Jäger (2007), and Xu, Dowman, and Griffiths (2013), which is discussed at some length in Douven and Gärdenfors (2018) and which provides evidence for holding that INFORMATIVENESS, PARSIMONY, and LEARNABILITY are operative in the cognitive domain.

Rationality and Vagueness

According to Douven and Gärdenfors (2018), natural concepts are those represented by the cells of an optimal partitioning of a similarity space. In other words, the natural partitioning into concepts of a given domain is the one that is most rational, in that it is the one that a good engineer would provide if she were asked to design our conceptual apparatus for us. So, on this account, there is a clear link between rationality and categorization. On the proposal to be made now, however, that account only describes the first step toward categorization. It is as if an architect had come up with a first rough sketch of your to-be-built house, reflecting a rational division of the available space, based on your needs and desires, but leaving out some important finer details. The filling-in of those details is itself subject to rationality principles.

I am actually not sure whether "details" is the right word in the present case. Here, the details concern vagueness and prototypes, which are of central importance to categorization. Douven and Gärdenfors (2018) explicitly bracket the former issue. And while, on their account, there is a connection between optimal categorization and prototypes—most notably, via REPRESENTATION and CONTRAST—prototypes themselves are not yet in the picture; for Douven and Gärdenfors' purposes, exactly where the prototypes end up being located in a space is not relevant to what makes concepts natural.

In this section, I consider what a straightforward extension of Douven and Gärdenfors' account of naturalness might look like, where the extension should be sensitive both to

the fact that most concepts are vague and to the fact that conceptual spaces are equipped with prototypes. The claim is not just that the account of vagueness from section 3 fits naturally with the account of naturalness from section 4, but that their combination flows from rationality considerations and that, consequently, vagueness can be thought of as resulting from those considerations.

Specifically, I argue that (i) it is rational to equip conceptual spaces with prototypical regions rather than with just prototypical points; (ii) it is rational to put each prototypical region at a location such that the prototypes are as representative of their concepts as possible and as different as possible from the prototypes of the other concepts in the space; and (iii) it is rational to obtain the final carving-up of a similarity space by constructing a collated Voronoi tessellation, in the way explained in section 3, on the basis of the prototypical regions placed in line with (ii). Clauses (i) and (ii) together form the second step of categorization, and clause (iii) is the third and final step of categorization.

The Second Step of Categorization In Douven and Gärdenfors' (2018) proposal, REPRESENTATION and CONTRAST refer to prototypes, which throughout their paper are supposed to appear as isolated points in a conceptual space. We know from section 3 that this can only be an idealization, and that it is more realistic to think of conceptual spaces as having prototypical regions. The latter claim is justified by a wealth of evidence, both from introspection and from experimental studies, showing that going by people's judgments, many concepts have multiple "best" instances. Importantly, that many concepts have prototypical regions (as opposed to prototypical points) is not

just true as a matter of empirical fact; it is a fact for which there exists a rational explanation.

In the psychological literature it is often said that we construct prototypes by abstracting from concrete instances we have encountered, as if we create a summary of those instances (see Gärdenfors 2000, Ch. 4.5). If so, learning a concept *C* is strictly a matter of being shown certain items, which are labeled as *C* by one's parents or teachers, whence one should try to figure out what makes those items relevantly similar.

No doubt this is part of the process of concept acquisition. But it is far from the whole story. Most notably, it completely ignores the educational role that the term "typical" and others play in the practice of teaching the meaning of a word. We *often* use "typical" when we want to convey the meaning of a word to a child: we do not just tell her that that thing over there is yellow; we tell her that the thing is *typically* yellow, or that this lemon's taste is *typically* sour, and so on. In other words, often we help a child to acquire a concept not just by showing it various items falling under that concept, labeling them with whatever predicate we use for the concept, and hoping that the child will somehow "get" what a best representative of the concept looks like; but by directly presenting some best representative to it, stressing that this *is* a best representative of the relevant concept. That practice will greatly help to speed up the learning process.

Naturally, multiple prototypes for a concept will offer more opportunities for pointing out what is typical for that concept. And the more such opportunities we have, the

more we will be able to speed up the learning process. In particular, we are then likely to teach the concept much faster than if there were just one prototype, with correspondingly fewer opportunities to use the word "typical" in relation to the concept, so that we have to rely much more on the child or student to discover for herself the important commonalities of the items we are grouping under the same label.

In fact, suppose there were exactly one BLUE prototype. In color space, this would correspond to a single point, which has measure 0. The chance that we would ever encounter exactly that shade, and thus be able to point it out to a child or student, would be essentially nil. By contrast, the prototypical BLUE region as identified in Douven et al. (2017) has a positive measure, and we can be as good as certain that we all have encountered plenty of shades of blue falling into that region.[9]

It is worth mentioning that, so far, all empirical evidence on prototypical regions is consistent with the assumption that such regions are convex (see, e.g., Douven 2016 and Douven et al. 2017). LEARNABILITY provides independent and a priori support for that assumption. After all, if we can assume that prototypical regions are convex, then the learning of which regions in a space represent the prototypes becomes very efficient: by learning of just a handful of shades that they are typical instances of, say,

[9] To be entirely precise, Douven et al. (2017) give a number of different estimates of the prototypical BLUE region in CIELUV space, but what is said above holds given any of those estimates.

BLUE, we then automatically learn that every shade represented by a point inside the convex hull of those typical instances is a typical instance of BLUE as well.

The second step of categorization also concerns the question of where in a given space—already equipped with the partitioning resulting from the first step—to *place* the prototypical regions. Here, too, there are *rational* decisions to be made: we want the prototypical regions to be representative of their concepts, and we also want them to be sufficiently distinct from each other, to facilitate memorization and to limit the chance of misclassifications, as previously explained. It is a matter of rational design to ensure that prototypical regions *can* be located in a space so that they are both representative and contrastive, as REPRESENTATION and CONTRAST jointly imply, precisely because it is rational to *have* prototypical regions that are representative and contrastive.

As an aside, it is to be noted that our discussion of structure F above made it clear that rationality may leave some leeway here. It can happen that, by making the prototypical regions less central, and so less representative, they become more distinct from one another. And if such a trade-off is to be made, there may be no single best way to do it. Possibly, different rational people will make different trade-offs in such cases, which would also explain why we find some individual variation in people's responses to questions asking them to designate the most typical item or items falling under a given concept (see Berlin and Kay 1969; Douven et al. 2017).

The Third Step of Categorization There is a third, and final, step of categorization in my proposal here, namely, that of generating a collated Voronoi tessellation from the

prototypical regions, as described in section 3. This procedure remains unaltered here. It is worth pointing out, however, that the procedure has a rational basis as well.

Defending his proposal of combining the conceptual spaces framework with prototype theory and the technique of Voronoi tessellations, Gärdenfors (2000, 89) convincingly argues that it yields a "cognitively economical way of representing information about concepts," the reason being that it is much more cost-effective to keep in memory the locations of the prototypes and to compute the categorization of any given point in the space by finding the prototype nearest to it than to remember the category of each point in a space.

Since we can just as easily measure distances between points as those between regions, or between regions and points, Gärdenfors' motivation for appealing to the technique of Voronoi tessellations is also valid for the extended technique of collated Voronoi tessellations. Admittedly, as can easily be shown, the procedure of just looking which prototypical region a given point is nearest to will not always tell you whether that point is a clear case or a borderline case. For that, one should ask whether the point is closer to *every* point in that region than it is to *any* point in *any* other prototypical region in the same space. But answering *that* question is probably *not* very cost-effective. So, when dealing with prototypical regions, the designated procedure can at best serve as a heuristic.

However, the procedure should still be good enough for practical purposes. And in any event, we may not be able to improve upon it, given that (as briefly mentioned in section 3) psychological measures tend to be imprecise. So, while following the

heuristic may not always give an exact answer as to which concept a given point falls under, trying to answer this question in the more involved way will not always yield a precise answer either—because of the inherent inexactitude of measuring distances in whichever space the point lies—and may in general not even yield a more exact answer than the one we get from the heuristic.

Discussion and Directions for Future Work In short, the proposal is that categorization can be thought of as consisting of three steps: in the first step, design criteria and constraints—which are at bottom rationality principles—determine a provisional partition of a given space; in the second step, prototypical regions are put into place, guided by the desiderata of representativeness and contrastiveness, where the fact that we are referring to regions has a rational motivation itself; and in the third step, the final conceptual architecture is achieved by generating a collated Voronoi tessellation from the prototypical regions, a procedure that also has a rational backing, as explained above. The process of going through these steps is illustrated in Figure 5 for an abstract two-dimensional similarity space. (To forestall misunderstanding, there is in general no reason to expect all prototypical regions to have the same size and shape—as the second panel of Figure 5 might seem to suggest—nor that the boundaries regions obtained in the final step will always be "well-aligned" with the borderlines obtained in the first step, as a comparison of the left and right panels might suggest.)

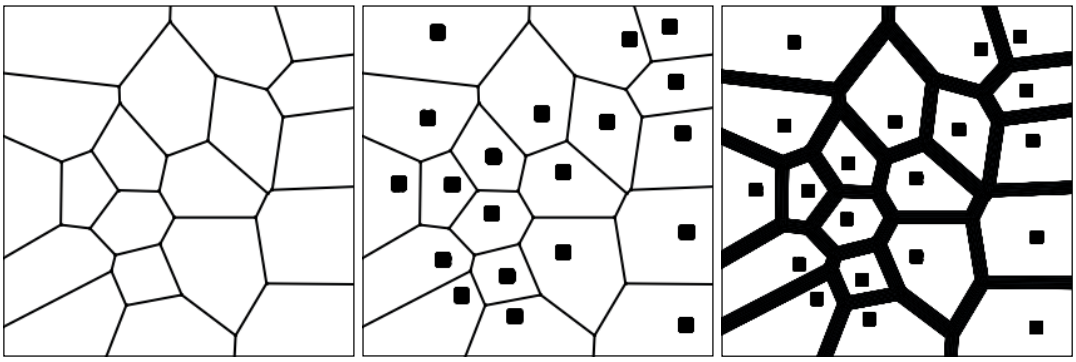

Figure 5: Results of determining the rough conceptual architecture of a given similarity space (left), the locations of the prototypical regions in that architecture (middle), and the collated Voronoi tessellation generated by those regions (right).

It cannot be stressed enough that what I have tried to accomplish in this paper is *not* to present a plausible phenomenology of the process of categorization. Rather, I have tried to give content to the notion of an optimal cognitive architecture. Douven and Gärdenfors (2018) have already argued that the notion of naturalness as it pertains to concepts is to be understood in terms of optimally partitioned similarity spaces. However, they admit that their proposal cannot be the complete story. In particular, vagueness is explicitly set aside, and the location of prototypes receives no discussion. The current proposal is meant to round out Douven and Gärdenfors' account of naturalness, or at least to take it some important steps further. Natural concepts are those that reside in optimally designed conceptual structures, where it is now seen that the vagueness of concepts also results from the optimality—that is, *rationality*—of the design.

I want to remain neutral, at least for now, on how optimality of the kind at issue is achieved. It is clear that we do not consciously take design decisions to arrive at our concepts. For all that has been said, the verdict that *this* is a good spot for placing a prototype may well manifest itself "from the inside" as an overwhelming sense of typicality. But then again, much or even all of what leads us to have the concepts we have may be hard-wired into our physiology, or derive from mechanisms that are hard-wired. To repeat, the point I want to make is that we can make sense of the notion of an optimally designed conceptual architecture, and that we can do so in a way that makes vagueness a matter of rationality rather than some defect for which we may or may not be able to find an excuse. But the rationality involved is not necessarily *ours*. Rather—to come back to an image from the introduction—the point is that if somehow logicians and linguists had devised our language, they would, considering our physiology as well as the world we inhabit, probably have come up with a language that exhibits vagueness much in the way and to the extent that spoken languages actually do.

In closing, I would like to mention that while important parts of the proposal already enjoy empirical support—there is the experimental work on vagueness cited in section 3, and Douven and Gärdenfors (2018) discuss at length evidence for their proposal concerning naturalness—so far little research has been done on the question of which principles may determine the locations of prototypes or prototypical regions in conceptual spaces. As mentioned, Douven and Gärdenfors leave open the possibility that the design criteria and constraints actually at work in structuring conceptual spaces are not exactly the ones they present in their paper, or are not restricted to those presented. Similarly, it is plausible that considerations of contrastiveness and

representativeness play a role in determining where to locate prototypes in a space, but this is by no means certain.

One way to find out would be to see whether we can get good predictions of where the prototypical regions are to be found in a given space by assuming that their locations strike the best balance between (i) being as centrally as possible located in the respective concepts and (ii) being as distant from each other, on average, as is compatible with their still lying in their concepts. To give a concrete example, suppose we knew how the basic color terms partition CIELAB space and also where in that space the prototypical regions are located. That constellation of prototypical regions could then be compared, in terms of both representativeness and contrastiveness, with other possible constellations, and we might find that on balance none of those other constellations do as well as the actual one.

More concretely still, note that we could randomly relocate each prototypical region within its concept, and that we could consider, in simulations, thousands or perhaps millions of such random relocations. If it then turned out that those relocations which result in higher scores on one count (representativeness or contrastiveness) tend to result in lower scores on the other count, that would offer strong support for the part of our proposal that so far is still largely unsubstantiated.

Contrary to what many might think, the partition of CIELAB space that this test would require is not yet known, nor is the constellation of the prototypical regions in that

space.[10] This is mainly due to the fact that color-naming studies have hitherto been conducted only on the "hull" of that space, typically by using the 330 Munsell chips that were used in the World Color Survey—chips that are all at maximum saturation for their hue–value combination. But color-naming studies, as well as studies asking for typicality judgments that use stimuli sampled from throughout the space, are currently underway, and the data from those studies should enable us to perform the test just described, and should thereby allow us to verify or falsify, as the case may be, claims concerning the design principles behind the locations of prototypical regions.[11]

References

Anderson, J. R. (1990) *The Adaptive Character of Thought*. New York: Psychology Press.

Berlin, B. and Kay, P. (1969) *Basic Color Terms*. Stanford CA: CSLI Publications.

[10] Douven et al. (2017) present the empirically determined locations of the BLUE and GREEN prototypical regions in CIELUV space; the locations in that space of other prototypical regions are still unknown.

[11] I am greatly indebted to Richard Dietz, Daniel Lassiter, Christopher von Bülow, and an anonymous referee for very valuable comments on previous versions of this paper. I am also grateful to an audience at the Ruhr University Bochum for stimulating questions and remarks.

Borg, I. and Groenen, P. (2010) *Modern Multidimensional Scaling* (2nd ed.). New York: Springer.

Cook, R. S., Kay, P., and Regier, T. (2005) The World Color Survey Database: History and Use. In H. Cohen and C. Lefebvre (eds.) *Handbook of Categorization in Cognitive science*, 223–42. Amsterdam: Elsevier.

Decock, L. and Douven, I. (2014) What is Graded Membership? *Noûs* 48, 653–82.

Douven, I. (2016) Vagueness, Graded Membership, and Conceptual Spaces. *Cognition* 151, 80–95.

Douven, I. (2017) Clustering Colors. *Cognitive Systems Research* 45, 70–81.

Douven, I. and Decock, L. (2017) What Verities May Be. *Mind* 126, 385–428.

Douven, I., Decock, L., Dietz, R., and Égré, P. (2013) Vagueness: A Conceptual Spaces Approach. *Journal of Philosophical Logic* 42, 137–60.

Douven, I. and Gärdenfors, P. (2018) What Are Natural Concepts? A Design Perspective. *Mind & Language*, in press.

Douven, I., Wenmackers, S., Jraissati, Y., and Decock, L. (2017) Measuring Graded Membership: The Case of Color. *Cognitive Science* 41, 686–722.

Gärdenfors, P. (2000) *Conceptual Spaces: The Geometry of Thought*. Cambridge MA: MIT Press.

Gärdenfors, P. and Zenker, F. (2013) Theory Change as Dimensional Change: Conceptual Spaces Applied to the Dynamics of Empirical Theories. *Synthese* 190, 1039–58.

Jäger, G. (2007) The Evolution of Convex Categories. *Linguistics and Philosophy* 30, 551–64.

Jraissati, Y. and Douven, I. (2017) Does Optimal Partitioning of Color Space Account for Universal Color Categorization? *PLOS ONE* 12(6): e0178083, https://doi.org/10.1371/journal.pone.0178083.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data*. Hoboken NJ: Wiley.

Kemp, C. and Regier, T. (2012) Kinship Categories Across Languages Reflect General Communicative Principles. *Science* 336, 1049–54.

Krumhansl, C. L. (1978) Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship Between Similarity and Spatial Density. *Psychological Review* 85, 445–63.

Lewis, D. K. (1983) New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61, 343–77.

Malacara, D. (2002) *Color Vision and Colorimetry: Theory and Applications*. Bellingham WA: SPIE Press.

Marzen, S. E. and DeDeo, S. (2017) The Evolution of Lossy Compression. *Journal of the Royal Society Interface* 14:20170166, http://dx/doi.org/10.1098/rsif.2017.0166.

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

Regier, T., Kay, P., and Khetarpal, N. (2007) Color Naming Reflects Optimal Partitions of Color Space. *Proceedings of the National Academy of Sciences USA* 104, 1436–41.

Rosch, E. (1973) Natural Categories. *Cognitive Psychology* 4, 328–50.

Rosch, E. (1975) Cognitive Reference Points. *Cognitive Psychology* 7, 532–47.

Valentine, T., Lewis, M. B., and Hills, P. J. (2016) Face-Space: A Unifying Concept in Face Recognition Research. *Quarterly Journal of Experimental Psychology* 69, 1996–2019.

van Rooij, R. and Jäger, G. (2007) Language Structure: Psychological and Social Constraints. *Synthese* 159, 99–130.

Verheyen, S. and Égré, P. (2018) Typicality and Graded Membership in Dimensional Adjectives. *Cognitive Science*, in press.

Xu, J., Dowman, M., and Griffiths, T. (2013) Cultural Transmission Results in Convergence towards Colour Term Universals. *Proceedings of the Royal Society B* 280:20123073.

Xu, Y. and Regier, T. (2014) Numeral Systems across Languages Support Efficient Communication: From Approximate Numerosity to Recursion. In P. Bello, M. Guarini, M. McShane, and B. Scassellati (eds.) *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin TX: Cognitive Science Society.

Xu, Y., Regier, T., and Malt, B. C. (2016) Historical Semantic Chaining and Efficient Communication: The Case of Container Names. *Cognitive Science* 40, 2081–94.